Degree Project in Sustainable Energy Engineering

Second Cycle, 30 credits

Can AI Replicate Expert Climate Policy Evaluations in a Swedish Context?

A Controlled Comparison of Open-Weight Models and RAG Architectures

Henning Ottervald Flygare

Author

Henning Ottervald Flygare ⟨heof@kth.se⟩
Energy System Analysis
Sustainable Energy Engineering
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden

Examiner

Associate Professor Francesco Fuso Nerini
Unit of Energy Systems
KTH Royal Institute of Technology

Supervisor

Associate Professor Åsa Persson
Unit of Energy Systems
KTH Royal Institute of Technology

Abstract

Large language models, AI systems that generate text, are increasingly used in climate policy analysis, but it is rarely tested whether their analysis agrees with expert assessment, whether giving them relevant documents helps, or whether common evaluation methods can tell good analysis from weak. This thesis tests all three for Swedish climate and energy policy.

Seven openly available models, run on local hardware, answered ten policy questions in four set-ups: without documents, or after searching a collection of Swedish policy documents once, through several sub-searches, or repeatedly while critiquing their own draft. The 280 answers were compared with reference answers generated by an AI model from the summary of the Swedish Climate Policy Council's 2026 report, held out from the collection, and scored by text-similarity measures, a citation check, two AI judges and a blinded human reader.

The four set-ups scored almost the same, and the more complex ones took several times as long and used more energy for no measurable gain. Because searching was changed together with the instruction to cite sources, the design cannot show cause and effect. The one clear benefit of searching was traceable citations. No method could reliably rank the usable answers: the two AI judges often disagreed, and the human reader favoured different set-ups than the judges. The AI judges, the citation check and the human reader did single out the same model as unusable, so such methods can screen out poor output but not select the best.

Read as climate and energy analysis, on four energy and land-use questions, the best answers named conflicts between sectors that the references left out, such as battery metals against reindeer herding and biomass wanted by both transport and industry, but did not rank or quantify them, and the answers agreed most on the least useful point, a call for coordination. On Sweden's 2030 carbon sink commitment, most answers and the reference reached opposite verdicts largely because they used different releases of the same statistic, which changes how much emission reduction transport and heating must deliver. One answer took a contested forestry position as its starting point while appearing balanced. None of the evaluation methods caught either problem. On the electricity system, several answers covered the question more fully than the reference, which shows that expert-written references are needed before such tools can be judged fairly. The tools may support scoping and evidence tracing, with dated statistics and expert review, but not replace expert judgement.

Keywords

Retrieval-Augmented Generation; Large Language Models; Climate Policy Analysis; Energy System Analysis; LLM-as-Judge; Evaluation Methods; Swedish Climate Policy

Sammanfattning

Stora språkmodeller, AI-system som genererar text, används allt oftare i klimatpolitisk analys, men det prövas sällan om deras analys stämmer överens med expertbedömningar, om det hjälper att ge dem relevanta dokument, eller om vanliga utvärderingsmetoder kan skilja god analys från svag. Denna uppsats prövar alla tre frågorna för svensk klimat- och energipolitik.

Sju öppet tillgängliga modeller, körda på lokal hårdvara, besvarade tio policyfrågor i fyra upplägg: utan dokument, eller efter att ha sökt i en samling svenska policydokument en gång, genom flera delsökningar, eller upprepade gånger medan de granskade sina egna utkast. De 280 svaren jämfördes med referenssvar som en AI-modell tagit fram ur sammanfattningen av Klimatpolitiska rådets rapport för 2026, som hölls utanför samlingen, och poängsattes med textlikhetsmått, en kontroll av källhänvisningar, två AI-domare och en blindad mänsklig bedömare.

De fyra upplägen fick nästan samma poäng, och de mer komplexa tog flera gånger så lång tid och använde mer energi utan mätbar vinst. Eftersom sökningen ändrades samtidigt med instruktionen att ange källor kan designen inte visa orsak och verkan. Den enda tydliga fördelen med sökning var spårbara källhänvisningar. Ingen metod kunde tillförlitligt rangordna de användbara svaren: de två AI-domarna var ofta oense, och den mänskliga bedömaren föredrog andra upplägg än domarna. AI-domarna, källkontrollen och bedömaren pekade dock ut samma modell som oanvändbar, så de kan gallra bort dåliga svar men inte välja ut de bästa.

Lästa som klimat- och energianalys lyfte de bästa svaren på fyra energi- och markanvändningsfrågor fram konflikter mellan sektorer som referenssvaren utelämnade, till exempel batterimetaller mot rennäring och biomassa som efterfrågas av både transporter och industri, men varken rangordnade eller kvantifierade dem, och svaren var mest eniga om den minst användbara punkten, en uppmaning till samordning. Om Sveriges åtagande för kolsänkan till 2030 drog de flesta svaren och referensen motsatta slutsatser, till stor del för att de använde olika versioner av samma statistik, vilket påverkar hur mycket utsläppen måste minska inom transporter och uppvärmning. Ett svar utgick från en omstridd ståndpunkt i skogsdebatten men framstod ändå som balanserat. Ingen av utvärderingsmetoderna fångade något av problemen. För elsystemet täckte flera svar frågan mer fullständigt än referensen, vilket visar att expertskrivna referenssvar behövs innan sådana verktyg kan bedömas rättvist. Verktygen kan stödja kartläggning av frågor och spårning av underlag, med daterad statistik och expertgranskning, men inte ersätta expertbedömning.

Nyckelord

Retrieval-Augmented Generation; stora språkmodeller; klimatpolitisk analys;
energisystemanalys; LLM-som-domare; utvärderingsmetoder; svensk klimatpolitik

Acknowledgements

I would like to thank my supervisor, Åsa Persson, whose advice has been invaluable. Furthermore, her work in the field of effective climate policy has greatly inspired both me and this thesis.

I wish to express my utmost gratitude to the following researchers, whose feedback has contributed greatly to the structuring and quality of the work:

- Adis Dzebo at the SEI for discussions regarding the processing of policy papers and ensuring high-quality results.
- Sina Sheikholeslami at KTH Royal Institute of Technology for the insights into LLM framework structuring and state-of-the-art model advice.
- Charlie Wilson at the University of Oxford for taking the time to discuss the potential of ML methods in climate-policy assessment, a topic outside the scope of this work that I will pursue further in my upcoming research internship.

I would also like to thank the KTH Climate Action Centre, my employer throughout my second-cycle studies. The perspectives shared by colleagues and the centre's dedication to SDG 13 have inspired me to pursue a career contributing to effective measures in the face of increasing global warming.

Henning Ottervald Flygare

Stockholm, September 2026

Note on AI assistance

Large language models were used in the preparation of this thesis in two capacities. First, as coding assistants during the development of the experimental pipeline and evaluation framework. Second, as writing tools to support drafting and editing of selected sections. All experimental design, analytical interpretations, and conclusions are the author's own.

Nomenclature

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BECCS	Bio-Energy with Carbon Capture and Storage
BERT	Bidirectional Encoder Representations from Transformers
CAPMF	Climate Actions and Policies Measurement Framework
CBAM	Carbon Border Adjustment Mechanism
CCS	Carbon Capture and Storage
CI	Confidence Interval
DACCS	Direct Air Carbon Capture and Storage
DiVA	Digitala Vetenskapliga Arkivet (Digital Scientific Archive)
EDGAR	Emissions Database for Global Atmospheric Research
ESR	Effort Sharing Regulation
EU	European Union
FRL	Forest Reference Level
HyDE	Hypothetical Document Embeddings
IPCC	Intergovernmental Panel on Climate Change
IRR	Inter-Rater Reliability
JSON	JavaScript Object Notation
LLM	Large Language Model
LULUCF	Land Use, Land-Use Change and Forestry
MaxDiff	Maximum Difference Scaling
ML	Machine Learning
MMTEB	Massive Multilingual Text Embedding Benchmark
MoE	Mixture of Experts
NLP	Natural Language Processing

OECD	Organisation for Economic Co-operation and Development
PUE	Power Usage Effectiveness
RAG	Retrieval-Augmented Generation
RISE	Research Institutes of Sweden
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RQ	Research Question
SDG	Sustainable Development Goal
SEB	Scandinavian Embedding Benchmark
SEI	Stockholm Environment Institute
SMHI	Swedish Meteorological and Hydrological Institute
SOU	Statens Offentliga Utredningar
VTI	Swedish National Road and Transport Research Institute

Contents

1 Introduction **1**
- 1.1 Background . . . 1
- 1.2 Theory and State of the Art . . . 2
  - 1.2.1 Methodologies for Evaluating Climate Mitigation Policies . . . 2
  - 1.2.2 Language-Based AI and Policy Evaluation . . . 2
  - 1.2.3 Evaluating AI-Generated Text . . . 3
  - 1.2.4 AI Hallucination and Mitigation . . . 4
  - 1.2.5 Context-Memory Conflict . . . 5
  - 1.2.6 Synthesis and Knowledge Gap . . . 5
- 1.3 Aim . . . 6
- 1.4 Research Questions . . . 6
- 1.5 Delimitations . . . 7

2 Methodology **8**
- 2.1 Qualitative LLM Framework . . . 8
  - 2.1.1 Corpus and Data Sources . . . 8
  - 2.1.2 Document Preprocessing and Indexing . . . 9
  - 2.1.3 Language Model Configuration . . . 10
  - 2.1.4 Experimental RAG Tiers . . . 11
  - 2.1.5 Benchmark Design . . . 12
- 2.2 Evaluation Framework . . . 14
  - 2.2.1 Automated Text and Semantic Metrics . . . 14
  - 2.2.2 LLM-as-Judge Scoring . . . 14
  - 2.2.3 Statistical Analysis . . . 15
  - 2.2.4 Citation Traceability . . . 16
  - 2.2.5 Human Evaluation and Judge Calibration . . . 17
- 2.3 Prompt Sensitivity Analysis . . . 18
  - 2.3.1 Context Format: Tabular versus Linearised . . . 18
  - 2.3.2 Judge-Prompt Language . . . 19

3 Results **20**
- 3.1 Overview of Experimental Runs . . . 20
- 3.2 Aggregate Performance Across Conditions . . . 21
  - 3.2.1 Per-Model Condition Means . . . 21
  - 3.2.2 Cross-Model Condition Ranking . . . 22
  - 3.2.3 Within-Model and Pooled Significance Tests . . . 23
- 3.3 Per-Question Performance . . . 25
- 3.4 Dimension-Level Breakdown . . . 25
- 3.5 Citation Behaviour . . . 26

3.5.1 Closed-Universe Citation Use . . . 26
3.5.2 Citation Traceability . . . 27
3.6 Overlap with the Gold Standard . . . 28
3.6.1 Lexical and Semantic Overlap . . . 28
3.7 Agreement Between the Two Automated Judges . . . 29
3.8 Preliminary Human Evaluation . . . 29
3.8.1 Human Condition Ranking . . . 30
3.8.2 Human–Judge Agreement . . . 31
3.9 Efficiency and Cost . . . 33
3.9.1 Energy Use and Carbon Footprint . . . 33
3.10 Prompt Sensitivity Analysis . . . 35
3.10.1 Context Format: Tabular versus Linearised . . . 35
3.10.2 Judge-Prompt Language . . . 36
3.11 Documented Failure Mode: gemma-4-26b-a4b@4bit . . . 38
3.12 The Answers as Climate and Energy Analysis . . . 39
3.13 Summary of Findings . . . 43

4 Discussion 45
4.1 RQ1: Alignment with an Expert-Derived Reference . . . 45
4.2 RQ2: How Does Retrieval Architecture Affect Quality? . . . 46
4.3 RQ3: Validity of Evaluation Methods . . . 48
4.4 The gemma-4-26b-a4b@4bit Failure Case . . . 49
4.5 Limitations . . . 50
4.6 Ethical and Sustainability Considerations . . . 51

5 Implications 54
5.1 Implications for Policy Investigators . . . 54
5.1.1 What the Technology Supports Today . . . 54
5.1.2 Corpus Design Decisions That Determine Output Quality . . . 54
5.1.3 Evaluating and Procuring Policy AI . . . 55
5.2 Implications for the Climate and Energy Community . . . 55

6 Conclusions & Future Work 58
6.1 Conclusions . . . 58
6.2 Future Work . . . 59

Bibliography 61

List of Figures

3.1 Cross-model condition means for the final run. From left: composite judge score (`medel`), ROUGE-L F1, BERTScore F1, and citation traceability. Error bars show ± 1 standard deviation across all scored cells. The composite, ROUGE-L and BERTScore panels are near-flat across conditions; only citation traceability separates no-RAG (zero by construction) from the three RAG conditions. 22

3.2 Condition effect relative to the no-RAG baseline, broken down by benchmark question and averaged across analyst models. Positive bars indicate that a RAG condition outperformed the same model's no-RAG answer; the sign is mixed across questions. 23

3.3 Left: per-cell agreement between Judge A (`microsoft/phi-4`) and Judge B (`nvidia/nemotron-3-nano-4b`), Spearman $\rho = 0.37$ ($n = 279$); points coloured by condition. Right: pooled condition effect relative to no-RAG from a linear mixed-effects model with random intercepts for model and question. All three RAG conditions sit below the no-RAG baseline; none is statistically significant. 30

3.4 Per-condition human MaxDiff utility (horizontal) against automated LLM-judge mean score (vertical), with regression line. Both axes use the 69 items with complete judge scores, so the judge means differ from the full-run values in Table 3.1 in both level and ordering, and the human utilities differ slightly from the 70-item values in Section 3.8.1, with the same ordering. That the ordering itself shifts between the full run and this subset is consistent with the resolution argument of Section 3.2.2. The negative slope is a four-point fit and is descriptive only. 31

3.5 Agreement between the human rater and the automated judge on best pick, worst pick, and implied ordered pairs, with 95% confidence intervals. The dashed line marks the 25% chance baseline for the four-option best/worst picks. 32

3.6 Quality against resource cost. Left: mean composite judge score against mean wall-clock time per answer, with the feedback tier plotted at its full standalone cost, the agentic draft plus the refinement increment (Table 3.8). Right: the same scores against the mean number of retrieved chunks placed in context. No-RAG sits furthest left on both panels and, for the capable models, also highest. 34

3.7 Left: per-condition mean composite under the Swedish and English judge rubrics. Right: mean per-dimension change (English minus Swedish), pooled over both judges. 38

List of Tables

3.1 Mean composite judge score per model and condition. Means are computed across the ten benchmark questions. Standard deviations are reported in parentheses. The `gemma-4-26b-a4b@4bit` no-RAG cell reflects nine observations (one judge failure; see Section 3.11); all other cells reflect ten. 21

3.2 Friedman test per analyst model (four conditions, ten benchmark questions). W denotes Kendall's coefficient of concordance. n is the number of complete blocks. The "Sig. pairs" column lists the Bonferroni-corrected Wilcoxon comparisons reaching $p_{\text{adj}} < 0.05$; all favour no-RAG. 24

3.3 Cross-model mean composite score per benchmark question and condition. The condition with the highest mean for each question is bold-faced. 25

3.4 Cross-model mean of each judge dimension by condition (scale 1–10), averaged over both judges. All seven analyst models are included. 26

3.5 Mean citation traceability by model and condition, with the mean number of citation attempts per answer given in parentheses. An attempt is any reference the model emits, valid or invalid [Källa N] or Harvard-style. A traceability of 1.00 indicates that every attempt resolves to an actually-provided chunk. The ratio is undefined where a model makes no attempt at all and is reported as zero in that case, which is why the attempt counts are needed to read the zeros. By construction, the no-RAG column is zero for all models. 27

3.6 Cross-model mean overlap metrics by condition, computed on citation-stripped answers against the gold standard. 28

3.7 Per-model agreement between the human rater's best/worst picks and the automated judge (n items per model). Preliminary, single rater. 32

3.8 Resource cost per answer by retrieval condition, six capable models (excluding `gemma-4-26b-a4b@4bit`; see Section 3.11). The feedback row gives the standalone cost of that tier, being the agentic draft (191.1 s) plus the refinement increment (139.8 s). Energy and emissions assume a mean draw of 70 W and a grid intensity of 21.3 gCO_2e/kWh, and cover operational energy only. 35

3.9 Exact-figure extraction accuracy (%) by presentation format, for eight numeric probes drawn from three tables in the held-out report. The penalty column is prose accuracy minus table accuracy, in percentage points; positive values would indicate that tabular formatting costs accuracy. 36

3.10 Per-condition mean composite score under the Swedish and English judge rubrics, pooled over both judges on the 62 paired units. Ranks are within column. 37

3.11 Analytical propositions in the answers to the four energy and land-use questions. Ref. gives the position of the reference answer: • advances the proposition, × rejects or contradicts it, ○ does not address it. Models gives the number of the seven human-preferred answers that advance the proposition, with the number that explicitly reject it in parentheses where non-zero. Restating the question without taking a position does not count as advancing a proposition, so `gemma-4-26b-a4b@4bit` registers on none of the fourteen rows (Section 3.11). 40

C.1 Analyst panel (the experimental subject). "Active" is the number of parameters engaged per token for mixture-of-experts (MoE) models. Quantisation is reported as encoded in the LM Studio model identifier; where the identifier carries no precision suffix the model ran at the default precision of the published build, which was not recorded at run time. The mixture of precisions is a source of non-comparability between rows and is noted in Section 4.5. The helper, judge, and gold-standard models are held distinct from this panel. 7

C.2 Pipeline and run parameters. All inference ran locally through LM Studio on Apple Silicon. 7

C.3 BERTScore F1 by model and condition. Differences of this magnitude are not interpretable as differences in semantic fidelity; see Section 3.6. No condition is marked as leading, because the metric does not support such a reading. . . . 8

C.4 Mean wall-clock time per answer (seconds) by model and condition, as recorded. The feedback column is the *increment* over the agentic draft, which that tier reuses rather than regenerating; the standalone cost of a feedback answer is the agentic and feedback columns summed. Section 3.9 explains the distinction and Table 3.8 reports the standalone figures. 8

D.1 Per-condition scores for the worked example (gemma-4-31b-it@4bit, Q3). Each judge value is that judge's mean over the five dimensions; the composite is the mean of the two judges. The human rater's best and worst picks are shown in the final column. 9

E.1 Friedman omnibus test per analyst model (composite score, four conditions as repeated measures over ten questions). Post-hoc pairwise comparisons (Table E.2) were computed only where the omnibus test was significant. 18

E.2 Complete Bonferroni-corrected Wilcoxon signed-rank pairwise comparisons for the four models with a significant Friedman test (composite score). Δ is the mean paired difference (first minus second condition); r_{rb} is the matched-pairs rank-biserial effect size; p_{adj} is Bonferroni-corrected across the six pairs. Bold marks $p_{\text{adj}} < 0.05$. 19

E.3 Mean composite score per condition, by judge, across the 279 scored cells. The final column gives the condition mean after each judge's scores are standardised to zero mean and unit variance. 19

Chapter 1

Introduction

1.1 Background

In 2015, 195 countries adopted the Paris Agreement's goal of limiting global average temperature increase to well below 2 degrees Celsius above pre-industrial levels, aiming to keep it below 1.5 degrees to minimise the negative impact of global warming [1]. By 2023, the Intergovernmental Panel on Climate Change (IPCC) expected the 1.5-degree goal to be exceeded within the 21st century, while the 2-degree goal became increasingly difficult to meet, with each increment intensifying the impact [2]. With this, overshoot scenarios are being increasingly considered; however, they risk creating feedback loops that make returning temperatures to the stated goals infeasible [3].

In a major study assessing 1 500 separate policy interventions in 41 countries over two decades, only 63 were found to have a major impact, further stressing the urgent need for efficient and effective climate policy [4]. This scarcity of high-impact policy underscores both the difficulty of evaluation at scale and the value of tools that can support such analysis more systematically.

Artificial Intelligence (AI) is increasingly being integrated into public sector governance, and the Organisation for Economic Co-operation and Development (OECD) recommends the introduction of both Machine Learning (ML) and Large Language Models (LLMs) in the early stages of policy creation, among other areas [5]. Progress has been made in using ML to inform climate policy via data assessments concerning both climate impact and rate of change [6, 7]. Despite this, major gaps remain between how LLMs communicate climate change at a surface level compared with more technically demanding policy domains [8]. Pilot studies using LLMs for policy evaluation have nonetheless indicated substantial promise: one study achieved 85% accuracy compared to expert analysis [9].

The intersection of AI and climate policy therefore represents a critical research frontier. In particular, the reliability and design of LLM-based policy analysis tools, and the methods used to evaluate them, remain open questions that this thesis seeks to address.

1.2 Theory and State of the Art

1.2.1 Methodologies for Evaluating Climate Mitigation Policies

The evaluation of climate mitigation policy has traditionally relied on quantitative break-detection methods and expert-led qualitative review. A recent large-scale study evaluated 1 500 policies over a 24-year period across 41 countries, using statistical methods to identify significant deviations from national emissions trends attributable to policy action, drawing on the Climate Actions and Policies Measurement Framework (CAPMF) and Emissions Database for Global Atmospheric Research (EDGAR) datasets [4].

At a finer level, the CAPMF dataset was utilised by D'Arcangelo et al. for clustered analysis of national mitigation strategies, linking policy cluster characteristics to emission outcomes through regression [10]. More commonly, expert analysis consists of structured literature reviews, as in Wang et al.'s three-step classification of EU carbon neutrality instruments and their distributional effects [11].

The benchmark for the Stockholm Environment Institute (SEI) AI reader (see Section 1.2.2) was established through in-house policy expert analysis, whose judgements were then compared to the LLM's output [9]. This highlights how human expert analysis remains the current de facto gold standard for policy evaluation, and motivates the need for reliable automated alternatives.

1.2.2 Language-Based AI and Policy Evaluation

While traditional methods excel at identifying measurable emissions effects of policies in retrospect, climate policy is inherently textual and semantic, requiring the analysis of legislative intent, implementation mechanisms, and governance gaps before quantitative outcomes are realised.

The SEI AI reader, developed to support climate literature reviews, used a curated analytical framework and structured prompting to extract policy-relevant data from academic papers. Comparison against expert analysis yielded 85% accuracy and 69–90% reliability, indicating that LLMs can operate meaningfully in this domain [9].

Complementing these extraction-focused applications, Belete introduced an integrated framework combining Natural Language Processing (NLP) with Retrieval-Augmented Generation (RAG) to systematically evaluate global climate policies [12]. By transforming unstructured policy text into semantic vector embeddings via Sentence-BERT, the study moved beyond keyword matching to capture conceptual similarity across diverse legal frameworks. The analysis quantified a significant "implementation gap" in global governance: while nations exhibited

high semantic convergence on aspirational targets such as "net-zero," similarity scores dropped sharply when analysing enforcement mechanisms. Sentiment analysis further found that neutral, regulatory-specific language correlated more strongly with emissions reductions than aspirational framing.

Taken together, these studies demonstrate that LLMs can engage substantively with policy text. However, neither study systematically compared retrieval architectures, and both relied on limited or ad hoc evaluation protocols, leaving open the question of which retrieval design best supports expert-quality analysis and how such quality should be measured.

1.2.3 Evaluating AI-Generated Text

The choice of evaluation method for AI-generated text is consequential, as different metrics capture different aspects of quality and can diverge significantly in their conclusions.

N-gram and Surface Metrics

The earliest automated evaluation metrics compare candidate and reference text by surface-level overlap. Recall-Oriented Understudy for Gisting Evaluation (ROUGE), introduced by Lin [13], computes recall across n-gram matches between a candidate and one or more reference texts. ROUGE-L, based on the longest common subsequence, has become a standard metric for summarisation and question-answering tasks. Despite their widespread use, n-gram metrics are known to be insensitive to paraphrasing and semantic equivalence: a factually correct response that uses different vocabulary from the reference will score poorly even if it conveys the same information.

Semantic Embedding Metrics

BERTScore [14] addresses these limitations by computing cosine similarity between contextual token embeddings of the candidate and reference, derived from a pretrained language model. Because it compares representations rather than surface forms, BERTScore is sensitive to semantic equivalence and has been shown to correlate more strongly with human judgements of fluency and relevance than n-gram alternatives, particularly for open-ended generation tasks.

LLM-as-a-Judge

Zheng et al. [15] established through MT-Bench that capable LLMs can serve as scalable evaluators for open-ended generation tasks, achieving meaningful agreement with human expert assessment. Liu et al. extended this with G-Eval [16], a structured prompting approach in which the judge first reasons step-by-step through each evaluation criterion before assigning a score,

improving consistency over direct scoring. However, LLM judges exhibit documented biases: self-preference (inflating scores for outputs from the same model family by 1–2 points) [15], position bias (favouring answers presented first), and length bias (preferring longer responses regardless of content quality) [17]. Awareness and explicit mitigation of these biases is therefore an integral part of any LLM-as-a-judge evaluation protocol.

RAG-Specific Evaluation

The metrics above judge a generated text against a reference. Retrieval-augmented systems raise a further question, namely whether the generated claims are actually supported by what was retrieved, and a distinct family of frameworks has grown up to answer it. RAGAs decomposes RAG quality into faithfulness, answer relevance and context relevance, and scores them without reference answers [18]. ARES trains lightweight judge models for the same dimensions and calibrates them against a small human-annotated sample [19]. Both treat grounding as a property to be estimated by a model. The deterministic citation audit used in this thesis (Section 2.2.4) sits in the same family but takes the opposite approach: because the analyst model may cite only by integer labels drawn from a closed set of provided chunks, grounding can be verified exactly rather than estimated, at the cost of measuring only whether a citation resolves and not whether it supports the claim.

Human Evaluation

Human evaluation remains the ultimate benchmark against which automated metrics are validated. Expert raters provide the most ecologically valid assessment of policy-analytical quality, but introduce their own limitations: high cost, slow throughput, and susceptibility to inter-rater inconsistency. Inter-rater reliability (IRR), typically quantified using Spearman's ρ for ranked data or Cohen's κ for categorical agreement, is therefore reported alongside human evaluation results in rigorous studies to characterise the reliability of the assessment process itself [15].

1.2.4 AI Hallucination and Mitigation

Large language models generate text by extrapolating from statistical regularities in their training data rather than by consulting a verifiable record, and their factual reliability therefore varies with how well a given topic is represented in that data. Factual errors in LLM output are commonly referred to as hallucinations, and evaluating them requires qualitative validation methods distinct from those applicable to narrower supervised models.

One of the primary structural mitigations is RAG, in which query-based retrieval supplies the model with relevant context, enabling factually grounded responses. Basic RAG is simple to implement but relies on data quality, and additions such as pre- and post-retrieval optimisation

further increase reliability [20]. Prompt engineering provides a complementary mitigation: attributing retrieved passages to their sources has been shown to align model responses to the supplied evidence rather than its parametric knowledge, and enables output verification when citations are preserved [21].

Model maturity has also been argued to reduce hallucination frequency, although the evidence base remains largely observational; King's comparative commentary on GPT-based systems reports a marked improvement over earlier model generations but does not constitute a systematic evaluation [22]. Taken together with the architectural mitigations above, these developments have lowered, without eliminating, the barrier to deploying LLM-based advisory tools in specialised domains such as climate policy.

1.2.5 Context-Memory Conflict

Retrieval is usually presented as strictly additive: the model gains access to evidence it would otherwise lack. A substantial literature complicates that picture by studying what happens when retrieved context and parametric knowledge disagree, a situation termed context-memory conflict [23]. Longpre et al. showed that substituting entities in retrieved passages frequently fails to change model answers, indicating heavy reliance on parametric memory even when the supplied context contradicts it [24]. Later work finds the balance to be conditional rather than fixed: models are highly receptive to context presented as the sole coherent evidence, but exhibit strong confirmation bias towards their parametric knowledge when supportive and contradictory evidence appear together [23].

Two implications follow for the present study. First, injected context is not free. Where a model's parametric knowledge of a domain is already adequate, retrieved passages compete with it rather than supplement it, and the outcome depends on the relevance and coherence of what is retrieved. Second, the volume of context matters independently of its content: language models attend unevenly across long inputs, with material in the middle of a long context used markedly less reliably than material at either end [25]. Retrieval tiers that maximise coverage by inserting more passages therefore trade one failure mode for another.

This literature is directly relevant to the result reported in Chapter 3, and it is set out here so that the finding can be read against it rather than as a surprise. It does not, however, anticipate the specific question this thesis asks, which is whether an evaluation apparatus of the kind commonly used in this setting is capable of detecting such an effect at all.

1.2.6 Synthesis and Knowledge Gap

The literature demonstrates that LLMs can engage substantively with climate policy text, achieving accuracy rates comparable to human expert review in constrained settings. RAG is the

established mechanism for grounding these outputs in domain-specific evidence and reducing hallucination. Evaluation frameworks ranging from automated surface metrics to structured LLM judges have been developed to assess open-ended generation quality.

However, three critical gaps remain. First, existing applications of LLMs to policy analysis have not systematically compared retrieval architectures: studies tend to employ a single RAG configuration without ablating retrieval design choices. It is therefore unknown whether the quality gains associated with more sophisticated retrieval strategies such as hypothetical document expansion, agentic sub-query decomposition, or self-critique feedback loops hold consistently in the specific context of expert policy analysis. Second, the evaluation methods applied to LLM-generated policy analysis have been inconsistent and often limited to a single metric or a single evaluator type, making it difficult to draw robust conclusions about quality or to compare findings across studies.

Third, the evaluation instruments themselves are rarely validated against one another on the task at hand. Automated judges, overlap metrics and human raters are typically reported as though they measure the same underlying quantity, and where they are compared it is usually to establish that a judge is adequate rather than to ask what follows if the instruments disagree.

This study addresses these gaps together. By holding the corpus, model, and benchmark questions constant while varying the retrieval architecture across four conditions of increasing sophistication, it provides, to the author's knowledge, the first controlled ablation of RAG design choices for climate policy analysis in a Swedish context and language. By applying a multi-layered evaluation framework, combining automated text and semantic metrics, RAG-pipeline diagnostics, structured LLM-as-judge scoring, and a blinded single-rater human evaluation, it also examines which evaluation methods yield consistent and interpretable conclusions for this task type.

1.3 Aim

The aim of this work is to evaluate the capacity of open-weight LLMs, enhanced by increasingly sophisticated RAG architectures, to produce Swedish climate policy analysis that aligns with a reference derived from an expert body's published assessment. In doing so, the study simultaneously examines the validity and mutual consistency of automated and human evaluation methods as applied to AI-generated policy analysis.

1.4 Research Questions

1. To what extent can an open-weight LLM enhanced by RAG produce climate policy analysis that aligns with a reference derived from the published assessment of an expert body?

2. How does retrieval architecture, ranging from no retrieval to agentic multi-step retrieval with self-critique, affect the quality and accuracy of LLM-generated policy analysis?
3. Which of the following evaluation methods demonstrate the greatest validity and mutual consistency when assessing AI-generated policy analysis?
    - Automated text-overlap metrics
    - Semantic embedding metrics
    - RAG-pipeline diagnostics
    - LLM-as-judge scoring
    - Blinded human evaluation

RQ1 is deliberately framed as alignment with an expert-*derived* reference rather than as a match to expert assessment directly. The reference answers used here are generated by a large external model reading a held-out report authored by Klimatpolitiska rådet (Section 2.1.5); the underlying judgements are expert judgements, but the formulation against which candidates are scored is not. The distinction is material to how the RQ1 result may be interpreted and is returned to in the Discussion.

1.5 Delimitations

For the purpose of this work, no LLMs were trained or fine-tuned from scratch. All generative inference relied on pre-trained, state-of-the-art models accessed via a local Application Programming Interface (API), constrained and optimised through prompt engineering and RAG.

The study is geographically and linguistically scoped to Swedish climate policy documents and Swedish-language benchmark questions. Findings regarding retrieval architecture effectiveness and evaluation methodology consistency reflect this specific domain and should not be generalised to other policy areas or languages without further validation.

The qualitative evaluation is conducted on ten benchmark questions targeting distinct reasoning capabilities relevant to Swedish climate policy as of 2026. The question set was developed to probe specific analytical dimensions; conclusions drawn from the results should be interpreted within this scope.

Chapter 2

Methodology

This study employs a controlled experimental design to evaluate how closely an LLM enhanced by RAG can align with a reference derived from an expert body's published climate policy assessment, and whether the methods used to measure that alignment are themselves reliable. The experiment holds the corpus, model family, and benchmark questions constant while varying the retrieval architecture across four conditions of increasing sophistication (Section 2.1). A multi-layered evaluation framework (Section 2.2) then assesses the quality of each condition's outputs through complementary automated and human methods. Section 2.3 describes the two sensitivity checks that test whether the headline result depends on prompt design.

2.1 Qualitative LLM Framework

2.1.1 Corpus and Data Sources

The knowledge base is built from three publicly available Swedish climate policy document collections. In total, the corpus comprises 5 614 documents, of which 5 592 parsed successfully into the index (Section 2.1.2). Each collection is openly accessible from its respective public repository; the low-level retrieval mechanics are not reproduced here, but the sources, selection criteria, and composition are given below so that an equivalent corpus can be reassembled. The date of final retrieval from all sources was the 19th of March 2026.

- **Statens Offentliga Utredningar (SOU)**
  All 630 completed government inquiries from the Ministry of Climate and Enterprise (Klimat- och näringslivsdepartementet), publicly available through the Swedish Government's official channels [26]. These cover the formal policy-development process and constitute primary source material for Swedish climate legislation.

- **Klimatpolitiska rådet**
  36 documents from Sweden's independent Climate Policy Council [27]: 8 annual reports and 28 commissioned background papers, covering the council's publications up to, but excluding, the 2026 annual report.

- **DiVA**
  4 948 full-text research publications and agency reports from the DiVA research portal [28], providing Swedish and international empirical context for the policy documents. The selection profile is described below.

Whereas the SOU and Klimatpolitiska rådet collections were taken in full, DiVA indexes a far larger and more heterogeneous body of work and was therefore narrowed by a fixed selection profile. Records were drawn from DiVA's public full-text export interface for eight Swedish public-sector and research organisations active in climate and environmental policy: Naturvårdsverket, IVL Svenska Miljöinstitutet, SMHI, RISE, Trafikverket, VTI, Havs- och vattenmyndigheten, and Nordiska ministerrådet. A record was retained only if it had an available full text, carried a publication year of 1990 or later, was of type report, evaluation, research dataset, or peer-reviewed journal article, and contained at least one of the keywords *klimat*, *policy*, *omställning*, *utsläpp*, *styrmedel*, or *färdplan* in its title, abstract, subject terms, or affiliations. Of 33 626 unique records, 5 221 passed this profile and 5 114 were successfully downloaded as full text, of which 4 948 entered the corpus after deduplication and parse-quality filtering.

The corpus is therefore dominated by academic literature (DiVA), with the government inquiries (SOU) and the council material forming smaller but topically central collections. The Klimatpolitiska rådet annual report for 2026 was deliberately excluded from the indexed corpus and reserved as the evaluation gold standard (Section 2.1.5).

2.1.2 Document Preprocessing and Indexing

Raw PDFs were converted to structured Markdown using Docling [29], a document parser that preserves headings, tables, and mixed-quality scanned content.

The Markdown files were chunked by a purpose-built structural chunker that walks each document by heading hierarchy (levels H1–H4) and packs adjacent paragraphs under the same section towards a target of 900 characters, with a hard ceiling of 1 400 and a floor of 220 characters so that no header-only or stub chunks are produced; oversized paragraphs are split at sentence boundaries. Every chunk is prefixed with a breadcrumb naming the document and its section path, so that the embedding model and the downstream analyst model both receive section context. A corpus-wide pre-pass removes boilerplate paragraphs appearing verbatim in five or more distinct files, and tables of contents and reference lists are dropped by title pattern. This produced an indexed collection of 869 133 chunks in a local Qdrant vector database.

Semantic indexing used BAAI/bge-m3 [30] a multilingual dense retrieval model selected via the Massive Multilingual Text Embedding Benchmark (MMTEB) [31] and the Scandinavian Embedding Benchmark (SEB) [32], filtering for models evaluated on government and legal

domain text in both Swedish and English. The model produces 1 024-dimensional normalised embeddings and was run in half-precision on Apple Silicon hardware.

2.1.3 Language Model Configuration

All generative inference ran locally through LM Studio, which exposed an OpenAI-compatible API on `localhost`. Models were assigned to four distinct roles, kept separate so that each effect could be attributed cleanly.

- **Analyst panel (the experimental subject).** Seven open-weight models from three families (Gemma, Granite, Qwen), spanning both dense and mixture-of-experts architectures and ranging from roughly 4 B to 35 B total parameters. The full four-tier $\times$ ten-question protocol was run independently for each model. Using a panel rather than a single model addresses the well-documented sensitivity of LLM benchmarks to model choice and makes it possible to test whether the retrieval-tier ranking is stable across architectures; cross-model concordance is reported in Section 2.2.3.
- **Helper model.** A smaller, efficiency-optimised model (`google/gemma-4-e4b`) handled the computationally cheap subtasks: HyDE passage generation in Tier 2 and sub-query decomposition in Tiers 3–4. Holding this helper constant across the panel isolates retrieval-architecture effects from helper-model variance. One caveat attaches to this arrangement. The helper is drawn from the same checkpoint family as one member of the analyst panel (`gemma-4-e4b-it`), so for that model alone the helper generating HyDE passages and sub-queries is effectively the analyst itself. The isolation of retrieval-architecture effects from helper-model variance therefore holds strictly for six of the seven panel members, and results for `gemma-4-e4b-it` are reported with this in mind.
- **Judge panel.** Two LLM-as-judge models scored every answer independently: `microsoft/phi-4` (Judge A) and `nvidia/nemotron-3-nano-4b` (Judge B). Both were held strictly distinct from every analyst family (Gemma, Granite, Qwen) to prevent self-preference bias [15]; using two judges of different families and sizes also exposes the calibration variance a single judge would conceal. The composite score is the mean of the two judges' five-dimension averages.
- **Gold-standard generator.** A large external model (`llama-3.3-70b-instruct`), also outside the analyst panel, produced the reference answers (Section 2.1.5).

Two generation settings were held constant across all four conditions. The temperature was 0.3 for every policy-analysis response and 0.1 for the judge scoring pass. A shared Swedish-language system prompt instructed the analyst model to act as an independent climate policy analyst and to structure every answer in four sections: (1) Context, (2) Analysis, (3) Implications, and (4) Recommendations.

2.1.4 Experimental RAG Tiers

The experiment varies the retrieval architecture while holding the model, corpus, benchmark questions and prompts fixed, except that the no-RAG prompt forbids citations and the RAG prompts require them (Appendix B). The four tiers form a progressive ablation: each tier adds one architectural element to the preceding one, enabling attribution of performance differences to specific design choices.

- **Tier 1: No RAG (Baseline)**
  The model answers entirely from its pre-training knowledge with no retrieved context. It serves as the baseline for the retrieval tiers. Because the citation instruction changes together with retrieval, a difference between tiers 2–4 and this baseline reflects both (Section 4.2).

- **Tier 2: Single-Search RAG with HyDE**
  Standard single-shot retrieval is combined with Hypothetical Document Embeddings (HyDE) [33]. The helper model first generates a concise hypothetical answer to the benchmark question, which is then embedded and used as the retrieval query rather than the raw question text. Embedding a plausible answer improves recall in domain-specific corpora because the hypothetical answer occupies a closer position in the semantic space to relevant passages than the question itself does [20, 33]. The top 20 retrieved chunks are passed to the main model.

- **Tier 3: Agentic Librarian**
  The model acts as its own research librarian through iterative, retrieval-conditioned querying [34]. Rather than splitting the question in a single pass, the helper model issues up to five search queries one at a time; each new query is conditioned on the original question, the queries already issued, and a running digest of the chunks retrieved so far, so that it targets an aspect not yet well covered. Near-duplicate queries are detected and regenerated. Each query retrieves 10 chunks, deduplicated across queries, yielding a variable number of unique chunks for synthesis (2–39 across runs, 24 on average). This adaptive strategy increases coverage at the cost of retrieval precision, as each additional query introduces its own retrieval noise.

- **Tier 4: Feedback Loop**
  The Feedback Loop extends Tier 3 with a self-critique and targeted re-retrieval cycle based on the Reflexion framework [35]. The Tier 3 draft is passed back to the main model, which identifies up to four specific evidence gaps in its own response. A second retrieval pass fetches 10 new chunks per identified gap, deduplicated against the original pool. The main model then produces a refined final answer integrating the draft with the new evidence. To avoid redundant computation, the Tier 3 draft is passed directly to Tier

4 within the same run.

In the results, Tiers 1 to 4 are referred to as the no-RAG, single-RAG, agentic and feedback conditions; the abstract, Chapter 5 and Chapter 6 call them set-ups.

2.1.5 Benchmark Design

Ten Swedish-language benchmark questions were designed to test distinct reasoning capabilities relevant to expert policy analysis. Questions were developed in dialogue with domain experts to ensure they reflect the kinds of analytical judgements required in real policy review contexts and that their answers are non-trivial to retrieve without the specific gold-standard source.

- **Q1: Temporal and evaluative reasoning**
  Tests whether the model can bound its assessment to the 2022–2026 mandate period, draw on independent expert appraisal, and identify the principal shortcomings of the climate action plan presented in 2023.
- **Q2: Adversarial and normative judgement**
  Tests whether the model will reach and defend an explicit verdict on whether reducing the biofuel blending obligation (reduktionsplikten) was correct, and trace the second-order consequences, including investment freezes and bankruptcy risk in the biofuel sector, through to attainment of the national climate targets. The question is adversarial in that it demands an evaluative conclusion rather than a description.
- **Q3: Causal and mechanistic reasoning**
  Tests whether the model can explain how changed technical and economic conditions for CCS and BECCS in Sweden have propagated into revision of the national strategy for negative emissions, and enumerate the alternative carbon sinks under discussion.
- **Q4: Cross-domain synthesis**
  Tests whether the model can analyse how transport, energy and industrial policy reinforce or counteract one another in Sweden's pursuit of net-zero emissions by 2045, requiring identification of conflicts rather than parallel description of three sectors.
- **Q5: Multi-level governance**
  Tests whether the model understands the interaction between the EU climate framework (Fit for 55, CBAM, ESR) and Swedish national climate policy, including how EU regulation constrains national room for manoeuvre and shapes the transition work of municipalities.
- **Q6: Security-climate nexus**
  Tests whether the model can articulate specific synergies and conflicts of objective between the climate transition and the ongoing reinforcement of total defence, and issue

concrete recommendations for the next mandate period.

- **Q7: Land use and carbon sinks**
  Tests whether the model correctly characterises the state of the Swedish Land Use, Land-Use Change and Forestry (LULUCF) sector, assesses the prospects of meeting the EU commitment to 2030, and identifies the policy instruments required to secure a high net uptake of greenhouse gases in forest and land.
- **Q8: Sectoral governance (agriculture)**
  Tests whether the model can identify the central shortcomings of current Swedish agricultural policy from a climate perspective and formulate a strategy comprising target vision, policy instruments and enabling measures.
- **Q9: Energy system reasoning**
  Tests whether the model can produce an integrated analysis of the conditions for a fossil-free Swedish electricity system, covering the role of nuclear power, obstacles to wind expansion, the potential of energy efficiency, and grid capacity requirements, and rank the resulting priorities.
- **Q10: Cross-sectoral bioeconomy reasoning**
  Tests whether the model can motivate a national bioeconomy strategy and balance competition for biomass from forest and land against carbon uptake, biodiversity, biofuels, bio-based materials and carbon capture, requiring recognition of competing policy logics within a single sector.

The full Swedish wording of all ten questions is reproduced in Appendix A.

Gold-standard expected answers were derived from the Klimatpolitiska rådet annual report for 2026 [36], which was excluded from the indexed corpus prior to ingestion. The gold-standard answers were generated by a large external model (`llama-3.3-70b-instruct`), itself outside the analyst panel, which read the summary section of the excluded report and produced reference answers without access to any retrieved context. Using a large external model decouples the reference from the parametric knowledge of any analyst model and grounds it in the excluded source document. Correct answers in the experiment can therefore not be copied from the held-out report; they must come from the model's own knowledge or from reasoning over related evidence already present in the database.

One property of this procedure bounds what the comparison can establish. The reference model was given the report's summary section (Sammanfattning) rather than the full report, so the gold answers reflect the level of detail available in that section. Where a benchmark question addresses a topic treated only briefly in the summary, the corresponding reference is correspondingly thin, and a candidate answer that supplies additional correct detail may be scored no higher than one that does not. The ten gold answers average 484 words (range 420–572).

This limitation is revisited in Section 4.5.

2.2 Evaluation Framework

Each of the $7 \times 4 \times 10 = 280$ model–condition–question outputs is evaluated against its gold-standard answer using four complementary approaches. The multi-layered design serves two purposes: it provides convergent evidence of output quality when metrics agree, and it surfaces methodological divergence, answering RQ3, when they do not.

2.2.1 Automated Text and Semantic Metrics

Two automated overlap metrics compare each candidate answer against its gold standard. Both are computed on citation-stripped text: markers of the forms `[Källa X]` and `(Author, Year)` are removed before scoring, so that differing referencing styles do not create scoring artefacts.

The longest-common-subsequence variant of Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) [13] is computed to provide a sequence-sensitive recall metric that rewards order-preserving coverage of the gold standard. ROUGE-L is a standard benchmark in NLP evaluation, enabling external comparability of results.

To complement the surface-level metrics above, BERTScore [14] is computed using the multilingual `xlm-roberta-large` encoder (layer 17), without baseline rescaling. For each token in the candidate, a maximum cosine similarity is computed against all tokens in the gold-standard embedding; the final score is the mean of these maximum similarities. Because BERTScore compares contextual representations rather than surface forms, it captures semantic equivalence that ROUGE-L misses when the model paraphrases the reference content accurately. A strong multilingual encoder is used so that semantically equivalent Swedish paraphrases are scored as similar even when their surface forms differ.

2.2.2 LLM-as-Judge Scoring

Following the G-Eval structured prompting approach [16], each of the two judge models is instructed to reason step-by-step through each evaluation criterion before assigning a score, improving consistency over direct point-assignment. Each judge scores every candidate answer on five dimensions on a 1–10 scale, providing a one-sentence rationale per score; the composite score is the mean of the two judges' dimension averages:

- **Coverage (täckning):** The extent to which the answer addresses all key aspects present in the gold standard.

- **Factual accuracy (korrekthet):** The degree to which the answer's factual claims agree with the gold standard.
- **Argumentation (argumentation):** The logical coherence of the reasoning chain and whether claims are supported step-wise by the answer's own argument structure. This dimension assesses the quality of reasoning, not the presence of citation markers; citation traceability is measured separately and deterministically (Section 2.2.4).
- **Structure (struktur):** Whether the answer is logically organised and easy to follow.
- **Policy-analytical depth (policyinsikt):** The level of nuance and domain-relevant insight demonstrated beyond surface-level description.

Three known biases inherent to LLM judges are explicitly mitigated in the experimental protocol. *Self-preference bias*, the tendency for a model to inflate scores for outputs from its own model family [15], is mitigated by using two judge models, both from families (`phi` and `nemotron`) not represented in the analyst panel, and reporting their mean. Their pairwise agreement is itself reported (Section 2.2.3) as a measure of judge reliability. *Position bias*, the tendency to favour responses presented first, is mitigated by presenting a single candidate at a time rather than a paired comparison. *Length bias*, the tendency to award higher scores to longer responses regardless of quality [17], is examined post-hoc by correlating judge scores against candidate word counts to determine whether length is a confound in this setting.

2.2.3 Statistical Analysis

Differences across the four retrieval conditions are tested non-parametrically, as judge scores are ordinal (1–10) rather than continuous interval data. Because the four conditions are repeated measurements of the same items, tests for related samples are used throughout.

A Friedman χ^2 test [37] is applied simultaneously across all four conditions for each scored dimension, serving as the omnibus test of whether any significant differences exist among tiers. Operating on within-item ranks, it makes no assumptions about the shape of the underlying distributions.

Where the Friedman test is significant, pairwise post-hoc comparisons are conducted using the Wilcoxon signed-rank test [38], the non-parametric analogue of the paired t-test. Because all $\binom{4}{2} = 6$ pairwise comparisons inflate the risk of a false positive, a Bonferroni correction [39] is applied: each comparison is evaluated against an adjusted threshold of $\alpha/6$, controlling the family-wise error rate and making any reported difference more robust [40].

Effect sizes are reported as the matched-pairs rank-biserial correlation r_{rb} for each Wilcoxon comparison [41], computed as the difference between the proportion of favourable and unfavourable signed ranks. This statistic is bounded on $[-1, 1]$, is directly interpretable as the net

proportion of question-level comparisons favouring one condition, and is preferred here over $r = Z/\sqrt{N}$ because the block size (ten questions) is small enough that the normal approximation underlying Z is unreliable. Values are interpreted following Cohen's benchmarks of small ($r \approx 0.10$), medium ($r \approx 0.30$), and large ($r \approx 0.50$) effects [42].

This procedure is applied separately for each of the five judge dimensions and for the composite mean judge score, providing both dimension-level and aggregate statistical conclusions.

Two limitations of this testing procedure should be stated in advance. First, the Bonferroni correction is applied to the six pairwise comparisons within each analyst model but not across the seven models, so the family-wise error rate is controlled within model and not across the panel; reported per-model significances should be read accordingly. Second, with ten blocks per test the smallest attainable Bonferroni-corrected p-value is 0.012, so the procedure operates close to its resolution limit and cannot distinguish a large effect from a very large one. The pooled mixed-effects model described below is reported alongside precisely because it does not inherit these constraints.

To pool the evidence across models and questions without treating cells as independent, a linear mixed-effects model is additionally fitted to the composite score with condition as a fixed effect and crossed random intercepts for analyst model and question. Two reliability diagnostics accompany the judge scores: the agreement between the two judges (Spearman's ρ on per-cell composite scores), and a length-bias check correlating the composite score against candidate word count to verify that the judges are not simply rewarding longer answers.

2.2.4 Citation Traceability

Beyond the evaluation streams above, citation traceability is computed deterministically from the saved record of chunks shown to the model. For every retrieval-augmented condition, each *[Källa N]* reference in the candidate answer is parsed and classified against the closed set of N labels actually presented to the model:

1. Valid traceable references: the integer N matches a chunk shown to the model for this question;
2. Invalid out-of-universe references: the integer N matches no provided chunk and therefore cannot resolve to evidence;
3. Forbidden parenthetical references: Harvard-style "(Author, Year)" parentheses, which the citation contract disallows in both the no-RAG and RAG analyst prompts.

A traceability ratio is reported as the share of citation attempts that resolve to a valid chunk; a citation density is reported as valid references per sentence. Because the analyst model receives the candidate chunks together with their integer labels and is instructed to cite only by these la-

bels, every citation attempt can be verified post-hoc against the saved chunks-used record without re-querying the database. This separates the empirical question "are the model's citations grounded?" from the LLM-judge's qualitative assessment of argument quality, and addresses a known confound in MT-Bench-style evaluation in which judges cannot reliably distinguish fabricated citations from genuine ones.

2.2.5 Human Evaluation and Judge Calibration

A blinded human evaluation protocol was specified to validate the automated judge and to provide the ecological validity that automated metrics cannot offer. It was completed by a single rater, the author, working blind to the condition labels and without access to the reference answers. The schedule between the final run and submission did not allow independent raters to be recruited for a protocol of this size, so inter-rater reliability is not reported anywhere in this thesis. The consequences are set out in Section 4.5, and the human results are presented throughout as preliminary.

Two measures separate the rater from the experimental condition. First, the four conditions are randomly re-labelled A–D independently for each question, so that neither the tier nor its position in the presented order is recoverable across items. Second, citation markers are removed from the candidate answers before they are presented, using the same routine that strips citations for the automated overlap metrics (Section 2.2.1). This second measure matters because the citation contract differs by condition: the no-RAG prompt forbids citations while the RAG prompts require them (Appendix B), so unstripped answers would identify the no-RAG arm to the rater on sight.

The stripping is not complete, and the residual leak is quantified here rather than claimed against. The routine matches the single-integer form [Källa N] but not the comma-separated form [Källa N, M], and it does not match bare-year parentheticals of the form "(2023)". Of approximately 2 200 citation attempts across the 210 retrieval-augmented answers, 96% were removed, and 13 of the 70 presented items retained at least one visible marker in at least one answer: seven items through the comma-separated form and six through bare years. Every residual marker falls in a retrieval-augmented arm, so on those 13 items the condition was in principle recoverable. A sensitivity analysis re-computing the human condition ranking on the 57 unaffected items is reported in Section 3.8.1.

The rater selects the best and worst response per question (a MaxDiff, or best/worst scaling, design) and may flag unusable answers. Best/worst choices are converted into a per-condition utility as the count of best selections minus the count of worst selections, normalised by the number of items in which the condition appeared, so that the utility is bounded on $[-1, 1]$ and zero denotes indifference. The resulting condition ranking is compared against the automated judge.

One channel of information cannot be removed by stripping. Retrieval-grounded answers name specific instruments, programmes and figures drawn from the retrieved chunks, whereas unretrieved answers tend towards generic formulations of the same content, so a rater with domain familiarity may infer retrieval status from specificity alone even with all markers removed. The protocol removes the mechanical tell, not the substantive one. Whether this constitutes a confound or is instead part of the quality signal the rater ought to respond to is itself an open question, and it is returned to in Section 4.5.

Judge calibration is assessed by comparing the rater's choices with the automated judge at two levels: over the four condition aggregates (Spearman's ρ and Kendall's τ), and at the item level, as agreement between the rater's best and worst picks and the judge's highest-scored and lowest-scored conditions, together with implied pairwise agreement. High agreement would indicate that the automated judge is a valid proxy for the human rater's judgement in this task; low agreement indicates that automated and human notions of quality diverge, which is itself a substantive finding relevant to RQ3.

2.3 Prompt Sensitivity Analysis

The methodology described so far fixes the frameworks and the models, but every result reported here is also conditioned on the prompts that drive them: the wording of the judge rubric, and the form in which source material is placed in the context window. Within the time available these could not be varied exhaustively, so a smaller sensitivity analysis was run instead. It is indicative rather than complete. It establishes whether the headline result is plausibly an artefact of prompt design, but not the full shape of the prompt–performance surface. Both checks reuse the cached generations, gold standard and judge scores from the main batch, so only genuinely new model calls were issued.

2.3.1 Context Format: Tabular versus Linearised

Several of the quantitative facts in the corpus and in the held-out 2026 report are carried in Markdown tables rather than prose: the ESR emissions-gap table, the transport-scenario table, and the LULUCF/FRL bookkeeping table. If a generation model cannot read a figure out of a table it has been shown, then any answer depending on that figure is degraded regardless of retrieval quality. Three real tables from the report were used to construct eight single-number probe questions whose answer appears *only* in the table. Each probe was asked of all seven analyst models under three presentations of the same facts:

- the table exactly as the chunker preserves it, which is the real condition;
- the same rows flattened into Swedish sentences, an upper bound on what perfect table-to-prose conversion would achieve;

- a short prose stub with the table removed, a floor, since the fact is then absent from the context entirely.

Decoding was at temperature 0, giving $7 \times 8 \times 3 = 168$ calls. The quantity of interest is the gap between the second presentation and the first, read as the penalty a model pays for tabular formatting. A high score on the third would instead indicate that a model is answering from parametric memory rather than from the supplied context.

Scoring is deterministic. The emitted answer must contain the exact gold figure, with Swedish decimal commas and decimal points treated as equivalent. A naive substring test is not sufficient, however. Several of the analyst models reason step by step before answering and quote the supplied context back inside that reasoning, so the gold figure can appear in the output even when the model never commits to an answer. Two measures address this: the scorer removes table rows and verbatim context echoes from the output before matching, and the token budget is set high enough for a reasoning model to finish and answer rather than being cut off mid-trace. Where the budget is adequate the strict and naive scorers agree, and that agreement is reported as a check that the run is clean.

2.3.2 Judge-Prompt Language

The dual judge scores Swedish answers using a Swedish rubric (Section 2.2.2). If the same rubric translated into English moves the scores, or reorders the four conditions, then the judge is sensitive to prompt language and part of the reported effect is an artefact of that choice. A faithful English translation of the judge instructions was written, preserving the five dimensions, the 1–10 scale, the instruction to ignore citation markers, and the Swedish output keys, so that only the instruction language varies. A random sample of 62 of the 279 scored $(\text{model}, \text{question}, \text{condition})$ units was re-scored by both judges under the English rubric and compared to the cached Swedish scores for the same units. Candidate answers, gold standard, sampling and decoding parameters were held constant. Three comparisons are reported: a paired Wilcoxon signed-rank test on the composite score per judge, the mean per-dimension delta, and Spearman's ρ between the per-condition mean composites under the two languages, which tests whether the condition ranking survives.

Chapter 3

Results

This chapter reports the empirical results of the qualitative LLM experiment. All numbers refer to the consolidated final run, in which seven analyst models each generated answers to the ten benchmark questions under all four retrieval conditions, and every answer was scored by a dual independent judge panel (`microsoft/phi-4` and `nvidia/nemotron-3-nano-4b`) against a gold standard generated by `llama-3.3-70b-instruct`. The findings are presented strictly descriptively, with the exception of Section 3.12, which reads a subset of the answers as climate and energy analysis and compares their conclusions with those of the reference; interpretation, comparison to prior work, and assessment of the research questions are reserved for the Discussion chapter (Chapter 4).

The presentation proceeds from the aggregate to the granular. Section 3.1 summarises the experimental sample. Section 3.2 reports condition-level performance per model and across models, together with the within-model and pooled statistical tests. Section 3.3 examines variation across the ten questions, and Section 3.4 decomposes the composite judge score into its five constituent dimensions. Section 3.5 reports the deterministic citation audit, Section 3.6 the lexical and semantic overlap with the gold standard, and Section 3.7 the agreement between the two automated judges. Section 3.8 reports the preliminary human evaluation and its agreement with the automated judge. Section 3.9 reports wall-clock cost and the associated energy use, Section 3.10 the prompt-sensitivity checks, and Section 3.11 documents the single model that did not satisfy the floor of usable performance. Section 3.12 then reads the human-preferred answers to four questions as climate and energy analysis. A condensed summary of findings closes the chapter (Section 3.13).

3.1 Overview of Experimental Runs

The full design contains $7 \times 4 \times 10 = 280$ answer cells. All 280 were successfully generated; 279 were scored by both judges and enter the analysis. The single excluded cell (`gemma-4-26b-a4b@4bit`, `Q5_multilevel`, no-RAG) failed at the judge stage rather than at generation and is reported as missing data. Unlike earlier pilot runs, no generation-time failures occurred in the final run; the residual degradation of the `gemma-4-26b-a4b@4bit` model is documented

in Section 3.11.

Each answer was scored independently by two judge models of different families and sizes, `microsoft/phi-4` (Judge A) and `nvidia/nemotron-3-nano-4b` (Judge B), and the composite score reported throughout is the mean of the two judges' five-dimension averages, on the 1–10 scale. It is referred to as the composite in the text and labelled `medel` in the figures, following the pipeline's Swedish column names; the two terms denote the same quantity. Both judges were held strictly distinct from every analyst model in the panel (Section 2.2.2). The gold-standard reference answers were generated by a large external model (`llama-3.3-70b-instruct`), itself outside the analyst panel, to decouple the reference from the parametric knowledge of any analyst model.

Across the 279 scored cells the mean answer length is 437 words. The mean wall-clock time per answer is distributed unevenly across conditions (no-RAG fastest for all six capable models; see Section 3.9 and Figure 3.6). A formal length-bias check (Section 3.7) finds no association between answer length and judge score (Spearman $\rho = 0.04$, $p = 0.52$).

3.2 Aggregate Performance Across Conditions

3.2.1 Per-Model Condition Means

Table 3.1 reports the mean composite judge score (mean of the two judges' five-dimension averages, scale 1–10) for each model in each condition. Figure 3.1 displays the cross-model condition means alongside the text and citation metrics.

Table 3.1: Mean composite judge score per model and condition. Means are computed across the ten benchmark questions. Standard deviations are reported in parentheses. The `gemma-4-26b-a4b@4bit` no-RAG cell reflects nine observations (one judge failure; see Section 3.11); all other cells reflect ten.

Model	No-RAG	Single-RAG	Agentic	Feedback
gemma-4-e4b-it	8.23 (0.25)	7.32 (0.77)	7.44 (0.49)	7.38 (0.55)
granite-4.1-8b	7.51 (0.98)	8.24 (0.58)	8.12 (0.40)	8.07 (0.45)
gemma-4-31b-it@4bit	8.28 (0.41)	7.40 (0.97)	7.74 (0.55)	7.65 (0.24)
granite-4.1-30b	8.29 (0.29)	8.16 (0.55)	7.98 (0.35)	7.87 (0.43)
qwen3.6-27b-ud-mlx	8.28 (0.40)	8.08 (0.36)	8.00 (0.74)	7.82 (0.64)
qwen3.6-35b-a3b-ud-mlx	8.30 (0.31)	7.19 (1.91)	7.76 (0.83)	8.01 (0.43)
gemma-4-26b-a4b@4bit	5.16 (1.51)	6.75 (1.26)	5.95 (1.94)	6.91 (1.16)
Cross-model mean	*7.76*	*7.59*	*7.57*	*7.67*
Cross-model mean (excl. 26b@4bit)	*8.15*	*7.73*	*7.84*	*7.80*

Contrary to the expectation that retrieval would uniformly raise quality, no-RAG attains the highest composite score in five of the seven models (`gemma-4-e4b-it`, `gemma-4-31b-it@4bit`, `granite-4.1-30b`, `qwen3.6-27b-ud-mlx`, and `qwen3.6-35b-a3b-ud-mlx`).

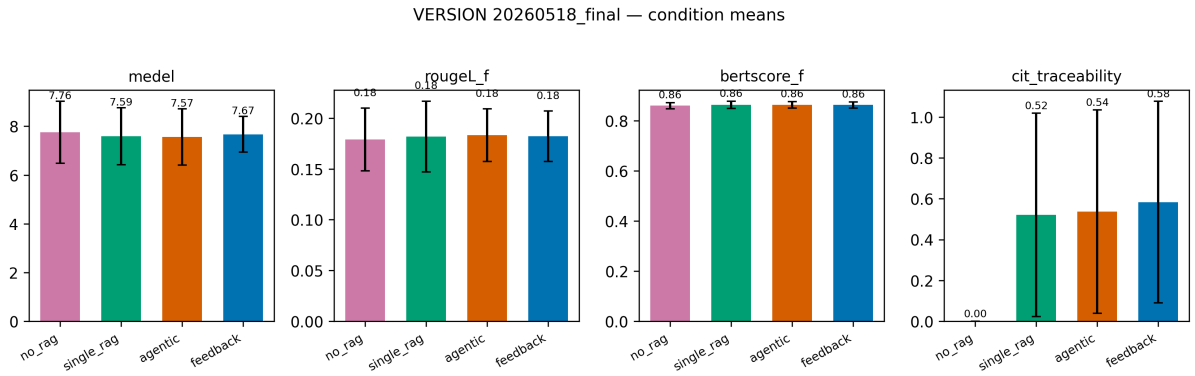

Figure 3.1: Cross-model condition means for the final run. From left: composite judge score (`medel`), ROUGE-L F1, BERTScore F1, and citation traceability. Error bars show ± 1 standard deviation across all scored cells. The composite, ROUGE-L and BERTScore panels are near-flat across conditions; only citation traceability separates no-RAG (zero by construction) from the three RAG conditions.

Only the two models with the weakest parametric baseline score highest under a RAG condition: `granite-4.1-8b` (single-RAG, $+0.73$ over its no-RAG baseline) and the failure-mode `gemma-4-26b-a4b@4bit` (feedback, $+1.75$). For the five capable models that peak without retrieval, the best RAG condition trails no-RAG by between 0.13 and 0.79 points.

Read as a frequency-of-success rate rather than a mean, 87% of the 279 scored outputs reach a composite of at least 7 on the 1–10 scale, rising to 93% for the six models above the failure case (`gemma-4-26b-a4b@4bit`); at the stricter threshold of 8 the overall rate falls to 49% (55% for those six). The stricter threshold also separates the conditions: no-RAG places 67% of its answers at 8 or above, against 39–47% for the three RAG tiers, consistent with its advantage on the composite mean. These rates are reported as internal descriptive statistics only. They are not comparable with the 85% expert agreement reported for the SEI policy reader [9], which measured agreement against expert labels on a different task, and no comparison with that figure is drawn here.

3.2.2 Cross-Model Condition Ranking

Pooling all scored cells, the cross-model condition means rank no-RAG (7.76), feedback (7.67), single-RAG (7.59), and agentic (7.57), a spread of only 0.19 points. Excluding the `gemma-4-26b-a4b@4bit` cells widens no-RAG's lead: no-RAG (8.15), agentic (7.84), feedback (7.80), single-RAG (7.73).

To assess agreement on the condition ranking when each model is treated as a single rater, Kendall's W was computed across the condition ranks of the seven models. The result is $W = 0.118$, indicating low agreement. With seven models ranking four conditions this coefficient is itself very noisy, so it is reported descriptively: low W here is as consistent with sampling noise around a small common effect as with genuine model-dependence, and it should

not be read as evidence for the latter. The corresponding mean ranks across models are 1.86 (no-RAG), 2.57 (agentic), 2.71 (single-RAG), and 2.86 (feedback), so no-RAG holds the best mean rank. First-place wins are distributed five times to no-RAG, once each to single-RAG and feedback, and never to agentic. The per-question breakdown of the condition effect is shown in Figure 3.2.

The stability of this ranking should be read with one further result in view. The prompt-sensitivity analysis of Section 3.10.2 re-scores a random subsample of 62 of the 279 scored units, and the cached Swedish scores for that subsample order the conditions differently: feedback (7.92), single-RAG (7.89), no-RAG (7.59), agentic (7.53), placing no-RAG third rather than first. The subsample was drawn without stratification across model and condition and is unbalanced, containing 12 no-RAG cells against 19 agentic cells. More consequentially, `qwen3.6-35b-a3b-ud-mlx`, the model with the largest no-RAG advantage in Table 3.1, contributes no no-RAG cell at all. Re-weighting the subsample so that each model contributes equally to each condition mean does not restore the ordering (no-RAG 7.61, still third), so composition explains part but not all of the shift. Either way the observation reinforces rather than undermines the reading offered here: a nominal ordering that does not survive the removal of four fifths of the data is not an ordering the instrument can resolve, and the substantive conclusion of this chapter is that the conditions are separated by less than the evaluation can measure.

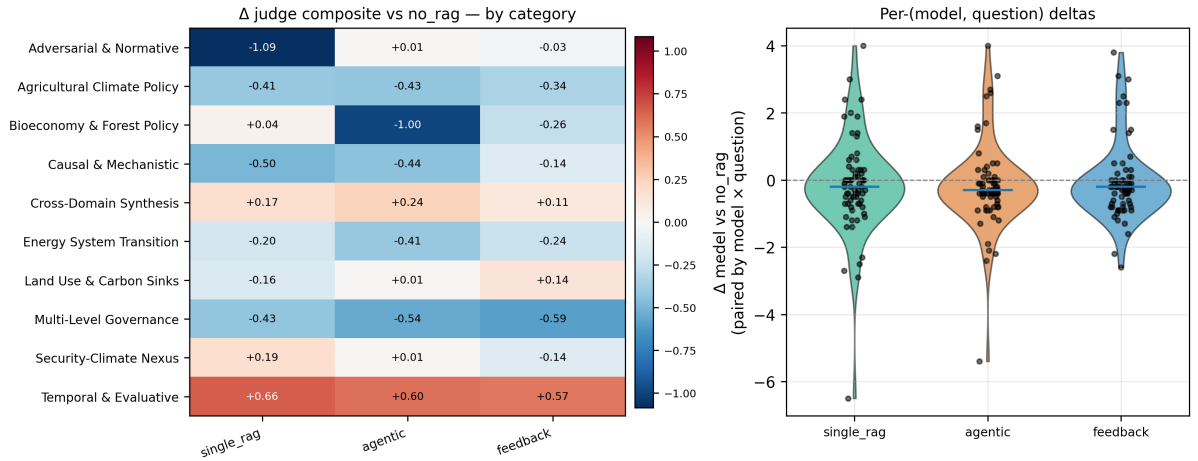

Figure 3.2: Condition effect relative to the no-RAG baseline, broken down by benchmark question and averaged across analyst models. Positive bars indicate that a RAG condition outperformed the same model's no-RAG answer; the sign is mixed across questions.

3.2.3 Within-Model and Pooled Significance Tests

A Friedman χ^2 test was computed for each analyst model with the four conditions as repeated measures across the ten questions (Table 3.2). Four of the seven models produce a test statistic that crosses the conventional $\alpha = 0.05$ threshold: `gemma-4-e4b-it` ($p < 0.001$), `gemma-4-`

`31b-it@4bit` ($p = 0.001$), `qwen3.6-35b-a3b-ud-mlx` ($p = 0.027$), and `granite-4.1-30b` ($p = 0.018$).

Critically, in *every* significant Wilcoxon signed-rank post-hoc comparison (Bonferroni-corrected across the six pairs), the direction of the effect favours no-RAG over the RAG condition: no-RAG significantly exceeds single-RAG, agentic and feedback for `gemma-4-e4b-it`; no-RAG exceeds single-RAG and feedback for `gemma-4-31b-it@4bit`; no-RAG exceeds feedback for `granite-4.1-30b`; and no-RAG exceeds agentic for `qwen3.6-35b-a3b-ud-mlx`. No RAG condition significantly exceeds no-RAG for any model.

Table 3.2: Friedman test per analyst model (four conditions, ten benchmark questions). W denotes Kendall's coefficient of concordance. n is the number of complete blocks. The "Sig. pairs" column lists the Bonferroni-corrected Wilcoxon comparisons reaching $p_{\text{adj}} < 0.05$; all favour no-RAG.

Model	χ^2	p	W	Sig. pairs ($p_{\text{adj}} < 0.05$)	n
gemma-4-e4b-it	19.07	< 0.001	0.636	no-RAG > single, agentic, feedback	10
granite-4.1-8b	6.04	0.110	0.201	none	10
gemma-4-31b-it@4bit	15.87	0.001	0.529	no-RAG > single, feedback	10
granite-4.1-30b	10.04	0.018	0.335	no-RAG > feedback	10
qwen3.6-27b-ud-mlx	6.73	0.081	0.224	none	10
qwen3.6-35b-a3b-ud-mlx	9.19	0.027	0.306	no-RAG > agentic	10
gemma-4-26b-a4b@4bit	4.33	0.228	0.160	none	9

To pool the evidence across models and questions while accounting for their structure, a linear mixed-effects model was fitted to the composite score with condition as a fixed effect and crossed random intercepts for model and question. The estimated condition effects relative to no-RAG are all negative (single-RAG, agentic and feedback each sit below the no-RAG baseline), and none is statistically significant at the pooled level (e.g. feedback $p = 0.595$, agentic $p = 0.217$). The pooled estimates are shown in the right panel of Figure 3.3.

Two limits of this specification should be noted. Random intercepts allow each model and question its own baseline but constrain the condition effect to be common across them, so the model tests only the average effect and cannot itself establish that the effect varies by model. The apparent variation visible in Table 3.1 is therefore described here rather than tested; a random slope for condition by model, or an explicit model by condition interaction, would be required to test it and is not fitted. Separately, judge scores are treated as ordinal for the non-parametric tests above (Section 2.2.3) but as interval here, since the composite is a mean over five dimensions and two judges. Both treatments are defensible in isolation and the tension between them is noted rather than resolved; the pooled model is reported as a complement to the rank-based tests, not as a replacement.

The combined picture is that the condition effect is small, that it is not distinguishable from zero when pooled, and that where within-model tests do reach significance they run against retrieval rather than for it.

3.3 Per-Question Performance

Table 3.3 reports the cross-model mean composite score for each of the ten benchmark questions in each condition.

Table 3.3: Cross-model mean composite score per benchmark question and condition. The condition with the highest mean for each question is bold-faced.

Question	No-RAG	Single-RAG	Agentic	Feedback
Q1_temporal	7.24	**7.90**	7.84	7.81
Q2_adversarial	7.24	6.16	**7.26**	7.21
Q3_causal	**6.99**	6.49	6.54	6.84
Q4_crossdomain	7.91	8.09	**8.16**	8.03
Q5_multilevel	**8.20**	7.77	7.66	7.61
Q6_security_climate	7.84	**8.03**	7.86	7.70
Q7_lulucf	7.67	7.51	7.69	**7.81**
Q8_agriculture	**8.20**	7.79	7.77	7.86
Q9_energy	**7.90**	7.70	7.49	7.66
Q10_bioeconomy	8.44	**8.49**	7.44	8.19

The per-question pattern is heterogeneous. No-RAG attains the highest mean on four questions (Q3, Q5, Q8, Q9), single-RAG on three (Q1, Q6, Q10), agentic on two (Q2, Q4), and feedback on one (Q7). The largest condition-to-condition spreads occur on Q2_adversarial (1.10 points, where single-RAG collapses to 6.16) and Q10_bioeconomy (1.05 points, where agentic drops to 7.44). The smallest spreads occur on Q4_crossdomain (0.25) and Q7_lulucf (0.30). Retrieval helps most clearly on the temporally bounded Q1, where the three RAG conditions all exceed no-RAG by 0.57–0.66 points; it hurts most clearly on Q2, which asks for an evaluative verdict on the reduction of the biofuel blending obligation (reduktionsplikten) and where single-RAG falls to 6.16. The corpus admits material published from 1990 onward and is not filtered or weighted by recency (Section 2.1.1), so one candidate explanation is that retrieval on this topic surfaces material describing the obligation before its reduction, which is accurate but temporally misaligned with a 2026 reference. This explanation is not tested here. Establishing it would require the retrieved-chunk analysis set out in Section 6.2, and it is offered as a hypothesis rather than a finding. The two questions with the smallest condition spreads, Q4 and Q7, are also two on which the answers diverge substantively from the reference and from one another; Section 3.12 reads them as analysis rather than as scores.

3.4 Dimension-Level Breakdown

The composite judge score is the unweighted mean of five dimensions: *täckning* (coverage), *korrekthet* (factual accuracy), *argumentation* (logical coherence), *struktur* (structure and readability), and *policyinsikt* (depth of policy reasoning). Table 3.4 reports the cross-model mean

for each dimension by condition.

Table 3.4: Cross-model mean of each judge dimension by condition (scale 1–10), averaged over both judges. All seven analyst models are included.

Dimension	No-RAG	Single-RAG	Agentic	Feedback
Täckning	7.70	7.71	7.61	7.74
Korrekthet	8.48	8.29	8.14	8.25
Argumentation	7.01	6.78	6.79	6.89
Struktur	8.01	7.71	7.82	7.89
Policyinsikt	7.59	7.47	7.48	7.59

The cross-condition spread is narrow on every dimension (≤ 0.34 points). No-RAG holds the highest mean on *korrekthet* (factual accuracy, 8.48), *struktur* (8.01) and *argumentation* (7.01), and ties with feedback for the highest mean on *policyinsikt* (7.59). The only dimension on which a RAG condition leads outright is *täckning* (coverage), where feedback (7.74) marginally exceeds no-RAG (7.70). In other words, under this judge panel retrieval does not improve perceived factual accuracy or reasoning quality; if anything it slightly depresses them, while leaving coverage roughly unchanged.

Two of the five dimensions are defined in closely related terms, *argumentation* as the coherence of the reasoning chain and *struktur* as the clarity of the argument's construction (Section 2.2.2), so it is worth asking whether they measure separable properties. Across the scored cells the rank correlation between them is Spearman $\rho = 0.47$, pooled over both judges. The association is substantial but well short of redundancy, so the two dimensions carry partly distinct information and the composite is not simply double-counting the formal properties of an answer.

3.5 Citation Behaviour

Citations are evaluated deterministically by the `citation_metrics()` routine against the saved `chunks_used` record (Section 2.2.4). The audit distinguishes valid closed-universe references ([Källa N] with N pointing to an actual provided chunk), invalid out-of-universe references, and Harvard-style parenthetical references, which the citation contract forbids in every condition.

3.5.1 Closed-Universe Citation Use

The mean number of valid [Källa N] references per answer is 9.86 (single-RAG), 9.41 (agentic), and 11.94 (feedback). The no-RAG condition produces no [Källa N] references by construction. The feedback condition's higher count reflects the gap-fill retrieval step appending additional labelled chunks to the citation universe (Section 2.1.4). Notably, no invalid out-of-universe

[Källa N] references occur in any condition in the final run, and Harvard-style fabrications are rare (0.16–0.19 per answer).

3.5.2 Citation Traceability

The traceability ratio cit_traceability $= n_{\text{Källa valid}}/(n_{\text{Källa valid}} + n_{\text{Källa invalid}} + n_{\text{Harvard}})$ is reported per model and condition in Table 3.5.

Table 3.5: Mean citation traceability by model and condition, with the mean number of citation attempts per answer given in parentheses. An attempt is any reference the model emits, valid or invalid [Källa N] or Harvard-style. A traceability of 1.00 indicates that every attempt resolves to an actually-provided chunk. The ratio is undefined where a model makes no attempt at all and is reported as zero in that case, which is why the attempt counts are needed to read the zeros. By construction, the no-RAG column is zero for all models.

Model	No-RAG	Single-RAG	Agentic	Feedback
gemma-4-e4b-it	0.00 (0.0)	1.00 (15.8)	1.00 (15.3)	1.00 (17.8)
granite-4.1-8b	0.00 (0.0)	0.35 (5.7)	0.00 (0.8)	0.00 (0.8)
gemma-4-31b-it@4bit	0.00 (0.0)	1.00 (26.9)	1.00 (22.2)	1.00 (29.7)
granite-4.1-30b	0.00 (0.0)	0.10 (1.1)	0.26 (3.1)	0.39 (7.5)
qwen3.6-27b-ud-mlx	0.00 (0.0)	0.40 (7.8)	0.50 (9.8)	0.70 (11.2)
qwen3.6-35b-a3b-ud-mlx	0.00 (0.0)	0.79 (13.0)	1.00 (15.8)	1.00 (17.8)
gemma-4-26b-a4b@4bit	0.00 (0.0)	0.00 (0.0)	0.00 (0.0)	0.00 (0.0)

Two models (`gemma-4-e4b-it` and `gemma-4-31b-it@4bit`) achieve traceability of 1.00 across all three RAG conditions, `qwen3.6-35b-a3b-ud-mlx` reaches 1.00 in agentic and feedback (0.79 in single-RAG), and `qwen3.6-27b-ud-mlx` rises to 0.70 in feedback. The two Granite models exhibit substantially weaker grounding: `granite-4.1-30b` ranges from 0.10 to 0.39 across RAG conditions, and `granite-4.1-8b` produces grounded citations only in single-RAG (0.35).

The zeros in the table arise for two behaviourally distinct reasons, which the attempt counts separate. `granite-4.1-8b` scores 0.00 in agentic and feedback while still making 0.8 attempts per answer: it cites, but exclusively in the forbidden Harvard form, emitting no [Källa N] tags at all. The failure-mode `gemma-4-26b-a4b@4bit` scores 0.00 because it makes no citation attempt of any kind in any condition, so its ratio is undefined rather than poor. A model that declines to cite and a model that cites unverifiably would both read as 0.00 on the ratio alone, and the distinction matters for any downstream use of grounding as a selection criterion. Consistent with this, no invalid out-of-universe [Källa N] reference occurs anywhere in the final run: where the models cite by chunk label at all, they cite labels that exist.

Citation traceability is thus the one axis on which the RAG conditions clearly and categorically outperform no-RAG, which offers no verifiable grounding at all. This advantage is not reflected in the judge composite, which still ranks no-RAG highest.

3.6 Overlap with the Gold Standard

Overlap with the gold standard is measured along two axes. Lexical overlap (ROUGE-L) captures order-sensitive surface agreement; semantic similarity (BERTScore) captures contextual-embedding agreement. All metrics are computed on the citation-stripped answer against the held-out Klimatpolitiska rådet 2026 reference.

3.6.1 Lexical and Semantic Overlap

Table 3.6 reports the cross-model means of ROUGE-L F1, BERTScore F1, and answer length per condition.

Table 3.6: Cross-model mean overlap metrics by condition, computed on citation-stripped answers against the gold standard.

Metric	No-RAG	Single-RAG	Agentic	Feedback
ROUGE-L F1	0.179	0.182	0.183	0.182
BERTScore F1	0.860	0.863	0.863	0.862
Ordantal (words)	463	395	442	449

Neither metric separates the conditions. ROUGE-L sits at approximately 0.18 F1 against references averaging 484 words, which is close to the floor for this pairing: at that level the statistic reflects the base rate of Swedish function-word overlap rather than any property of the candidate, and the 0.003–0.004 differences between conditions carry no information. It is retained here as a negative result and is not interpreted further. BERTScore behaves the same way for a different reason, discussed below. Answer length is highest in no-RAG (463 words) and lowest in single-RAG (395 words); the formal length-bias analysis (Section 3.7) finds no association between length and judge score.

At the per-model level the nominally highest condition is split almost evenly between single-RAG, agentic and feedback, with the within-model spread generally below 0.01 F1 points (Table C.3). That split is better read as noise than as a ranking, for the reason given next.

One property of the semantic metric should be borne in mind when reading these differences. BERTScore is computed here without baseline rescaling (Section 2.2.1), so its values are not anchored to a meaningful zero: two arbitrary Swedish texts of similar register will typically score in the low 0.80s on this encoder, and the entire observed range across all conditions and models (0.849–0.872) sits inside that band. Differences of 0.002–0.004 between conditions are therefore not interpretable as differences in semantic fidelity. The metric is reported here principally to demonstrate that it does not separate the conditions, rather than to rank them. Taken with ROUGE-L, the position is that neither overlap metric resolves the experimental manipulation, which is itself part of the answer to RQ3.

3.7 Agreement Between the Two Automated Judges

Each answer was scored independently by two judges. Judge A (`microsoft/phi-4`) is systematically stricter, with a cross-condition mean of approximately 7.1, while Judge B (`nvidia/nemotron-3-nano-4b`) is more lenient at approximately 8.2. Their per-cell agreement is moderate: Spearman $\rho = 0.37$ ($p < 10^{-10}$, $n = 279$), shown in the left panel of Figure 3.3. The two judges concur on broad ordering but disagree substantially on individual cells while differing by 1.1 points in mean calibration. This is the clearest return on the dual-judge design: a single-judge protocol would have reported one of these two calibrations as the score, with no way to detect that the other existed. Any absolute LLM-judge figure quoted without a second judge alongside it should be read with that in mind.

The pairing was also constrained. Both judges had to fit alongside a 27–35B analyst model in local memory, so Judge B scores outputs from models several times its size, and the analysis below treats Judge A as the primary instrument with Judge B as a robustness check on it.

The two judges also differ in dispersion. Across the 279 scored cells the standard deviation of the composite is 1.59 for Judge A against 0.97 for Judge B, so Judge B is both more lenient and more compressed. Because a compressed judge contributes less rank information to the mean, it is worth checking whether the condition ranking reported throughout this chapter is in fact driven by Judge A alone. It is not. Taken separately, Judge A ranks the conditions no-RAG (7.15), feedback (7.14), single-RAG (7.12), agentic (7.01), and Judge B ranks them no-RAG (8.37), feedback (8.21), agentic (8.13), single-RAG (8.06). Both judges independently place no-RAG first and feedback second, and differ only in the ordering of the two lowest conditions, which are separated by 0.11 points under Judge A and 0.07 under Judge B. Since the two judges differ by more than a point in mean calibration and by a factor of 1.6 in dispersion, the composite was additionally recomputed with each judge's scores standardised to zero mean and unit variance before averaging. Under that transformation the ranking is unchanged: no-RAG ($+0.11$), feedback ($+0.02$), single-RAG (-0.06), agentic (-0.06). The dual-judge design therefore exposes a substantial calibration asymmetry without that asymmetry driving the reported ordering.

A length-bias check correlating the composite judge score against answer word count returns Spearman $\rho = 0.04$ ($p = 0.52$, $n = 279$), indicating that, unlike the bias documented for some judge models in prior work, length is not a meaningful confound in this setting.

3.8 Preliminary Human Evaluation

A blinded human evaluation was conducted following the protocol in Section 2.2.5. The four conditions were randomly re-labelled A–D independently for each question, and the rater, the

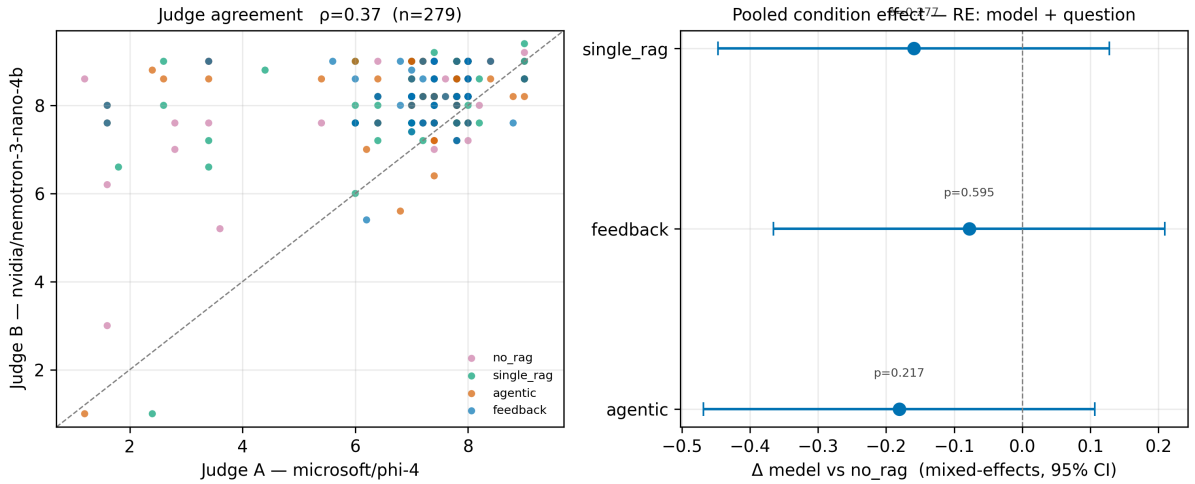

Figure 3.3: Left: per-cell agreement between Judge A (`microsoft/phi-4`) and Judge B (`nvidia/nemotron-3-nano-4b`), Spearman $\rho = 0.37$ ($n = 279$); points coloured by condition. Right: pooled condition effect relative to no-RAG from a linear mixed-effects model with random intercepts for model and question. All three RAG conditions sit below the no-RAG baseline; none is statistically significant.

author working blind to the labels, selected the best and worst response per item and flagged unusable answers. All results reported here come from this *single* rater. The condition ranking uses the rater's choices on all 70 items; comparisons with the automated judge use the 69 items that have a judge score for every condition (one no-RAG cell failed at the judge stage). No inter-rater reliability statistic accompanies these figures (Section 2.2.5), and they are reported throughout as preliminary.

3.8.1 Human Condition Ranking

Aggregating the human best/worst choices into a MaxDiff-style utility per condition yields the ranking feedback ($+0.114$) $>$ agentic ($+0.043$) $>$ single-RAG (0.000) $>$ no-RAG (-0.157). This differs sharply from the automated judge's ranking, in which no-RAG places first. The human rater therefore prefers the retrieval-augmented answers, and most prefers the feedback condition, precisely where the automated judge prefers the ungrounded baseline.

Because citation stripping was incomplete on 13 of the 70 presented items (Section 2.2.5), and because every residual marker falls in a retrieval-augmented arm, this ranking was recomputed on the 57 items that carried no visible marker in any answer. The result is materially unchanged: feedback ($+0.140$) $>$ agentic ($+0.018$) $>$ single-RAG (0.000) $>$ no-RAG (-0.158). The no-RAG utility in particular moves by 0.001. The rater's preference for the retrieval-augmented conditions is therefore not an artefact of the residual citation markers, although it remains exposed to the specificity channel that stripping cannot remove (Section 4.5).

3.8.2 Human–Judge Agreement

Treating each condition's human utility and automated-judge mean as paired observations, the two condition-level orderings are negatively associated: Spearman $\rho = -0.40$ and Kendall $\tau = -0.33$. These coefficients are computed over only the four condition aggregates ($n = 4$); they therefore cannot reach significance and would flip on a single swap, so they are reported as a descriptive summary rather than a test, with the item-level agreement below as the more reliable measure. Figure 3.4 plots the condition utilities against the judge means and shows the negative slope, with no-RAG isolated in the high-judge/low-human corner and feedback in the low-judge/high-human corner.

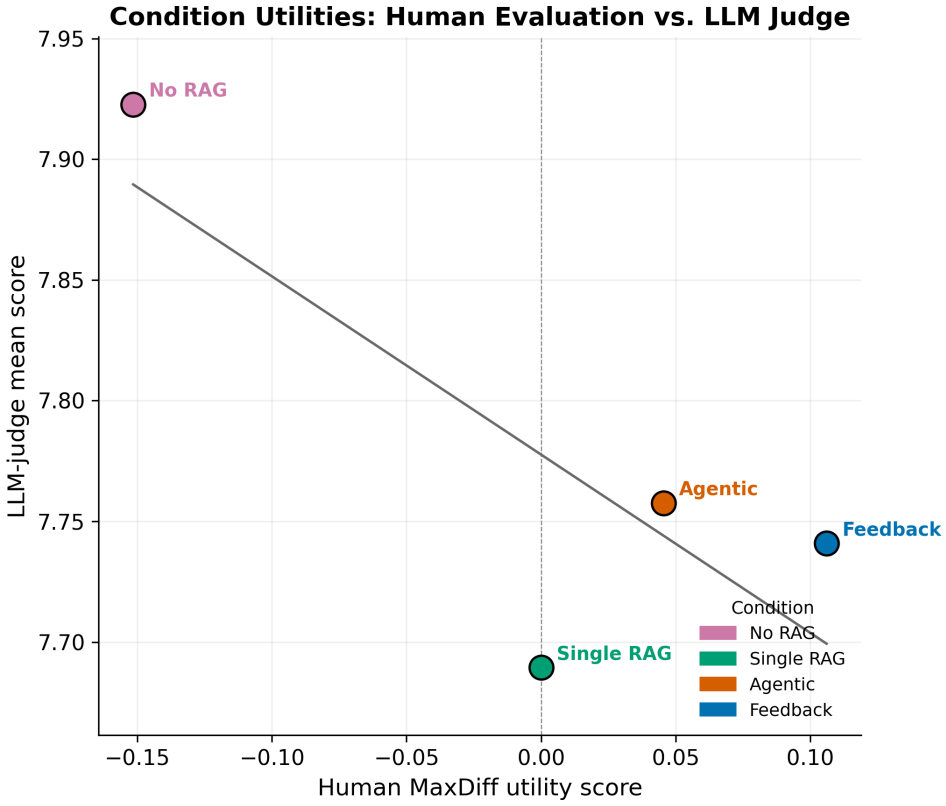

Figure 3.4: Per-condition human MaxDiff utility (horizontal) against automated LLM-judge mean score (vertical), with regression line. Both axes use the 69 items with complete judge scores, so the judge means differ from the full-run values in Table 3.1 in both level and ordering, and the human utilities differ slightly from the 70-item values in Section 3.8.1, with the same ordering. That the ordering itself shifts between the full run and this subset is consistent with the resolution argument of Section 3.2.2. The negative slope is a four-point fit and is descriptive only.

At the item level, agreement between the human's single best pick and the judge's highest-scoring condition is 26% (95% CI 16–36%), barely above the 25% chance baseline for four options; worst-pick agreement is 36% (95% CI 25–48%); and implied-pairwise agreement across all ordered condition pairs is 44.5% (95% CI 39–50%). These rates are visualised in Figure 3.5. Per-model best-pick agreement (Table 3.7) is highest for `qwen3.6-35b-a3b-ud-mlx` (40%) and otherwise clusters at or below 33%.

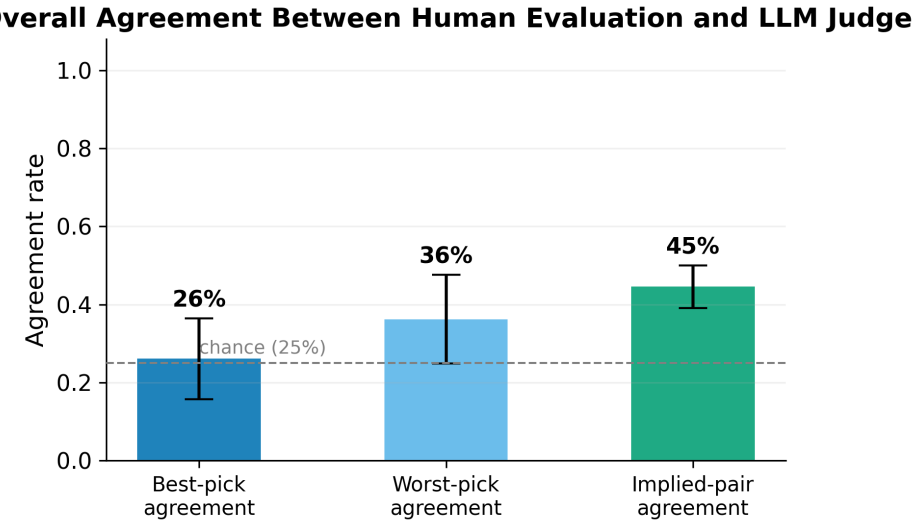

Figure 3.5: Agreement between the human rater and the automated judge on best pick, worst pick, and implied ordered pairs, with 95% confidence intervals. The dashed line marks the 25% chance baseline for the four-option best/worst picks.

Table 3.7: Per-model agreement between the human rater's best/worst picks and the automated judge (n items per model). Preliminary, single rater.

Model	n	Best-pick agr.	Worst-pick agr.
qwen3.6-35b-a3b-ud-mlx	10	40%	20%
gemma-4-26b-a4b@4bit	9	33%	44%
gemma-4-e4b-it	10	30%	40%
gemma-4-31b-it@4bit	10	20%	40%
granite-4.1-8b	10	20%	50%
qwen3.6-27b-ud-mlx	10	20%	40%
granite-4.1-30b	10	20%	20%

The interval on the best-pick rate spans the chance baseline, so the correct reading is not that the rater and the judge disagree systematically but that this comparison detects no relationship between them in either direction. Combined with the inter-judge agreement of $\rho = 0.37$ and the flatness of both overlap metrics, the position at the level of ranking usable answers is that none of the instruments applied here demonstrates validity as a ranker. The condition-level inversion reported above is consistent with that reading but does not strengthen it, resting as it does on four aggregates and one rater who also designed the experiment.

One result stands clear of all this. The rater flagged 12 of the 280 answers as unusable. These fall at a median 4.7th percentile of the automated judge's scores, and 8 of the 12 belong to `gemma-4-26b-a4b@4bit`, the model that the composite (Table 3.1) and the deterministic citation audit (Table 3.5) independently identify as failing. The flags are distributed evenly across the four retrieval conditions, three in each, so they track the model rather than the manipulation. Three instruments of quite different kinds, a graded LLM rubric, a deterministic grounding check, and

a human reader, therefore converge on the same failure case while disagreeing about everything above it. Detecting unusable output is the one evaluation task on which this instrument set demonstrably works, and because it turns on agreement rather than on ranking, it does not depend on the number of raters.

3.9 Efficiency and Cost

Wall-clock time per answer is broken down by model in Table C.4 and visualised against quality in Figure 3.6. The no-RAG condition is the fastest for every model except `gemma-4-26b-a4b@4bit`, whose no-RAG time is inflated by repetition loops.

One property of the instrumentation must be understood before these figures are read. The feedback tier reuses the agentic draft rather than regenerating from scratch (Section 2.1.4), and its timer starts after that draft has been received. The recorded feedback time is therefore the *increment* over agentic, covering gap-query generation, gap retrieval and the refinement pass, and not the cost of producing a feedback answer from nothing. Two quantities follow, and they answer different questions. Summing the four recorded times gives the true wall-clock of the experiment as run, because the agentic draft was computed once and used twice; no work is double-counted. However, the cost of *choosing* the feedback tier in deployment is the agentic time plus the feedback increment, since a system configured for that tier must produce the draft as well. Across the six capable models this is $191.1 + 139.8 = 330.9\,\text{s}$, which makes feedback the most expensive condition by a wide margin rather than, as the recorded increment alone would suggest, the cheapest of the three retrieval tiers.

On that basis, and among the retrieval conditions, single-RAG is the cheapest, agentic next at 3.2 times the unretrieved baseline, and feedback the most expensive at 5.6 times.

Two of the models with the fewest active parameters (`gemma-4-e4b-it`, a 4B dense model, and `qwen3.6-35b-a3b-ud-mlx`, a mixture-of-experts model engaging 3B parameters per token) sit in the favourable region of the cost-quality plot, holding composite scores above 7 in every condition while answering in under 80 seconds in all but the feedback tier. Because no-RAG is both the fastest and, under this judge panel, the highest-scoring condition for the capable models, the cost-quality trade-off in the final run does not favour the more elaborate retrieval tiers.

3.9.1 Energy Use and Carbon Footprint

Because the conditions ran on identical hardware, the times above convert into a resource figure. This is reported as an accountability exercise rather than a finding: the absolute quantities are negligible and nothing in the thesis rests on them.

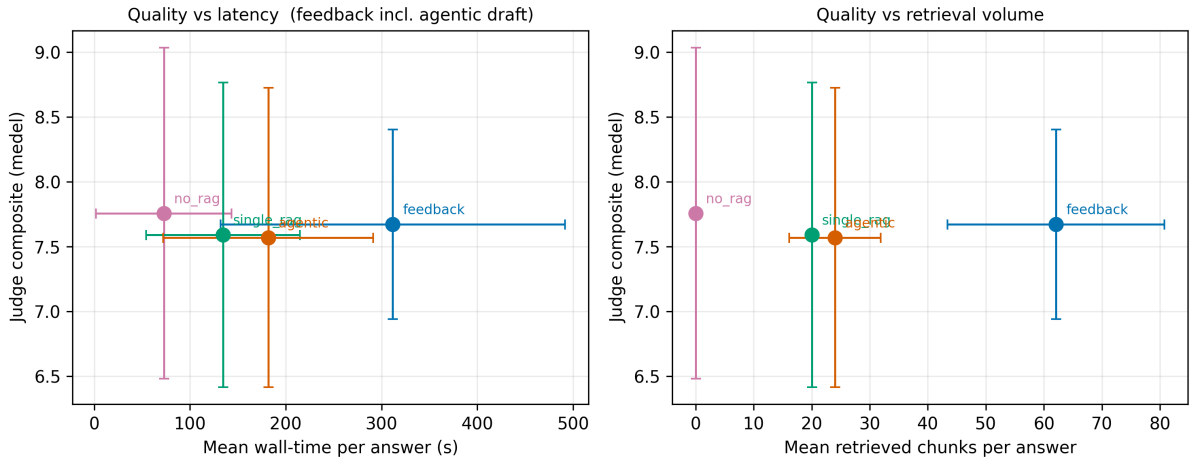

Figure 3.6: Quality against resource cost. Left: mean composite judge score against mean wall-clock time per answer, with the feedback tier plotted at its full standalone cost, the agentic draft plus the refinement increment (Table 3.8). Right: the same scores against the mean number of retrieved chunks placed in context. No-RAG sits furthest left on both panels and, for the capable models, also highest.

The operational term follows the Green Algorithms model [43], $E = t \times P \times \text{PUE}$ and $C = E \times I$, and the embodied term the Software Carbon Intensity specification [44], which amortises manufacturing over the device lifetime. All inference ran on one Mac Studio (Apple M1 Max, 64 GB) in Stockholm during the spring of 2026, at PUE $= 1.0$, the value Green Algorithms prescribes for desktop work. Apple publishes wall-measured draw of 11 W idle and 115 W at maximum load for the M1 Max model [45]; power was not instrumented during the runs, so a duty-cycle-weighted mean of 70 W is assumed, which is the largest uncertainty in the estimate. Grid intensity is the flow-traced Swedish mean for 2025, 21.3 gCO_2e/kWh [46]. Runtime is read from the saved run manifests rather than assumed, so the figures below follow from the recorded artefacts and the constants cited here.

Generation across all pilot and final runs totals 60.5 hours of measured wall-clock, of which the final scored run is 10.1 hours; corpus parsing adds an estimated 41.5 hours. Embedding, judge scoring and the sensitivity re-runs were not instrumented, so the total of 102 hours is a lower bound. At 70 W that is 7.1 kWh, giving operational emissions of 0.15 kg CO_2e.

The embodied term depends on an allocation choice. Apple's life-cycle assessment puts the closest published configuration (M1 Max, 32 GB memory, 512 GB storage) at 262 kg CO_2e, of which 66% is production, transport and end-of-life processing over an assumed four-year first-owner period [47]; amortising that across the 102 hours would add 0.50 kg. The machine used here was bought second-hand, however, and under the cut-off allocation conventional in life-cycle assessment the production burden stays with the first owner, leaving a reuse phase with approximately none of it. The total is therefore a range, roughly 0.15 to 0.66 kg CO_2e, or at most four kilometres of car travel [43].

The distribution matters more than the total. Had the machine been new, manufacturing would

have accounted for 77% of the footprint, so on a grid as clean as Sweden's the dominant emission from local inference is building the machine rather than running it. Counting only operational energy therefore measures the smaller half of the problem, and buying the hardware second-hand removed more emissions from this project than any plausible software optimisation could have.

Table 3.8 gives the cost of each tier across the six capable models, excluding `gemma-4-26b-a4b@4bit` because its no-RAG figure of 161.4 s comes from repetition loops rather than useful generation and inflates the baseline. Feedback is reported at its full standalone cost, the agentic draft plus the refinement increment, for the reason given above. The time ratios are exact; their conversion to energy assumes constant mean power across conditions, which is only approximately true.

Table 3.8: Resource cost per answer by retrieval condition, six capable models (excluding `gemma-4-26b-a4b@4bit`; see Section 3.11). The feedback row gives the standalone cost of that tier, being the agentic draft (191.1 s) plus the refinement increment (139.8 s). Energy and emissions assume a mean draw of 70 W and a grid intensity of 21.3 gCO_2e/kWh, and cover operational energy only.

Condition	s/answer	Relative	Wh/answer	gCO_2e/answer
No-RAG	59.4	1.00	1.16	0.025
Single-RAG	149.1	2.51	2.90	0.062
Agentic	191.1	3.22	3.72	0.079
Feedback	330.9	5.57	6.43	0.137

Read against the quality results, the ordering is unfavourable to elaboration at every step. The agentic tier costs 3.2 times the baseline and never attains the highest composite score for any analyst model; the feedback tier costs 5.6 times the baseline and ranks second on quality by a margin the evaluation cannot resolve. The one benefit retrieval does buy, verifiable citation grounding, is already available at the single-RAG tier, where two models reach full traceability, for 78% of the agentic cost and 45% of the feedback cost. At ten questions none of this matters; across 100 000 answers the premium over the baseline would be roughly 5.5 kg CO_2e for agentic and 11.2 kg for feedback in Sweden, and 54 and 111 kg respectively at an illustrative EU-27 average of 210 gCO_2e/kWh, for no quality return measurable by any instrument used here.

3.10 Prompt Sensitivity Analysis

3.10.1 Context Format: Tabular versus Linearised

Table-fact extraction accuracy is reported in Table 3.9. Six of the seven analyst models answered all eight probes correctly from the table, and did so equally from the linearised prose. Averaged over models, accuracy is 98% from the table and 95% from prose, a format penalty of -3.6 percentage points. The tabular presentation is therefore very slightly *easier*, not harder. With

the table removed no model produced a correct figure at all, confirming that the correct answers are being read out of the supplied context rather than recalled from pre-training.

The entire departure from ceiling is one model. `gemma-4-26b-a4b@4bit` scores 88% from the table and 62% from prose, and is the only model that ever failed a probe. Its deficit is therefore not a table-reading deficit, since it is worse on prose than on the table, and it is consistent with the broader instruction-following failure documented for this model in Section 3.11. Every completion in the experiment that exhausted the token budget also belongs to this model. Of the 56 model–probe pairs, only four were answered correctly under one presentation but not the other, and all four are this model.

The practical consequence is that the corpus-wide caveat this check was designed to detect does not apply. Where a RAG answer misses a quantitative fact that was retrieved, the explanation lies in retrieval or in generation, not in the markdown table format. The scope of the claim is narrow, however. The probes are single-number extractions from a short, abridged table presented on its own, which is an easier task than locating the same figure inside a full retrieved chunk set. The check rules out format as a bottleneck; it does not certify numerical fidelity in the benchmark answers. Two of the eight probe golds are also values that appear in another row of the same table, so a wrong-row read would still score as correct on those two, biasing the figures upward slightly.

Table 3.9: Exact-figure extraction accuracy (%) by presentation format, for eight numeric probes drawn from three tables in the held-out report. The penalty column is prose accuracy minus table accuracy, in percentage points; positive values would indicate that tabular formatting costs accuracy.

Model	Table	Prose	Removed	Penalty (pp)
gemma-4-e4b-it	100	100	0	0
granite-4.1-8b	100	100	0	0
gemma-4-31b-it@4bit	100	100	0	0
granite-4.1-30b	100	100	0	0
qwen3.6-27b-ud-mlx	100	100	0	0
qwen3.6-35b-a3b-ud-mlx	100	100	0	0
gemma-4-26b-a4b@4bit	88	62	0	−25
Mean	98	95	0	−3.6

3.10.2 Judge-Prompt Language

Translating the judge rubric from Swedish to English lowers scores systematically and by a small amount. On the 62 paired units, Judge A falls from a mean composite of 7.17 to 6.99 ($\Delta = -0.18$, Wilcoxon $p = 0.014$) and Judge B from 8.29 to 8.07 ($\Delta = -0.23$, $p < 0.001$). Both shifts are statistically detectable, and both are under a quarter of a point on a 1–10 scale.

The shift is not spread evenly across dimensions (Figure 3.7). Pooled over judges, the English

rubric scores *policyinsikt* 0.59 points lower, *korrekthet* 0.28 lower and *täckning* 0.23 lower, leaves *struktur* essentially unchanged at -0.02, and scores *argumentation* marginally higher at $+0.12$. The dimensions that lose most are those requiring the judge to reason about Swedish policy substance rather than about the form of the answer, which is the pattern one would expect if the English rubric slightly degrades the judge's engagement with Swedish-language domain content.

At the condition level the two languages agree on the shape of the result but not on the exact ordering (Table 3.10). Under the Swedish rubric the order is feedback (7.92), single-RAG (7.89), no-RAG (7.59), agentic (7.53); under English it is single-RAG (7.73), feedback (7.69), no-RAG (7.41), agentic (7.32). Spearman's ρ between the two sets of condition means is 0.80, and the entire departure from $\rho = 1$ is the transposition of the top two conditions, which are separated by 0.03 points in Swedish and 0.04 points in English, far below the resolution of the instrument. The bottom two hold their order under both languages, separated by 0.06 and 0.09 points respectively.

This is a mixed outcome. The judge is language-sensitive. Absolute scores are not invariant to the rubric language, and the nominal condition ranking is not stable either, so any statement of the form "condition X ranks first" is contingent on the prompt language and should not be made. Note further that the Swedish column of Table 3.10 does not reproduce the ordering of the full run reported in Section 3.2.2, placing no-RAG third rather than first on this subsample; that discrepancy is discussed there, and the two observations point the same way. What the check does support is that the substantive conclusion of this thesis does not depend on either choice. The four conditions span only 0.39 points in total under the Swedish rubric, and the conclusion drawn from that is precisely that they are not separated by an amount the evaluation can resolve (Section 3.2.3). The English re-judge reproduces exactly that picture: the same tight cluster, the same absence of a usable gap, with every condition shifted down by roughly 0.2 points. A rank order that flips under a 0.03-point perturbation, or under the removal of four fifths of the data, is consistent with the null finding, not evidence against it.

Table 3.10: Per-condition mean composite score under the Swedish and English judge rubrics, pooled over both judges on the 62 paired units. Ranks are within column.

Condition	Swedish	Rank	English	Rank
Feedback	7.92	1	7.69	2
Single-RAG	7.89	2	7.73	1
No-RAG	7.59	3	7.41	3
Agentic	7.53	4	7.32	4
Spearman ρ (condition means, Swedish vs English) $= 0.80$				

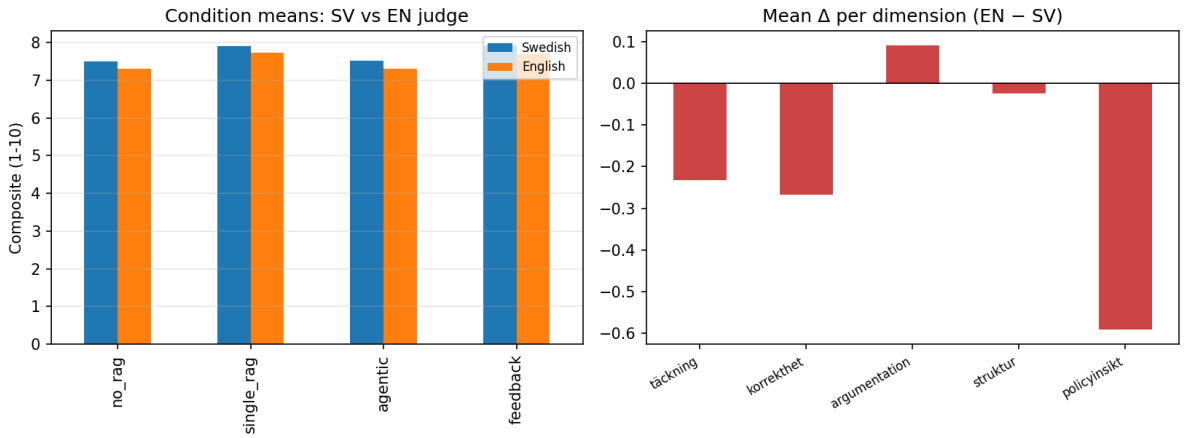

Figure 3.7: Left: per-condition mean composite under the Swedish and English judge rubrics. Right: mean per-dimension change (English minus Swedish), pooled over both judges.

3.11 Documented Failure Mode: `gemma-4-26b-a4b@4bit`

The Gemma-4 26B mixture-of-experts model at 4-bit quantisation did not satisfy the floor of usable analytical output. Its no-RAG mean composite is 5.16 on the 1–10 scale, more than two points below every other model's no-RAG score, and its three RAG conditions attain 5.95–6.91. It is one of only two models, with `granite-4.1-8b`, for which all three RAG conditions score above no-RAG, and it emits no well-formed [Källa N] tags in any condition, so its citation traceability is zero throughout (Table 3.5). One of its forty cells failed at the judge stage (no-RAG, Q5).

The absolute scores in the final run (5–7) are higher than in earlier pilot runs because Judge B (`nvidia/nemotron-3-nano-4b`) is markedly more lenient; the qualitative degradation is unchanged. Two failure modes remain visible in the answer corpus: repetition loops, in which the model emits the same sentence or paragraph fragment until the token budget is exhausted (the no-RAG answer to Q1 runs to roughly twenty thousand characters of repeated text), and language drift, in which the answer partially switches from Swedish to another language mid-paragraph. Gemma-specific chat-template markers (`<start_of_turn>`, `<end_of_turn>`) are removed by the post-processing routine `_strip_leaks()` but indicate the same underlying instability.

This model is retained in the cross-model analysis as a documented failure case rather than excluded. Its condition-relative pattern, in which RAG improves on the no-RAG baseline, is the opposite of the five models that peak without retrieval. It is reported here but should be read in the light of its very low and highly variable absolute scores.

The blinded human evaluation corroborates this assessment independently. Of the 12 answers

the rater flagged as unusable across the whole experiment, eight belong to `gemma-4-26b-a4b@4bit`, that is, a fifth of its 40 cells. The remaining four flags are split between `granite-4.1-8b` (two) and `qwen3.6-35b-a3b-ud-mlx` (two), each on the adversarial Q2, in the single-RAG condition, or both, while the other four analyst models received no unusable flags at all. A human rater, scoring blind, therefore converges on the same model that the composite score and the citation audit identify as failing.

3.12 The Answers as Climate and Energy Analysis

Every result so far treats an answer as something that receives a score. This section instead reads some of the answers as climate and energy policy analysis. The question is not how closely they resemble the reference, but what they conclude about the Swedish system, and whether a sector analyst could use those conclusions.

The reason for doing this is the ceiling described in Sections 2.1.5 and 4.2. The reference answers were generated from the summary of the held-out report rather than its full text, so an answer that adds relevant content cannot score higher than one that does not, and none of the instruments in this study picks up the difference. The only way to see what that difference consists of is to read the answers.

Procedure

Four benchmark questions were chosen because their subject matter is directly about energy and climate systems: Q4 (cross-domain synthesis), Q5 (multi-level governance), Q7 (land use and carbon sinks) and Q9 (energy system reasoning). For each of the seven analyst models, the answer the blinded human rater picked as best (Section 2.2.5) was used, giving 28 answers alongside the four reference answers.

From these texts a set of analytical propositions was identified: claims about how the Swedish system works or what should be done, of the kind an assessment would either support or reject. For each proposition, Table 3.11 records the position of the reference answer, which stands in for the expert view of the held-out Klimatpolitiska rådet report [36], and how many of the seven answers advance or reject it. The text after the table checks the main points of agreement and disagreement against official statistics and against what the questions actually ask.

Two limits apply. The propositions were identified and coded by the author, who also designed the benchmark and rated the answers. And the comparison covers four of ten questions, one condition per model, from a single generation run.

Table 3.11: Analytical propositions in the answers to the four energy and land-use questions. Ref. gives the position of the reference answer: • advances the proposition, × rejects or contradicts it, ○ does not address it. Models gives the number of the seven human-preferred answers that advance the proposition, with the number that explicitly reject it in parentheses where non-zero. Restating the question without taking a position does not count as advancing a proposition, so `gemma-4-26b-a4b@4bit` registers on none of the fourteen rows (Section 3.11).

Proposition	Ref.	Models
Q4 Transport, energy and industry		
Transport electrification and industrial decarbonisation compete for the same fossil-free electricity and network capacity	○	6/7
The principal obstacle is coordination failure, remediable by a cross-sectoral coordinating body or plan	•	6/7
Competition for biomass feedstock and for critical minerals makes the sectors materially rival, not merely poorly coordinated	○	3/7
Infrastructure planning premised on traffic forecasts rather than climate targets actively counteracts the transition	○	1/7
Q5 EU and national governance		
EU regulation now constrains national room for manoeuvre rather than merely supplementing it	•	6/7
Municipal capacity and financing, not EU stringency, is the binding constraint on local implementation	○	4/7
Q7 Land use and carbon sinks		
The 2030 LULUCF commitment is within reach on current trends	•	2/7 (4)
Harvest rate is the most effective short-term lever on net uptake	○	2/7 (1)
A shortfall in the sink transfers abatement effort to the ESR sectors, principally transport and heating	○	4/7
The measures currently proposed are quantitatively insufficient, and the shortfall can be stated numerically	○	2/7
Q9 Electricity system		
Network capacity, not generation capacity, is the binding constraint	○	3/7
Municipal veto rights are the principal obstacle to wind expansion	○	2/7
Energy efficiency should be treated as a primary resource rather than a residual	○	6/7
The question's request for a ranked set of priorities is answered	×	3/7

Where the answers and the reference agree

The answers agree with each other and with the reference mostly on general points. Six of the seven answers to Q4 (the seventh is the failure case in Section 3.11) recommend a body or plan to coordinate the sectors, as does the reference. The same six note that the sectors compete for the same fossil-free electricity and grid capacity, and on Q5 the same six agree that EU regulation now limits national policy rather than only adding to it.

This agreement tells a reader little. A call for coordination needs no evidence, fits any period, and does not help choose between options, which is what an assessment is for. The answers agree most where agreeing costs nothing. Someone reading a single answer would have no way to tell a well-founded consensus from a default one, while most of the propositions that could actually inform a decision are advanced by three answers or fewer.

Where the answers and the reference disagree

The clearest disagreement concerns whether Sweden is on track for its 2030 LULUCF commitment. The reference says that net uptake rose sharply over 2022–2024 and that the outlook has improved. Four of the seven answers say the opposite, and only two agree, both with conditions. `qwen3.6-27b-ud-mlx` reports that net uptake fell to about 31 Mt CO_2e in 2023, against a long-term average near 55 Mt, and `gemma-4-31b-it@4bit` notes that the only LULUCF measure in the government's action plan, rewetting drained peatland, adds about 0.1 Mt against a required increase of nearly 4 Mt from the 2016–2018 level [48].

Both views match an official release of the statistic. The official figure for 2023 was 31 Mt when first reported [49]. In the next release, the three most recently reported years (2021–2023) were recalculated with new data from the National Forest Inventory, as the reporting method prescribes, and the sector was reported at 54 Mt for 2024, with a difference of 8 Mt between 2023 and 2024 [50]. Much of the gap between the two figures therefore comes from recalculation rather than from change in the forest. The answers and the reference are not split between right and wrong analysis, but between an older and a newer release of the same statistic.

This matters for the energy system, not only for forestry. Sweden's national framework allows extra uptake in forests and land to count towards part of the target for the sectors outside the EU emissions trading system. Whether the sink delivers therefore affects how much emissions must be cut in transport and heating instead. Four of the seven answers make this link explicitly, and the reference does not. Which release of the statistic an analysis uses changes what it implies for transport and heating.

On Q9, the reference is the weaker text. The question asks for a joint analysis of nuclear power, obstacles to wind power, the potential of energy efficiency and the need for grid capacity, and asks which priorities matter most. The reference does not discuss nuclear power or wind per-

mitting, mentions energy efficiency and grid expansion only in passing as possible results of a broad energy agreement, and gives no priorities; its recommendations concern a parliamentary energy agreement and a bioeconomy strategy. `gemma-4-31b-it@4bit` identifies grid capacity as the main bottleneck, names municipal vetoes as a key obstacle to wind power, and ends with an explicit order of priorities. `qwen3.6-27b-ud-mlx` argues for putting energy efficiency first and warns that a one-sided focus on new reactors risks delaying other needed investments.

Judged against what the question asks rather than against the reference, the comparison is simple. The question names four elements and asks for priorities. The reference covers two of the four briefly and sets no priorities. All six functioning answers discuss all four elements, and three of them set out an explicit order of priorities. By the standard the question itself sets, these answers are more complete than the reference. No instrument in this thesis can show this: the automated scores measure similarity to the reference, and the human rater compared the answers only with each other.

A second Q7 answer shows a different kind of problem. `granite-4.1-30b` notes criticism that the EU LULUCF regulation penalises active forestry, because it counts carbon uptake but not the substitution benefit of wood products, and then builds its answer on that view: it recommends rewarding harvesting for long-lived wood products and judges the target reachable if policy rewards substitution. It is the only answer that rejects the idea that a lower harvest rate is the most effective short-term lever. Substitution is a real position in the Swedish forest debate, and the answer's sources do support it, but it is a contested view, and here it becomes the starting point of the analysis rather than one view among several. For an assessment body whose authority depends on independence, this kind of output is of little use and hard to spot, because nothing in its form sets it apart from balanced analysis.

What the answers add that the reference does not

The answers add real content, mostly in the cross-sectoral material Q4 was designed to draw out. `gemma-4-31b-it@4bit` identifies conflicts between goals that the reference does not mention: mining metals for batteries versus reindeer herding and biodiversity; infrastructure planning based on traffic forecasts rather than climate targets; and the greater exposure of an electrified transport system to extreme weather and cyber-attacks compared with a fuel-based one. `qwen3.6-27b-ud-mlx` raises competition for biomass between transport and industry, and the use of electric car batteries to support the grid.

Each of these is a known concern in the Swedish literature, and each is specific enough to check. None of them raises the score of the answer that contains it, because the reference does not mention them. The most specific propositions in Table 3.11, those advanced by three answers or fewer, are mostly ones the reference leaves out, and they come largely from the same two or three models.

These answers are weaker at ranking the conflicts or putting numbers on the trade-offs. After naming the competition for biomass, `qwen3.6-27b-ud-mlx` recommends that it be coordinated; after naming the conflict between mining and reindeer herding, `gemma-4-31b-it@4bit` recommends identifying such conflicts early. The answers map the conflicts more fully than the reference does, but do no better at deciding between them.

Relation to the evaluation findings

None of this shows up in the instruments used in this study. The difference in data behind the Q7 split affects no score: the overlap metrics cannot see it, the citation audit shows where a claim came from but not whether a figure is current, and the rater, reading blind, picked answers on both sides of the split as best for their model. The framing adopted in the `granite-4.1-30b` answer is not something any of the instruments is designed to detect. In the other direction, the cross-sectoral conflicts a sector reader would value most are not rewarded, because the reference they are scored against does not contain them. This sharpens the finding summarised in Section 3.13: on the questions a policy body cares most about, the difficulty is not that these instruments measure the relevant properties poorly, but that they are not designed to measure them.

3.13 Summary of Findings

The principal descriptive findings of the final run are as follows.

At the aggregate level, the four conditions are separated by only 0.19 points on the 1–10 composite scale, and *no-RAG is nominally highest* (no-RAG 7.76, feedback 7.67, single-RAG 7.59, agentic 7.57). No-RAG wins the most first-place model rankings (five of seven) and holds the best mean rank (1.86), though cross-model concordance is low (Kendall's $W = 0.118$). Four of seven models show a significant within-model condition effect, and every significant pairwise comparison favours no-RAG over a RAG condition. A pooled mixed-effects model places all three RAG conditions below no-RAG, none significantly.

At the question level, no-RAG wins four questions, single-RAG three, agentic two, and feedback one; retrieval helps most on the temporally bounded Q1 and hurts most on the adversarial Q2.

At the dimension level, no-RAG leads on factual accuracy, structure and reasoning, and the only dimension on which a RAG condition edges ahead is coverage.

On citation behaviour, the RAG conditions provide grounded, verifiable [Källa N] citations (traceability up to 1.00 for three models) where no-RAG can provide none; no invalid out-of-universe citations occur, and Harvard-style fabrications are rare. This is the one axis on which

RAG categorically outperforms the baseline, and it is invisible to the judge composite.

On overlap with the gold standard, ROUGE-L sits at its floor and BERTScore within its unrescaled baseline band; neither resolves the conditions.

On evaluation methodology, no instrument applied here demonstrates validity as a ranker of usable answers. The two automated judges agree only moderately (Spearman $\rho = 0.37$) and differ by more than a point in mean calibration; the overlap metrics are flat; and the preliminary human evaluation agrees with the judge at chance on individual items (best-pick 26%, 95% CI 16–36%, against a 25% baseline). What the evidence supports is that evaluator choice moves the apparent ranking at least as much as the experimental manipulation does, not that human and automated preferences are systematically opposed.

Detection of unusable output is the exception, and it is robust. The twelve answers the rater flagged fall at a median 4.7th percentile of the judge's scores, and eight of them belong to the model that the composite and the citation audit independently identify as failing. Three independent instruments agree on the floor and disagree above it.

On efficiency, no-RAG is the fastest condition for every capable model and, for the capable models, also the highest scoring. Counted at their standalone cost, the agentic tier is 3.2 times the baseline and feedback, which must produce the agentic draft before refining it, is 5.6 times. The elaborate tiers therefore carry substantial cost without a measured quality return on this benchmark.

Read as climate and energy analysis (Section 3.12), the human-preferred answers to four energy and land-use questions agree with the reference mainly on general points, such as the need for coordination, and name cross-sectoral conflicts that the reference leaves out without ranking or quantifying them. On the 2030 LULUCF commitment, four of the seven reach the opposite verdict to the reference, largely because they rely on an earlier release of the same statistic, and one builds its analysis on a contested position in the Swedish forestry debate. On Q9, the six functioning answers cover all four elements the question names, where the reference covers two. None of these differences is picked up by the instruments used in this study.

Chapter 4

Discussion

The three research questions are taken in order below, but they are not of equal weight. RQ1 and RQ2 concern a manipulation that this design cannot cleanly isolate and an effect smaller than the evaluation can resolve; RQ3, which asks which evaluation methods are valid, is where the transferable contribution lies, and Section 4.3 should be read as the substantive core of the chapter.

4.1 RQ1: Can RAG-Augmented Open-Weight Models Align with an Expert-Derived Reference?

In their best condition the six capable analyst models score within a remarkably narrow band, between 8.23 and 8.30 on the composite scale, and five of the six reach that score *without* retrieval (Section 3.2.1). On its face, this suggests that 4–35B open-weight models already support fluent, well-structured Swedish climate policy reasoning, and that the benchmark ceiling is set by the evaluation instrument rather than by either the model or corpus access: a band of 0.07 points across six models of widely differing size is more plausibly a property of the judge than of the models.

Two caveats temper that reading. First, the reference is itself machine-generated, produced by `llama-3.3-70b-instruct` reading the held-out source summary, so the evaluation rewards proximity to a fluent LLM paraphrase rather than to expert judgement, and a confident no-RAG answer is well placed to match it. This is why RQ1 is framed as alignment with an expert-*derived* reference rather than as a match to expert assessment: the underlying judgements in the held-out report are expert judgements, but the formulation against which candidates are scored is not, and the gap between the two cannot be measured from within this design. Section 3.12 shows what that gap can look like: on Q9 the reference covers only two of the four elements the question names, and those only briefly, while all six functioning answers cover all four. Second, and more decisively, the scores themselves are not a reliable ranking. Section 4.3 sets out the evidence that no instrument applied here resolves the conditions, and a benchmark score is only as meaningful as the instrument producing it.

RQ1 therefore cannot be settled from this design. Whether these systems approach expert-level analysis is a question about their agreement with expert judgement, and what was measured is their similarity to a machine-generated paraphrase of an expert document, scored by an apparatus that cannot reliably tell the candidates apart. Answering the original question requires expert-written references (Section 6.2); what can be said here is that the models produce fluent, well-structured Swedish policy prose at a level this evaluation cannot distinguish between them on.

4.2 RQ2: How Does Retrieval Architecture Affect Quality?

There is no evidence that retrieval improved judged quality under this rubric, and every effect that reaches significance runs against it. Four of the seven models show a significant within-model condition effect, every significant pairwise comparison favours no-RAG (Table 3.2), and the pooled mixed-effects model places all three RAG tiers below the baseline, none significantly. Only the two models with the weakest parametric baseline benefit at all: `granite-4.1-8b` (single-RAG, $+0.73$) and the failure-mode `gemma-4-26b-a4b@4bit` (feedback, $+1.75$), which suggests that retrieval substitutes for absent parametric knowledge rather than adding to sufficient knowledge.

The four conditions nonetheless span only 0.19 points, and the sharper statement, defended in Section 3.2.2 and again in Section 3.10.2, is that they are separated by less than this instrument can resolve. What follows is therefore an account of why an effect of that size might arise, not an explanation of an established one.

Two features of the design qualify the comparison before any mechanism is considered. The first is that the no-RAG and RAG arms differ in two respects, not one: Tier 1 receives a prompt forbidding citations, while Tiers 2 to 4 receive one requiring them (Appendix B). The manipulation is thus retrieval together with mandated citation-laden prose, set against no retrieval with citation forbidden, and this design cannot separate them. Tellingly, no-RAG's lead includes *struktur* (8.01 against 7.71) and *argumentation* (7.01 against 6.78), precisely those a reader would expect dense citation markers to depress, notwithstanding the rubric's instruction to disregard them, while the dimension on which retrieval does best, *täckning*, is the one least sensitive to register. Separating the two effects requires a fifth arm in which retrieved context is supplied under the no-RAG citation contract, and the absence of that arm is the principal internal-validity limitation of the study rather than a residual caveat.

The second concerns the reference. The judge scores candidates by their match to it, and it is fluent, uncited prose produced by a large model reading a source document without retrieval (Section 2.1.5), so the rubric rewards resemblance to exactly what the no-RAG condition pro-

duces. It was also generated from the report's summary rather than its full text, which means a candidate supplying additional correct detail can score no higher than one that does not, and retrieved answers are the ones that supply additional detail. This is not merely a ceiling on precision; it is a mechanism capable of producing the observed result on its own. The defensible claim is therefore that retrieval reduced measured similarity to an unretrieved, summary-derived reference under this rubric. Whether it reduced analytical quality is not something this design can determine.

Subject to all of that, the most likely mechanism is retrieval noise, and the direction is what the context-memory conflict literature would predict (Section 1.2.5): where parametric knowledge is already adequate, injected passages compete with it rather than supplement it, and attention over long contexts degrades as more are added [23, 25]. The tiers insert 20 chunks at single-RAG, 2 to 39 at agentic (mean 24) and 9 to 59 at feedback (mean 38), and each forces the model to phrase around them and cite them, trading confident parametric prose for hedged synthesis. Consistent with this, the agentic tier never wins a model ranking, as every additional sub-query is another chance for a tangential chunk to enter the context. The result is better described as that literature appearing in a Swedish policy setting than as a surprise. It remains an inference from the pattern of outcomes: no recall, precision or relevance annotation was computed for the retrieved chunks, so retrieval quality itself is nowhere measured in this thesis.

One case documents the mechanism concretely. Appendix D reproduces the single-RAG answer to Q3, which asks how *changed* conditions for CCS and BECCS have altered the national strategy for negative emissions. All of its citations resolve to provided chunks, so the audit certifies it as perfectly traceable; the twelve chunks it cites come from sources dated between 2013 and 2022. Every one predates the change the question asks about. The model is grounded, verifiably, in a superseded state of the world. Traceability and currency are different properties, and an audit that establishes the first says nothing about the second.

That caveat granted, citation grounding remains the one categorical benefit of retrieval. The RAG conditions emit verifiable closed-universe [Källa N] references, with traceability of 1.00 for three models and no out-of-universe references anywhere in the final run, where no-RAG can offer none (Section 3.5). The judge composite is blind to this, ranking ungrounded no-RAG highest. Notably, most of that benefit is already obtained at the cheapest retrieval tier: single-RAG achieves traceability of 1.00 for two of those three models, and 0.79 for the third, at 45% of the cost of feedback, which must produce the agentic draft before it can refine it (Table 3.8).

4.3 RQ3: Which Evaluation Methods Are Most Valid and Consistent?

The most substantive contribution of this study concerns evaluation rather than retrieval: the choice of evaluator, not the choice of retrieval tier, is the largest driver of the conclusions one would draw.

The two automated judges agree only moderately (Spearman $\rho = 0.37$) and differ by more than a full point in calibration (`microsoft/phi-4` ≈ 7.1, `nvidia/nemotron-3-nano-4b` ≈ 8.2). A single-judge design would have silently inherited whichever calibration happened to be chosen. A further channel deserves mention: the judges were held outside the analyst families to remove self-preference, but the reference itself is `llama-3.3-70b-instruct` output and scoring is match-to-reference, so stylistic proximity to that model's register is an unexamined route by which a candidate could gain score. Nothing here tests for it. The length-bias check is null ($\rho = 0.04$); indeed no-RAG produces the longest answers, so the judge's preference for it is not a verbosity artefact.

A related assumption in the judge protocol deserves to be made explicit. The rubric instructs both judges to disregard citation markers entirely (Appendix B), and the deterministic audit is conducted separately for exactly this reason. However, instructing a model to ignore a salient surface feature is not equivalent to establishing that the feature has no effect on its scoring, and no manipulation check was run to verify compliance. The no-RAG advantage on *struktur* (8.01 against 7.71) is consistent both with genuinely cleaner argument construction and with citation-laden text reading as less fluent despite the instruction. Testing this would require re-scoring citation-stripped RAG answers under the same rubric, a check that was not performed here and that should accompany any replication.

The preliminary human evaluation agrees with the judge at chance on individual items (best-pick 26%, 95% CI 16–36%, against a 25% baseline, $n = 69$). Because that interval spans the baseline in both directions, the honest reading is that the comparison detects no relationship rather than an opposed one. The condition-level aggregates do run in opposite directions (Section 3.8.2), but that comparison has four points and one rater behind it and carries no weight here.

Assembling these, the defensible conclusion is narrower than the inversion story, and stronger for it: *no evaluator used here demonstrated validity as a ranker*. Two automated judges agree at $\rho = 0.37$ and differ by more than a point in calibration; ROUGE-L sits at its floor and BERTScore inside its unrescaled baseline band; the single human rater agrees with the judge at chance. Every instrument applied to these answers fails to resolve the conditions, and the apparent ranking one obtains depends on which instrument is used. That is a claim about the

measurement apparatus rather than about human and machine taste, it requires no second rater, and it is the finding this thesis actually establishes.

Two exceptions stand out against that picture, and both are useful. The deterministic citation audit measures a real, verifiable property, grounding, that none of the graded metrics captures, and it is the most reliable single signal precisely because it is not a matter of judgement. Its limits are equally clear: as Appendix D shows, an answer can be perfectly traceable and comprehensively out of date.

The second exception is detection of the floor. The rater's twelve unusable flags fall at a median 4.7th percentile of the judge's scores, and eight of them fall on the model that the composite and the citation audit independently identify as broken. Three instruments of quite different kinds converge on the same failure case while disagreeing about everything above it. An evaluator that reliably identifies unusable output and cannot rank usable output can be used to screen but not to select, and that distinction is the practical form of the RQ3 answer.

The practical implication is that LLM-as-judge scores for open-ended policy analysis should be reported with at least two independent judges, validated against human raters before any ranking is treated as substantive, and accompanied by a deterministic grounding audit.

4.4 The `gemma-4-26b-a4b@4bit` Failure Case

At 4-bit quantisation the 26B mixture-of-experts model fell below the floor of usable output: its no-RAG composite of 5.16 sits more than two points beneath every capable model, and retrieval lifts it only to 5.95–6.91 (Section 3.11). The failure is structural rather than a uniform loss of quality, appearing as repetition loops, mid-paragraph language drift, and an inability to emit well-formed citations (traceability zero throughout).

Two interpretive points follow. First, a model can fall below the floor of this task: the analytical task is not merely performed worse but cannot be performed at all. Models in the panel with the same or fewer active parameters did not fail (Table C.1), so this floor is not set by size alone, which bears directly on resource-constrained or fully local deployment. Second, the failure is corroborated by three independent signals, the composite score, the zero-traceability audit, and the blinded human rater, who assigned eight of twelve unusable flags to this one model. That its composite still lies between 5 and 7, higher than in the pilot runs because Judge B is more lenient (Section 3.11), underlines the RQ3 caution that absolute LLM-judge scores are not comparable across judge calibrations.

4.5 Limitations

Five limitations bound what this study establishes, and they are ordered by how much they constrain the conclusions.

- **The independent variable is confounded.** Retrieval was varied together with the citation contract, so the comparison is retrieval plus mandated citation against no retrieval with citation forbidden (Section 4.2). This is the principal threat to internal validity, and a fifth arm supplying retrieved context under the no-RAG contract would resolve it.

- **The reference caps every comparison.** The gold answers are machine-generated rather than expert-written, and were produced from the held-out report's summary rather than its full text. The judge then scores candidates by their match to them, which structurally favours the register of unretrieved output and prevents a candidate supplying additional correct detail from being credited for it. The lower scores under retrieval cannot be separated from this feature of the instrument.

- **Generation variance is unmeasured.** Each cell is a single draw at temperature 0.3 with no recorded seed, so individual answers are not exactly reproducible and within-cell variance is nowhere estimated. This conditions every inferential statement in Chapter 3: with a total condition spread of 0.19 points, the effects may well sit inside the generation noise, which is consistent with the reading offered throughout but is not demonstrated. A probe of three to five samples per question for one model would settle the question.

- **The human evaluation rests on one rater, the author, across 69 items.** The final run completed roughly two weeks before submission, which left too little time to recruit independent raters for a 280-answer protocol; the rating tool accepts multiple rating sets without modification. Randomising the condition labels and stripping citation markers (Section 2.2.5) removes the mechanical route by which the rater could identify the retrieval condition, but stripping was incomplete on 13 of 70 items, all in retrieval-augmented arms, and it cannot remove the substantive tell: grounded answers name specific instruments and figures where unretrieved answers offer generic equivalents, so a rater with domain familiarity may infer the condition from specificity alone. Recomputing the ranking on the 57 unaffected items leaves it materially unchanged (Section 3.8.1). No inter-rater reliability can be computed from one rater, the rater is not a formal climate-policy expert, and the conclusions in Section 4.3 are framed so as not to depend on the human ranking.

- **The substantive comparison is small and not independent.** Section 3.12 covers 28 answers to four purposively selected questions, and the propositions were identified and coded by the same person who designed the benchmark and carried out the human rating. It shows that the patterns described occur and that no instrument used here detects them;

it does not show how common they are, and the coding is open to challenge.

Five further constraints are noted more briefly. Retrieval quality itself was never measured, so the attribution of the effect to retrieval noise is an inference from the outcome pattern rather than an observation. The Bonferroni correction controls the family-wise error rate within each analyst model but not across the panel of seven, and with ten blocks per test the procedure operates at its resolution limit (Section 2.2.3), so the per-model significances indicate a consistent direction rather than seven independent confirmations. The analyst models did not all run at the same numerical precision (Table C.1), which is a source of non-comparability between rows of the per-model tables, and the helper model that generated HyDE passages shares a checkpoint family with one panel member, so the separation between helper and analyst does not hold for `gemma-4-e4b-it`. Finally, the DiVA portion of the corpus was selected by a fixed keyword and publication-type filter, which over-includes tangential records and misses relevant documents outside the chosen terms and organisations; the corpus approximates the relevant literature rather than sampling it exhaustively.

The benchmark is ten targeted questions on Swedish climate policy as of 2026, and the findings should not be generalised beyond that scope without a broader, independently developed question set.

4.6 Ethical and Sustainability Considerations

The work bears on sustainability in two distinct ways: through the resources it consumes, and through the governance consequences of the tools it evaluates.

Counting the instrumented steps, which give a lower bound, the work behind this thesis consumed roughly 7 kWh of electricity and emitted between 0.15 and 0.66 kg CO_2e, the range depending on how manufacturing emissions are allocated (Section 3.9.1). That is at most four kilometres of car travel, and no environmental claim of substance rests on it. Three points nonetheless follow.

The first is that manufacturing dominates. Had the machine been bought new, embodied emissions would have been about three quarters of the total, so on a grid as clean as Sweden's the electricity a local machine draws is a minor term against the emissions incurred in building it, and efficiency arguments counting only energy measure the smaller half of the problem. It follows that reuse is the largest lever available to an individual researcher: this machine was bought second-hand, so under the conventional cut-off allocation the production burden sits with the first owner and this project inherits close to none of it. Extending the machine's service life removed more emissions here than any plausible software optimisation could have. Location outweighs architecture on the same logic, since the identical workload at an illustrative EU-27 average intensity of 210 gCO_2e/kWh would emit roughly ten times as much operationally.

The second is that none of this is where the interesting waste lies. Producing a feedback answer requires the agentic draft first, so its true cost is 5.6 times the unretrieved baseline, with agentic at 3.2 times, and neither returns a quality improvement measurable by any metric applied here. At ten questions the difference is immaterial. Across 100 000 answers it is an avoidable premium of roughly 11 kg CO_2e for feedback and 5.5 kg for agentic in Sweden, and about ten times those figures at the illustrative EU-27 intensity. Architectural elaboration that cannot be shown to improve output adds cost without a measured return, which argues for evaluating retrieval sophistication empirically before adopting it, on resource grounds as well as on quality grounds.

The third point is methodological. Running the pipeline locally rather than through hosted inference is what made any of this measurable: the wall-clock of every generation was recorded, the hardware was known, and the resulting figures are attributable. Centralised APIs generally expose none of that, so a study of this kind conducted through them could report latency but not energy, and could not have noticed that the feedback tier's recorded time understates its true cost by more than half.

Turning from resources to governance, climate policy evaluation informs decisions with substantial distributional consequences, and the legitimacy of those decisions rests partly on the traceability of the analysis behind them. The clearest governance-relevant result here is that the evaluation regimes examined disagree about which outputs are better, and that the automated judge nominally ranked first the condition offering no verifiable grounding at all. An organisation procuring a policy-analysis tool on the basis of automated quality scores could therefore select against traceability while believing itself to be selecting for quality. The experiment observed this pattern, if only by a margin the evaluation cannot resolve. The recommendation that follows, namely at least two independent evaluators, a human in the loop, and a deterministic grounding audit, should be read as an accountability safeguard rather than a methodological nicety.

Nothing in these results supports substituting automated analysis for expert judgement. The defensible role is narrower and more useful: a first-pass evidence synthesis in which every claim resolves to an openable source passage, produced as an auditable draft for an expert to verify and revise. Framing the technology this way keeps responsibility for the analytical judgement with an identifiable person, which matters when the analysis feeds a democratic process.

On provenance, all source material is publicly available and institutionally published, comprising government inquiries, council reports and research publications, and contains no personal data. The corpus construction is documented in full (Section 2.1.1) so that an equivalent corpus can be reassembled and the results contested. The 2026 Klimatpolitiska rådet report was excluded from the index before ingestion and used solely to derive references, and the pipeline made no external network calls, so no source material left local storage at any point. That last

property is what makes the approach usable for embargoed or pre-publication material.

On reproducibility, the corpus can be reassembled from the sources and selection profile given in Section 2.1.1, and the model configuration, prompts and decoding parameters are reproduced in Appendices B and C. Two gaps should be stated rather than glossed. Generation ran at temperature 0.3 without recorded seeds, so individual answers cannot be regenerated exactly, only resampled from the same distribution. The pipeline itself is not published alongside this report; the analysis notebooks, the saved run manifests, the judge caches and the carbon calculation script are available from the author on request, and any replication should begin from those rather than from a reimplementation of the descriptions here.

Finally, the findings are scoped to Swedish-language climate policy, ten questions, and one corpus. Presenting them as general conclusions about the utility of retrieval augmentation would be a misuse of the work, and the delimitations in Section 1.5 should be read as binding rather than formal.

Chapter 5

Implications

Chapter 4 set out what this design can and cannot show. This chapter considers what the findings mean in practice, first for bodies that carry out or commission policy analysis, and then for the climate and energy community whose questions the benchmark covers. The second section is more interpretive, and it says so where it goes beyond what the evidence shows.

5.1 Implications for Policy Investigators

Each point below starts from a result in this study and considers what it means for the work of a body such as Klimatpolitiska rådet or an SOU inquiry. More speculative directions are left to Future Work (Section 6.2).

5.1.1 What the Technology Supports Today

The technology is best used as a tool for finding and tracing evidence, not as a replacement for the analyst. The one clear benefit found under RQ2 is verifiable citation, not better analysis. A realistic near-term use is a first draft of the evidence on a question, drawn from the Swedish corpus (SOU, Klimatpolitiska rådet, DiVA), in which each claim links to a source passage that can be opened and checked. The result is a draft to be reviewed, not a finished judgement. Section 3.12 adds a caution: a citation shows where a claim came from, but not whether it is still current.

Fully local use is possible for sensitive material. The whole pipeline ran on local hardware, over a local Qdrant index of 869 133 chunks, with no external API calls, and the 2026 Klimatpolitiska rådet report was kept out of the index (Section 2.1.1). For an agency working with embargoed or unpublished material, this shows that traceable Swedish policy synthesis can be done without data leaving the building.

5.1.2 Corpus Design Decisions That Determine Output Quality

Recency needs to be built into retrieval, not left to the choice of documents. Q2, which asks for a verdict on the reduction of the biofuel blending obligation, is the question where retrieval does

worst, with single-RAG falling to 6.16 (Section 3.3). The corpus includes material from 1990 onward with no weighting for recency, which is a weakness for a question about a recent policy change. Q7 shows the same problem reversing the verdict on an EU commitment (Section 3.12). A policy-analysis corpus needs dates attached to its documents and retrieval that takes them into account. At the same time, a focused Swedish corpus keeps national policy detail that web-scale retrieval would dilute. Topic and recency are separate questions, and only the first was controlled here.

The type of source matters as much as its topic. About 88% of the corpus is academic and agency literature from DiVA, while the questions and reference answers come from a policy council's assessment. The retrieved material is therefore mostly background research, while the reference offers evaluative judgement. Matching the type of source material to the type of output wanted is as important a design choice as choosing the topic.

5.1.3 Evaluating and Procuring Policy AI

Automated quality scores are not a safe basis for choosing a tool. No evaluator used here could reliably rank the usable answers, and the automated judge nominally ranked the ungrounded condition first, with no way of noticing that it was ungrounded (Section 4.3). Tools for policy analysis should not be bought or assessed on automated quality scores alone, and any such score should be paired with a deterministic grounding check that does not depend on judgement.

Review effort should go where an error would change the conclusion. The comparison in Section 3.12 shows where to look. What passed every instrument was not invented content, but a conclusion built on an outdated statistic and a contested view used as the starting point of an answer. The first is caught by checking the date of each figure, the second by a reviewer who knows the debate. Neither is caught by looking for hallucinations.

A further caution follows from RQ3. Because changing the evaluator changes the apparent ranking, any agency using a language model to sort or score policy documents should treat a human in the loop, at least two independent evaluators and a deterministic grounding check as basic requirements, not options.

5.2 Implications for the Climate and Energy Community

The section above applies to anyone using these systems. This one is written for people who work on climate and energy policy, and asks what the results mean for how they might use AI in their own work. Each point rests on a specific result, and the limits of that evidence are stated at the end.

AI is most useful for mapping a problem, not for settling it. On Q4, the best answers listed

conflicts between goals that the reference did not mention, such as battery metals versus reindeer herding and biodiversity, biomass wanted by both transport and industry, and infrastructure planned from traffic forecasts rather than climate targets (Section 3.12). The same answers did not rank these conflicts or put numbers on them, and usually ended by recommending coordination. This suggests, though it does not prove, that such tools fit the early scoping stage of sector work, where the task is to list what a policy area contains, better than the later stage of deciding which goal should give way.

Every key number needs a date as well as a source. Energy and climate statistics are revised, and a citation does not show whether a figure is still current. On Q7, four of the seven answers and the reference reached opposite conclusions about Sweden's LULUCF commitment largely because they relied on different releases of the same statistic, and much of the difference between those releases came from recalculation rather than from change in the forest. The worked example in Appendix D shows the same issue in another form: an answer in which every citation is correct but every source predates the change the question asks about. Anyone using AI-assisted analysis in this field should record which release each key figure comes from, not only which document quotes it. This is good practice with or without AI, but AI tools make the cost of skipping it easier to see.

Agreement between AI answers is not evidence. Six of the seven answers to Q4, and the reference, agreed that the sectors need better coordination, while the points that would actually help a decision were each made by only one to three answers. Coordination appears to be what the answers fall back on when they have nothing more specific to say. Asking several models and treating their agreement as confirmation would give the most weight to the least useful conclusions.

A built-in position is harder to spot than an invented fact. The answer that would most mislead an independent assessment did not contain made-up facts. It took one side of the Swedish forestry debate as its starting point, with sources that support that side (Section 3.12), and it reads as balanced and well sourced. Only a reviewer who knows the debate would notice. For a body whose authority rests on independence, sector expertise is therefore not just helpful when reviewing AI output; it is the main thing the review needs.

Simple set-ups can be enough, and they use less energy. The agentic and feedback set-ups took 3.2 and 5.6 times as long per answer as the set-up without retrieval, with no gain in quality that any instrument here could measure, and the one benefit retrieval did bring, traceable citations, was already mostly available with the simplest retrieval set-up (Section 3.9). The recorded runs used roughly 7 kWh of electricity, a lower bound, on a single second-hand desktop computer (Section 3.9.1). For energy-sector organisations, the practical point is that private, traceable AI analysis can run on modest local hardware, and that more complex options should be tested against simple ones before they are adopted.

The sector needs its own test material before it can trust these tools. On Q9, the reference answer, produced by a large language model from a report summary, covered only two of the four elements the question named, briefly and without priorities, while several model answers covered all of it, yet every automated score in this thesis is measured against that reference. A good reference answer reflects a judgement about what a question is really asking, which is expert work. If energy agencies and research groups want to know whether these tools can do their kind of analysis, they will need reference answers written by their own experts, and ways of scoring that reward specific, well-sourced content rather than resemblance to a text.

These points rest on a small comparison: 28 answers to four of ten questions, one condition per model, from a single generation run, read and coded by the author, who also designed the benchmark and rated the answers. They show that these patterns occur and that the instruments used here do not detect them. They do not show how common the patterns are, do not rank the models on domain quality, and say nothing about systems larger or newer than the open-weight models tested here.

Chapter 6

Conclusions & Future Work

6.1 Conclusions

This study asked whether openly available language models, given access to Swedish climate policy documents, can produce analysis that agrees with an expert-derived reference, whether the way documents are retrieved makes a difference, and which evaluation methods can tell good analysis from weak. Seven models answered ten questions in four retrieval set-ups, and the 280 answers were scored against references generated from the summary of the held-out 2026 Klimatpolitiska rådet report, by text-similarity measures, a citation check, two language models acting as judges and a blinded human rater. Answers to four energy and land-use questions were also read as climate and energy analysis.

Four conclusions follow, the first most firmly supported and the second most relevant to climate and energy work.

First, no evaluation method could reliably rank the usable answers. The two automated judges agreed only moderately, the text-similarity measures barely moved, and the single human rater preferred different set-ups than the judge and agreed with it no better than chance on individual answers. Which set-up looks best therefore depends on which method does the scoring, and that dependence is at least as large as any difference the set-ups produced. One task was done reliably: the rater, the judge and the citation check agreed on the unusable output, with eight of the rater's twelve flags falling on the one model the other methods also identified as failing. Such methods can screen out poor output but not select the best.

Second, read as climate and energy analysis, the answers were better at mapping a policy problem than at settling it, and their most important weaknesses were ones no score detected. On the cross-sectoral question the best answers named conflicts the reference left out, such as battery metals against reindeer herding and biomass wanted by both transport and industry, but did not rank or quantify them, and the answers agreed most on the least useful recommendation, better coordination. On Sweden's 2030 commitment for the land-use carbon sink (LULUCF), four of the seven answers and the reference reached opposite verdicts, largely because they relied on different releases of the same statistic. Since the sink affects how much reduction transport and

heating must deliver, the release an analysis uses changes what it implies for the energy system. One answer took a contested position in the Swedish forestry debate as its starting point while reading as balanced. Neither problem involved invented facts. This reading covers 28 answers coded by the author, so it shows that these patterns occur, not how common they are.

Third, retrieval brought no measurable gain in scored quality, and the more complex set-ups cost more. The small lead of the set-up without retrieval did not hold in the random subsample re-scored to test the language of the judges' instructions, and two design features rule out a causal reading: retrieval was changed together with the instruction to cite sources, and the references were machine-generated from a summary, which favours text written without retrieval. The one clear benefit of retrieval was citations traceable to a source passage, although such a citation does not show whether a figure is still current (Appendix D). The agentic set-up, which splits a question into sub-questions, took 3.2 times as long per answer as no retrieval, and the feedback set-up, which also critiques and revises its draft, 5.6 times, with no measurable gain.

Fourth, this design cannot settle whether such models approach expert-level analysis, because the automated scores measure similarity to a machine-generated reference and the single human rater compared the answers only with each other. On the electricity-system question, several answers covered what the question asked more fully than the reference did. Judging these tools fairly requires reference answers written by experts. For the climate and energy community, the results suggest that such tools suit scoping a policy area and tracing its evidence, with every key statistic dated and the result reviewed by a sector expert, and do not replace expert judgement. The broader contribution is a reproducible framework for evaluating AI-assisted policy analysis, and the finding that the choice of evaluator within it is not a technicality but decides which answer looks best.

6.2 Future Work

Several directions follow directly from the findings.

The most immediate is to add independent raters, ideally with formal climate-policy expertise, compute inter-rater reliability (Cohen's κ on dimension bands and Spearman's ρ on rankings), and re-estimate judge calibration against the pooled human ratings. Only this can show whether human and automated evaluators diverge systematically or simply fail to agree. The rating instrument should also remove the citation marker forms missed here (Section 2.2.5). The same raters could extend the comparison in Section 3.12 to all ten questions and all four set-ups, and independently check the coding of propositions, which was done here by the author alone.

The second is better reference material. Reference answers written by climate and energy experts, or at least generated from the full held-out report rather than its summary, would remove the ceiling that the present references place on every comparison, and would test whether the

nominal advantage of the set-up without retrieval is partly an artefact of a fluent machine-written reference. Scoring should also reward specific, well-sourced content rather than resemblance to a text, for example through a judge rubric that explicitly rewards verifiable grounding.

Before further retrieval set-ups are tested, the variation between repeated generation runs should be measured. Several samples per model, question and set-up, even for one model and two set-ups, would show whether the small differences between set-ups lie inside or outside that noise. If they lie outside it, the most valuable addition is a fifth set-up that supplies retrieved context under the no-retrieval citation contract, separating the effect of retrieval from that of the citation requirement, followed by re-scoring the citation-stripped answers to test whether the judges disregard citation markers as instructed. If they lie inside it, effort is better spent on the evaluation methods than on further set-ups.

Retrieval quality was never measured. Manual annotation of the relevance and date of retrieved passages would turn the retrieval-noise account from inference into evidence and test whether older material displaces current figures, as documented in Appendix D and on the carbon sink question (Section 3.12). A corpus with dates attached to its documents, and retrieval that takes them into account, follows directly from this, and reranking, hybrid sparse-dense retrieval and tighter chunk budgets could then be tested against it. The feedback set-up's self-identified evidence gaps could also be used to map where the evidence base for a policy area is thin. A broader, independently developed benchmark and full-precision inference would test whether the findings hold beyond these ten questions and the quantised models used here.

The clearest direction for follow-on work, and the one this thesis was built to enable, is to move retrieval beyond semantic similarity towards documented policy effectiveness. An evaluator does not need the passage most similar to the query but the one most informative about which instruments actually lower emissions, which only a small fraction do at scale (63 of 1 500 in [4]). A planned continuation couples a localised neural network that estimates a draft policy's emission impact with iterative RAG that grounds and contextualises that estimate, extending the logic of this study from matching expert text towards forecasting achievable impact.

Bibliography

[1] UNFCCC. "The Paris Agreement". In: COP 21, 2015. URL: https://unfccc.int/documents/184656.

[2] IPCC. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Tech. rep. Geneva, Switzerland: Intergovernmental Panel on Climate Change, July 2023. URL: https://www.ipcc.ch/report/ar6/syr/ (visited on 11/10/2025).

[3] Carl-Friedrich Schleussner, Gaurav Ganti, Quentin Lejeune, Biqing Zhu, Peter Pfleiderer, Ruben Prütz, Philippe Ciais, Thomas L. Frölicher, Sabine Fuss, Thomas Gasser, Matthew J. Gidden, Chahan M. Kropf, Fabrice Lacroix, Robin Lamboll, Rosanne Martyr, Fabien Maussion, Jamie W. McCaughey, Malte Meinshausen, Matthias Mengel, Zebedee Nicholls, Yann Quilcaille, Benjamin Sanderson, Sonia I. Seneviratne, Jana Sillmann, Christopher J. Smith, Norman J. Steinert, Emily Theokritoff, Rachel Warren, Jeff Price, and Joeri Rogelj. "Overconfidence in climate overshoot". *Nature* 634.8033 (Oct. 2024), pp. 366–373. ISSN: 1476-4687. URL: https://www.nature.com/articles/s41586-024-08020-9 (visited on 11/10/2025).

[4] Annika Stechemesser, Nicolas Koch, Ebba Mark, Elina Dilger, Patrick Klösel, Laura Menicacci, Daniel Nachtigall, Felix Pretis, Nolan Ritter, Moritz Schwarz, Helena Vossen, and Anna Wenzel. "Climate policies that achieved major emission reductions: Global evidence from two decades". *Science* 385.6711 (Aug. 2024), pp. 884–892. URL: https://www.science.org/doi/10.1126/science.adl6547 (visited on 12/17/2025).

[5] OECD. *AI in policy evaluation: Governing with Artificial Intelligence*. Sept. 2025. URL: https://www.oecd.org/en/publications/governing-with-artificial-intelligence_795de142-en/full-report/ai-in-policy-evaluation_c88cc2fd.html (visited on 12/17/2025).

[6] T. Suresh Balakrishnan, Prabhakar Krishnan, U. Samson Ebenezar, M. Mohammed Nizarudeen, and N. Kamal. "Machine Learning for Climate Change Impact Assessment and Adaptation Planning". In: *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*. Mar. 2024, pp. 1–6. URL: https://ieeexplore.ieee.org/document/10545291 (visited on 12/17/2025).

[7] Harshita Jain, Renu Dhupper, Anamika Shrivastava, Deepak Kumar, and Maya Kumari. "AI-enabled strategies for climate change adaptation: protecting communities, infrastructure, and businesses from the impacts of climate change". *Computational Urban Science* 3.1 (July 2023), p. 25. ISSN: 2730-6852. URL: https://doi.org/10.1007/s43762-023-00100-2 (visited on 12/17/2025).

[8] Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Hübscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauß. *Assessing Large Language Models on Climate Information*. arXiv:2310.02932 [cs]. May 2024. URL: http://arxiv.org/abs/2310.02932 (visited on 12/17/2025).

[9] Kira Kappe and Adis Dzebo. *Can AI produce reliable and consistent data analysis?* Tech. rep. Stockholm Environment Institute, Aug. 2025. URL: https://www.sei.org/publications/can-ai-produce-reliable-consistent-data-analysis/ (visited on 12/17/2025).

[10] Filippo Maria D'Arcangelo, Tobias Kruse, and Mauro Pisu. "Identifying and tracking climate change mitigation strategies with a cluster-based assessment". *npj Climate Action* 3.1 (Oct. 2024), p. 86. ISSN: 2731-9814. URL: https://www.nature.com/articles/s44168-024-00158-6 (visited on 02/18/2026).

[11] Di Wang, Lijing Chen, and Liang Dong. "A critical review of climate change mitigation policies in the EU ——based on vertical, horizontal and policy instrument perspectives". *Journal of Cleaner Production* 467 (Aug. 2024), p. 142972. ISSN: 0959-6526. URL: https://www.sciencedirect.com/science/article/pii/S0959652624024211 (visited on 02/18/2026).

[12] Eskinder Belete. *AI-Powered Climate Policy Analysis Using Natural Language Processing and Large Language Models*. Zenodo. Mar. 2025. URL: https://zenodo.org/doi/10.5281/zenodo.18159387.

[13] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/ (visited on 05/08/2026).

[14] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. *BERTScore: Evaluating Text Generation with BERT*. Apr. 2019. URL: https://arxiv.org/abs/1904.09675v3 (visited on 05/08/2026).

[15] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*.

arXiv:2306.05685 [cs]. Dec. 2023. URL: http://arxiv.org/abs/2306.05685 (visited on 05/08/2026).

[16] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. Mar. 2023. URL: https://arxiv.org/abs/2303.16634v3 (visited on 05/08/2026).

[17] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. *Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators*. Apr. 2024. URL: https://arxiv.org/abs/2404.04475v2 (visited on 05/08/2026).

[18] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. "RAGAs: Automated Evaluation of Retrieval Augmented Generation". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/.

[19] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 338–354. URL: https://aclanthology.org/2024.naacl-long.20/.

[20] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. *Retrieval-Augmented Generation for Large Language Models: A Survey*. arXiv:2312.10997 [cs]. Mar. 2024. URL: http://arxiv.org/abs/2312.10997 (visited on 02/24/2026).

[21] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. "Context-faithful Prompting for Large Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14544–14556. URL: https://aclanthology.org/2023.findings-emnlp.968/ (visited on 02/24/2026).

[22] Michael R. King. "An Update on AI Hallucinations: Not as Bad as You Remember or as You've Been Told". *Cellular and Molecular Bioengineering* 18.6 (Dec. 2025), pp. 543–548. ISSN: 1865-5033. URL: https://doi.org/10.1007/s12195-025-00874-x (visited on 02/23/2026).

[23] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. *Knowledge Conflicts for LLMs: A Survey*. arXiv:2403.08319 [cs]. 2024. URL: http://arxiv.org/abs/2403.08319.

[24] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. "Entity-Based Knowledge Conflicts in Question Answering". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 7052–7063. URL: https://aclanthology.org/2021.emnlp-main.565/.

[25] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. "Lost in the Middle: How Language Models Use Long Contexts". *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 157–173.

[26] Regeringen och Regeringskansliet. *Statens offentliga utredningar*. Text. Sept. 2014. URL: https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/ (visited on 07/07/2026).

[27] Klimatpolitiska rådet. *Publikationer*. URL: https://www.klimatpolitiskaradet.se/publikationer/ (visited on 07/07/2026).

[28] *Digitala Vetenskapliga Arkivet*. URL: https://www.diva-portal.org/ (visited on 07/07/2026).

[29] Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. *Docling Technical Report*. Aug. 2024. URL: https://arxiv.org/abs/2408.09869v5 (visited on 05/10/2026).

[30] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. arXiv:2402.03216 [cs]. Dec. 2025. URL: http://arxiv.org/abs/2402.03216 (visited on 05/10/2026).

[31] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran Gv, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri

Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. *MMTEB: Massive Multilingual Text Embedding Benchmark*. Feb. 2025. URL: https://arxiv.org/abs/2502.13595v4 (visited on 04/06/2026).

[32] Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. *The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding*. June 2024. URL: https://arxiv.org/abs/2406.02396v1 (visited on 04/06/2026).

[33] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. *Precise Zero-Shot Dense Retrieval without Relevance Labels*. arXiv:2212.10496 [cs]. Dec. 2022. URL: http://arxiv.org/abs/2212.10496.

[34] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. *Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions*. arXiv:2212.10509 [cs.CL]. June 2023. URL: http://arxiv.org/abs/2212.10509 (visited on 06/23/2026).

[35] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. *Reflexion: Language Agents with Verbal Reinforcement Learning*. arXiv:2303.11366 [cs]. Oct. 2023. URL: http://arxiv.org/abs/2303.11366 (visited on 05/08/2026).

[36] Klimatpolitiska rådet. *Årsrapport 2026*. Tech. rep. 2026-00005/K. Klimatpolitiska rådet, Mar. 2026. URL: https://www.klimatpolitiskaradet.se/rapport-2026/ (visited on 05/08/2026).

[37] Milton Friedman. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance". *Journal of the American Statistical Association* 32.200 (Dec. 1937), pp. 675–701. ISSN: 0162-1459. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522 (visited on 07/07/2026).

[38] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". *Biometrics Bulletin* 1.6 (1945), pp. 80–83. ISSN: 0099-4987. URL: https://www.jstor.org/stable/3001968 (visited on 07/07/2026).

[39] Olive Jean Dunn. "Multiple Comparisons among Means". *Journal of the American Statistical Association* 56.293 (Mar. 1961), pp. 52–64. ISSN: 0162-1459. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1961.10482090 (visited on 07/07/2026).

[40] Richard A. Armstrong. “When to use the Bonferroni correction”. *Ophthalmic and Physiological Optics* 34.5 (2014), pp. 502–508. ISSN: 1475-1313. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/opo.12131 (visited on 07/07/2026).

[41] Catherine O. Fritz, Peter E. Morris, and Jennifer J. Richler. “Effect size estimates: Current use, calculations, and interpretation”. *Journal of Experimental Psychology: General* 141.1 (2012), pp. 2–18. ISSN: 1939-2222.

[42] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge, May 2013. ISBN: 978-0-203-77158-7.

[43] Loïc Lannelongue, Jason Grealey, and Michael Inouye. “Green Algorithms: Quantifying the Carbon Footprint of Computation”. *Advanced Science* 8.12 (2021), p. 2100707.

[44] Green Software Foundation. *Software Carbon Intensity (SCI) Specification, Version 1.1*. ISO/IEC 21031:2024. 2024. URL: https://sci.greensoftware.foundation/.

[45] Apple Inc. *Mac Studio power consumption and thermal output (BTU) information*. Wall-measured power, M1 Max 10-core CPU / 32-core GPU configuration. 2022. URL: https://support.apple.com/en-us/102027 (visited on 08/07/2026).

[46] Electricity Maps. *Electricity Grid Review 2025: Sweden*. Flow-traced (consumption-based) carbon intensity. 2026. URL: https://www.electricitymaps.com/grid-in-review-2025/sweden (visited on 08/07/2026).

[47] Apple Inc. *Product Environmental Report: Mac Studio*. Tech. rep. Life-cycle assessment per ISO 14040/14044; M1 Max, 32 GB, 512 GB configuration. Apple Inc., Mar. 2022. URL: https://www.apple.com/environment/pdf/products/desktops/Mac_Studio_PER_March2022.pdf (visited on 08/07/2026).

[48] European Parliament and Council of the European Union. *Regulation (EU) 2023/839 of the European Parliament and of the Council of 19 April 2023 amending Regulation (EU) 2018/841 as regards the scope, simplifying the reporting and compliance rules, and setting out the targets of the Member States for 2030, and Regulation (EU) 2018/1999 as regards improvement in monitoring, reporting, tracking of progress and review*. Official Journal of the European Union. OJ L 107, 21.4.2023, pp. 1–28. 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/839/oj.

[49] Naturvårdsverket. *Nettoutsläpp och nettoupptag av växthusgaser från markanvändning (LULUCF)*. Official statistics for 2023, page version reviewed 18 June 2025, archived by the Internet Archive. 2025. URL: https://web.archive.org/web/20250620170546/https://www.naturvardsverket.se/data-och-statistik/klimat/vaxthusgaser-nettoutslapp-och-nettoupptag-fran-markanvandning/ (visited on 09/15/2026).

[50] Naturvårdsverket. *Nettoutsläpp och nettoupptag av växthusgaser från markanvändning (LULUCF)*. Official statistics for 2024, published December 2025. 2025. URL: https://www.naturvardsverket.se/data-och-statistik/klimat/vaxthusgaser-nettoutslapp-och-nettoupptag-fran-markanvandning/ (visited on 09/15/2026).

Appendix Contents

A Evaluative Questions **1**

B Prompts **3**

B.1 Analyst system prompt (shared base) 3

B.2 HyDE prompt (Tier 2) 4

B.3 Iterative query prompt (Tier 3) 4

B.4 Feedback prompts (Tier 4) 4

B.5 Gold-standard generation prompt 5

B.6 LLM-as-judge prompt 5

C Model and Run Configuration **7**

C.1 Supplementary Result Tables 7

D Worked Example: Human–Judge Divergence **9**

D.1 Citation Traceability 13

E Full Statistical Results **18**

Chapter A

Evaluative Questions

Following are the ten questions shared by all models and conditions. The questions are written in Swedish to ensure alignment with the gold standard and to avoid translation losses.

1. Temporal and Evaluative Reasoning

   "Utvärdera den svenska regeringens samlade klimatpolitik under mandatperioden fram till valet 2026 baserat på oberoende expertis, och identifiera vilka huvudsakliga bristerna i den klimathandlingsplan som presenterades 2023 var."

2. Adversarial and Normative Judgement

   "Var det korrekt att sänka reduktionsplikten? Analysera effekterna av den sänkta reduktionsplikten på biodrivmedelsmarknaden och Sveriges klimatmål, inklusive eventuella investeringsstopp och konkursrisker."

3. Causal and Mechanistic Reasoning

   "Hur har ändrade förutsättningar för CCS/BECCS-teknik i Sverige påverkat den nationella strategin för negativa utsläpp, och vilka alternativa kolsänkor diskuteras?"

4. Cross-Domain Synthesis

   "Gör en tvärsektoriell analys av hur transport-, energi- och industripolitiken samverkar (eller motverkar varandra) i Sveriges försök att nå nettonollutsläpp till 2045."

5. Multi-Level Governance

   "Analysera samspelet mellan EU:s klimatramverk (Fit for 55, CBAM, ESR) och svensk nationell klimatpolitik. Hur påverkar EU-reglering det nationella handlingsutrymmet och kommunernas klimatomställningsarbete?"

6. Security-Climate Nexus

   “Hur kan och bör Sverige integrera klimatomställningen med den pågående förstärkningen av totalförsvaret? Identifiera de viktigaste synergierna och målkonflikterna samt ge konkreta rekommendationer för nästa mandatperiod.”

7. Land Use and Carbon Sinks

   “Redogör för läget för Sveriges markanvändningssektor (LULUCF), bedöm möjligheterna att nå EU-åtagandet till 2030, och beskriv vilka styrmedel som behövs för att säkerställa ett högt nettoupptag av växthusgaser i skog och mark.”

8. Sectoral Governance (Agriculture)

   “Analysera de centrala bristerna i Sveriges nuvarande jordbrukspolitik ur klimatperspektiv och formulera en strategi inklusive målbild, styrmedel och förutsättningsskapande åtgärder för jordbrukets bidrag till klimatmålen.”

9. Energy System Reasoning

   “Gör en sammanhållen analys av förutsättningarna för ett fossilfritt elsystem i Sverige: kärnkraftens roll, vindkraftens utbyggnadshinder, energieffektiviseringens potential och behovet av elnätskapacitet. Vilka är de viktigaste prioriteringarna?”

10. Cross-Sectoral Bioeconomy Reasoning

    “Varför behöver Sverige en nationell bioekonomistrategi, och hur bör konkurrensen om biomassa från skog och mark balanseras med hänsyn till kolupptag, biologisk mångfald, biodrivmedel, biobaserade material och koldioxidinfångning?”

Chapter B

Prompts

All prompts were issued in Swedish. The analyst system prompt and the four-section answer schema (*kontext*, *analys*, *implikationer*, *rekommendationer*) were held constant across the four retrieval conditions; only the citation contract differed between the no-RAG and RAG variants. Placeholders such as `{question}`, `{context}`, and `{draft}` are substituted at run time.

B.1 Analyst system prompt (shared base)

```
Du är en oberoende svensk klimat- och miljöpolitisk analytiker.

SPRÅK: Skriv hela svaret på svenska.

FORMAT: Svara med ett JSON-objekt med exakt dessa fyra nycklar:
  "kontext"          - bakgrund och frågans relevans (100-150 ord)
  "analys"           - huvudslutsats (150-200 ord)
  "implikationer"    - vad detta innebär för svensk klimatpolitik (100-150 ord)
  "rekommendationer" - konkreta policyförslag (100-150 ord)
```

No-RAG citation contract (appended to the base):

```
KÄLLHÄNVISNINGAR: Du har INGA externa källor. Skriv INGA (Författare, År)-
parenteser och INGA [Källa N]-taggar. Avstå hellre från ett påstående än att
uppfinna en källa.

STIL: Expertanalys i offentlig utredning. Saklig och neutral.
```

RAG citation contract (appended to the base; used in Tiers 2–4):

```
KÄLLHÄNVISNINGAR: Använd ENDAST [Källa N] där N matchar en tillhandahållen
källa. Använd ALDRIG (Författare, År)-stil. Hitta INTE på källor. Påstå bara
det som faktiskt står i källorna.

STIL: Expertanalys i offentlig utredning. Saklig och neutral.
```

The RAG user message is `"KÄLLOR:\n{context}\n\nFRÅGA: {question}"`; the no-RAG user message is the question alone.

B.2 HyDE prompt (Tier 2)

System: "Svensk klimatpolitisk expert. Specifikt, faktabaserat."

```
Skriv ett kortfattat, faktabaserat hypotetiskt svar (3-4 meningar) på frågan.
Lägg svaret i 'hypotetiskt_svar'. INGA tankesteg, ingen meta-text.

FRÅGA: {question}
```

B.3 Iterative query prompt (Tier 3)

System: "Informationsspecialist. Korta, distinkta svenska sökfraser." Issued once per round; {prev} lists queries already used and {digest} is a one-line summary of chunks retrieved so far.

```
FRÅGA:
{question}

REDAN ANVÄNDA SÖKFRASER:
{prev}

HITTILLS HITTAT:
{digest}

Formulera EN ny svensk sökfras som täcker en aspekt av frågan som ännu inte är
välbelagd i hittills hittade källor. Inga upprepningar. Lägg endast själva
sökfrasen i 'sökfras'.
```

B.4 Feedback prompts (Tier 4)

Gap identification (system: "Informationsspecialist. Korta svenska sökfraser."):

```
Utkastet nedan besvarar en fråga. Föreslå {N} svenska sökfraser som hämtar
information om aspekter utkastet saknar. Endast korta sökfraser (max 12 ord) -
inte rubriker, inte instruktioner, inte citat från utkastet, inte engelska.

FRÅGA: {question}

UTKAST:
{draft}
```

Refinement (system: the RAG citation contract above):

```
Förfina denna policyanalys med nytt bevismaterial. Du får BEHÅLLA befintliga
[Källa N]-taggar och LÄGGA TILL nya. Hitta INTE på källor utöver de
```

```
tillhandahållna.

UTKAST:
{draft}

YTTERLIGARE BEVIS:
{refine_ctx}

FRÅGA: {question}
```

B.5 Gold-standard generation prompt

Run once per question by `llama-3.3-70b-instruct` over the *Sammanfattning* of the held-out 2026 report (temperature 0.1).

```
Du är en analytiker som sammanfattar Klimatpolitiska rådets årsrapport 2026.
Svara ENBART baserat på rapporten du fått som kontext. INGA källhänvisningar -
rapporten är den enda källan. Inga [Källa]-markeringar. Inga parenteshänvisningar.

Skriv på svenska. Svara med ett JSON-objekt med dessa fyra nycklar:
  "kontext"          - bakgrund och frågans relevans (100-150 ord)
  "analys"           - rapportens huvudslutsatser (150-200 ord)
  "implikationer"    - vad detta innebär för klimatpolitiken (100-150 ord)
  "rekommendationer" - rådets konkreta rekommendationer (100-150 ord)

STIL: Objektiv expertanalys. Inga värderingsord.
```

User message: `"RAPPORT:\n{report}\n\nFRÅGA: {question}"`.

B.6 LLM-as-judge prompt

Each judge (`microsoft/phi-4`, `nvidia/nemotron-3-nano-4b`) scores one candidate at a time against the gold reference, temperature 0.1.

```
Du är en strikt akademisk granskare av policyanalys.
Jämför KANDIDATSVARET med REFERENSSVARET.
Ge poäng 1-10 på fem dimensioner:

  täckning      - täcker kandidaten referensens nyckelpoänger?
  korrekthet    - stämmer faktapåståendena med referensen?
  argumentation - är resonemanget logiskt sammanhängande och stödjer
                  slutsatserna stegvis sina premisser?
  struktur      - logiskt uppbyggt och lätt att följa?
  policyinsikt  - djup förståelse för policykontext och konsekvenser?
```

```
REGLER:
1. Poäng baseras enbart på matchning mot REFERENSEN.
2. motivering: 2-3 meningar på svenska.
3. IGNORERA källhänvisningar helt - varken [Källa N] eller (Författare, År)
   ska påverka något betyg.
```

User message: `"FRÅGA: {question}\n\nREFERENSSVAR: {gold}\n\nKANDIDATSVAR: {answer}"`.

Chapter C

Model and Run Configuration

Table C.1: Analyst panel (the experimental subject). "Active" is the number of parameters engaged per token for mixture-of-experts (MoE) models. Quantisation is reported as encoded in the LM Studio model identifier; where the identifier carries no precision suffix the model ran at the default precision of the published build, which was not recorded at run time. The mixture of precisions is a source of non-comparability between rows and is noted in Section 4.5. The helper, judge, and gold-standard models are held distinct from this panel.

Model	Family	Type	Params (active)	Quantisation
gemma-4-e4b-it	Gemma	Dense	4B (4B)	Not recorded
granite-4.1-8b	Granite	Dense	8B (8B)	Not recorded
gemma-4-31b-it@4bit	Gemma	Dense	31B (31B)	4-bit
granite-4.1-30b	Granite	MoE	30B (3B)	Not recorded
qwen3.6-27b-ud-mlx	Qwen	Dense	27B (27B)	Unsloth dynamic, MLX
qwen3.6-35b-a3b-ud-mlx	Qwen	MoE	35B (3B)	Unsloth dynamic, MLX
gemma-4-26b-a4b@4bit	Gemma	MoE	26B (4B)	4-bit

Table C.2: Pipeline and run parameters. All inference ran locally through LM Studio on Apple Silicon.

Component	Setting
Helper model (HyDE, query generation)	`google/gemma-4-e4b`
Judge A / Judge B	`microsoft/phi-4` / `nvidia/nemotron-3-nano-4b`
Gold-standard model	`llama-3.3-70b-instruct`
Embedding model	`BAAI/bge-m3` (1024-dim, half precision)
Vector store	Qdrant, 869 133 indexed chunks
Document parser	Docling
Chunking (target / max / min chars)	900 / 1 400 / 220
Tier 2 retrieval (top-k)	20
Tier 3 retrieval	up to 5 queries $\times$ 10 chunks
Tier 4 retrieval	Tier 3 pool + up to 4 gap queries $\times$ 10 chunks
Temperature (analyst / judge / gold)	0.3 / 0.1 / 0.1
Temperature (HyDE / query / gap)	0.3 / 0.4 / 0.3

C.1 Supplementary Result Tables

Table C.3: BERTScore F1 by model and condition. Differences of this magnitude are not interpretable as differences in semantic fidelity; see Section 3.6. No condition is marked as leading, because the metric does not support such a reading.

Model	No-RAG	Single-RAG	Agentic	Feedback
gemma-4-e4b-it	0.8588	0.8632	0.8612	0.8609
granite-4.1-8b	0.8620	0.8650	0.8652	0.8655
gemma-4-31b-it@4bit	0.8598	0.8604	0.8618	0.8575
granite-4.1-30b	0.8590	0.8592	0.8559	0.8539
qwen3.6-27b-ud-mlx	0.8686	0.8682	0.8680	0.8671
qwen3.6-35b-a3b-ud-mlx	0.8623	0.8557	0.8654	0.8646
gemma-4-26b-a4b@4bit	0.8486	0.8717	0.8618	0.8670

Table C.4: Mean wall-clock time per answer (seconds) by model and condition, as recorded. The feedback column is the *increment* over the agentic draft, which that tier reuses rather than regenerating; the standalone cost of a feedback answer is the agentic and feedback columns summed. Section 3.9 explains the distinction and Table 3.8 reports the standalone figures.

Model	No-RAG	Single-RAG	Agentic	Feedback
gemma-4-e4b-it	33.5	54.3	62.3	52.7
granite-4.1-8b	52.4	126.7	145.1	100.8
gemma-4-31b-it@4bit	81.4	228.6	297.7	224.6
granite-4.1-30b	67.7	230.3	308.7	214.8
qwen3.6-27b-ud-mlx	83.7	200.9	256.0	185.6
qwen3.6-35b-a3b-ud-mlx	37.6	53.9	76.9	60.1
gemma-4-26b-a4b@4bit	161.4	46.0	125.8	70.5

Chapter D

Worked Example: Human–Judge Divergence

This appendix presents the single item with the largest disagreement between the human rater and the automated judge among the capable models: `gemma-4-31b-it@4bit` on Q3 (the CCS/BECCS question, Appendix A). It illustrates both RQ3 findings at once. First, the human rater's *best* pick (Single-RAG) is the automated judge's *lowest*-scored condition. Second, the two judges disagree sharply on the retrieval-augmented answers: Judge A (the stricter `microsoft/phi-4`) scores Single-RAG near the floor while Judge B (`nvidia/nemotron-3-nano-4b`) scores it near the ceiling, on the identical text.

The answers are reproduced below with their [Källa N] markers intact, so that the citation-traceability analysis in Section D.1 can be followed. The versions presented to the human rater had these markers removed, with the exceptions documented in Section 2.2.5.

Table D.1: Per-condition scores for the worked example (gemma-4-31b-it@4bit, Q3). Each judge value is that judge's mean over the five dimensions; the composite is the mean of the two judges. The human rater's best and worst picks are shown in the final column.

Condition	Judge A (phi-4)	Judge B (nemotron)	Composite	Human pick
No-RAG	7.6	8.6	8.10	–
Single-RAG	2.6	9.0	5.80	best
Agentic	3.4	9.0	6.20	worst
Feedback	5.6	9.0	7.30	–

The two judges' five-dimension scores and rationales were as follows.

No-RAG. *Judge A* (täckning 7, korrekthet 8, argumentation 9, struktur 8, policyinsikt 6): "Kandidatdokumentet täcker flera nyckelpoänger från referensen, såsom den omvärderade strategin för negativa utsläpp och diskussionen av alternativa kolsänkor. Det stämmer överens med referensen i att betona en diversifierad riskhantering och inkludering av naturbaserade lösningar. Argumentationen är logiskt sammanhängande, med tydliga premisser som leder till slutsatser om behovet av en hybridmodell för kolsänkor. Strukturen i kandidatdokumentet är väl organiserad och lätt att följa. Dock saknas djupare policyinsikt specifikt relaterad till den bredare

EU-kontexten och de politiska rekommendationerna från referensen, vilket begränsar poängen på denna dimension." *Judge B* (täckning 9, korrekthet 10, argumentation 8, struktur 7, policyinsikt 9): "Kandidatens svar täcker de viktigaste nyckelpoängerna i referensen, inklusive förändrade förutsättningar och alternativa kolsänkor. Faktapåståenden är korrekta och stämmer med kontexten."

Single-RAG. *Judge A* (täckning 3, korrekthet 2, argumentation 2, struktur 4, policyinsikt 2): "Kandidatdokumentet täcker vissa aspekter av CCS/BECCS-teknikens roll i Sveriges strategi för negativa utsläpp och diskuterar alternativa kolsänkor. Det inkluderar information om bio-CCS, naturliga sänkor och andra teknologier som DACCS. Dock saknas en direkt koppling till den nationella strategin för att nå klimatmålen 2030 och 2040, vilket är centralt i referensdokumentet. Kandidatdokumentet ger inte heller en analys av bristerna i nuvarande politik eller rekommendationer om hur regeringen ska hantera dessa utmaningar, vilket begränsar dess täckning och korrekthet. Argumentationen är logisk men saknar den bredare kontexten av klimatpolitiska rådets årsrapport 2026. Strukturen i kandidatdokumentet är välorganiserad, med tydliga sektioner för kontext, analys och rekommendationer. Policyinsikten är begränsad eftersom det inte adresserar de specifika bristerna eller styrningsrekommendationerna som nämns i referensdokumentet." *Judge B* (täckning 9, korrekthet 10, argumentation 8, struktur 9, policyinsikt 9): "Kandidatens svar täcker de viktigaste nyckelpoängerna i referensen, inklusive CCS/BECCS-teknikens roll, negativa utsläppsmål till 2045 och alternativa kolsänkor som skogar och mark. Faktapåståenden är korrekta, särskilt angående bio-CCS som permanent negativt utsläpp ger och att kompletterande åtgärder får begränsad del av måluppfyllelsen. Argumentationen är logiskt sammanhängande men kan förbättras genom tydligare koppling mellan tekniska lösningar och politisk styrning."

Agentic. *Judge A* (täckning 3, korrekthet 2, argumentation 4, struktur 5, policyinsikt 3): "Kandidatdokumentet täcker vissa aspekter av CCS/BECCS-teknikens roll i Sveriges klimatpolitik men saknar en direkt koppling till den nationella strategin för negativa utsläpp som beskrivs i referensen. Det stämmer överens med referensen när det gäller betydelsen av BECCS och behovet av ekonomiska incitament, men missar att nämna de specifika bristerna i den nuvarande politiken eller rekommendationerna från Klimatpolitiska rådet. Argumentationen är logiskt strukturerad med tydliga implikationer och rekommendationer, även om det saknas en direkt koppling till de kritiska bristerna i den nuvarande politiken som referensen betonar. Strukturen är väl organiserad och lätt att följa, men policyinsikten är begränsad eftersom det inte fullt ut adresserar de bredare strategiska rekommendationerna från Klimatpolitiska rådet." *Judge B* (täckning 9, korrekthet 8, argumentation 10, struktur 9, policyinsikt 9): "Kandidatens svar täcker de viktigaste nyckelpoängerna i referensen, inklusive mål om nollutsläpp 2045, rollen för BECCS och alternativa kolsänkor. Faktapåståenden är nästan exakta, men det finns en liten avvikelse i att referenssvaret nämner 2030 som EU-mål medan kandidatens svar fokuserar på

2045. Argumentationen är logiskt sammanhängande och stödjer slutsatserna tydligt. Struktur är väl organiserad med tydliga sektioner. Policyinsikten är djupt förstådda, särskilt angående BECCS:s roll och begränsningarna hos naturliga kolsänkor."

Feedback. *Judge A* (täckning 6, korrekthet 5, argumentation 4, struktur 8, policyinsikt 5): "Kandidatdokumentet täcker vissa nyckelpoänger från referensen, såsom fokus på BECCS och dess betydelse för att uppnå negativa utsläpp. Det stämmer överens med referensen när det gäller den nationella strategin och de tekniska utmaningarna kring skala och infrastruktur. Dock saknas en direkt koppling till Klimatpolitiska rådets årsrapport 2026, som betonar brister i den nuvarande politiken och behovet av förbättrad styrning. Argumentationen är logisk men inte helt sammanhängande med referensens rekommendationer om koldioxidskatt och inblandningskrav. Strukturen i kandidatdokumentet är välorganiserad, vilket gör det lätt att följa. Policyinsikten är begränsad eftersom den inte fullt ut adresserar de bredare politiska rekommendationerna från referensen." *Judge B* (täckning 9, korrekthet 8, argumentation 9, struktur 10, policyinsikt 9): "Kandidatens svar täcker de viktigaste nyckelpoängerna i referensen, inklusive BECCS som central pelare, behovet av negativa utsläpp och alternativa kolsänkor. Faktapåståenden är nästan exakta, men vissa detaljer om EU:s 55-procentspaket saknas. Argumentationen är logiskt sammanhängande och stödjer slutsatserna tydligt. Struktur är väl organiserad med tydliga sektioner. Förståelsen för policykontext och konsekvenser är djupt, särskilt angående BECCS-kostnader, lagringslängd och behovet av internationellt samarbete."

Answer under No-RAG (the judge's highest-scored condition, composite 8.10):

Kontext
Sverige har ett ambitiöst mål om netto-nollutsläpp till 2045, där negativa utsläpp spelar en avgörande roll för att kompensera svårreducerade sektorer. Fokus har länge legat på BECCS (Bio-energy Carbon Capture and Storage), vilket innebär att koldioxid fångas in från biobränsleeldade kraftvärmeverk. Tekniken ses som en nyckel för att skapa negativa flöden i atmosfären tack vare Sveriges stora biomassatillgångar. De ändrade förutsättningarna inkluderar både tekniska framsteg, ökade investeringskostnader och ett förändrat regulatoriskt landskap inom EU. Detta har tvingat fram en omvärdering av hur snabbt storskalig implementering kan ske och i vilken utsträckning statliga stöd krävs för att locka privata investeringar till en marknad som ännu saknar fullt utbyggd infrastruktur för transport och lagring.

Analys
Den nationella strategin har skiftat från en optimistisk teknikdriven ansats till en mer diversifierad riskhantering. BECCS är fortfarande hörnstenen, men osäkerheten kring kostnadseffektivitet och lagringskapacitet har lett till att strategin nu betonar en bredare portfölj av lösningar. Analysen visar att ett ensidigt beroende av BECCS innebär en hög systemrisk om tekniken inte når kommersiell skala i tid. Som ett resultat diskuteras nu alternativa kolsänkor med större intensitet. Detta inkluderar främst naturbaserade lösningar såsom skogsförvaltning för ökad kolinlagring i mark och biomassa, samt tekniker som Biochar (biokol) där kol binds stabilt i jordbruksmark. Skillnaden ligger i att

naturbaserade sänkor ofta kan implementeras snabbare, medan BECCS erbjuder en mer mätbar och permanent reduktion per enhet. Slutsatsen är att Sverige rör sig mot en hybridmodell där tekniska och biologiska sänkor kompletterar varandra för att säkra måluppfyllelsen.

Implikationer
För svensk klimatpolitik innebär detta att man inte längre kan betrakta negativa utsläpp som en framtida bonus, utan som en integrerad del av den nuvarande utsläppsbudgeten. Det krävs en tydligare definition av vad som räknas som en permanent sänka för att undvika dubbelräkning och säkerställa miljömässig integritet. Politiskt innebär det en förskjutning mot att skapa ekonomiska incitament som inte bara premierar utsläppsminskningar, utan även aktiv kolinlagring. Det skapar även en spänning mellan skogsbrukets ekonomiska intressen och behovet av att bevara kolsänkor i naturen, vilket kräver en noggrann balansgång mellan produktion av bioenergi för BECCS och bevarande av ekosystem.

Rekommendationer
För det första bör staten etablera ett robust och långsiktigt ekonomiskt stödsystem, likt omvända auktioner, för att sänka risken vid investeringar i BECCS. För det andra bör ett nationellt ramverk för certifiering av kolinlagring i mark och biokol införas för att skapa en transparent marknad för negativa utsläpp. För det tredje rekommenderas en strategisk satsning på infrastruktur för koldioxidtransport i samarbete med nordiska grannländer för att sänka enhetskostnaderna. Slutligen bör forskningsanslag prioriteras mot optimering av biokolsproduktion i kombination med jordbruksnytta, då detta erbjuder en skalbar och kostnadseffektiv väg till negativa utsläpp som kan implementeras parallellt med de större industriella projekten.

Answer under Single-RAG (the human rater's best pick, yet the judge's lowest-scored condition, composite 5.80). Citation labels are those presented to the model; the cited sources are listed in Appendix D.1:

Kontext
Sverige har fastställt ett ambitiöst klimatmål om att nå nettonollutsläpp senast år 2045, för att därefter uppnå negativa utsläpp [Källa 1, Källa 3]. För att nå detta krävs att territoriella utsläpp minskar med minst 85 procent jämfört med 1990, där kvarvarande utsläpp kompenseras genom så kallade kompletterande åtgärder [Källa 4, Källa 15]. Dessa åtgärder är särskilt kritiska för sektorer där utsläppsminskningar är tekniskt svåra, såsom jordbruket och vissa delar av industrin [Källa 1, Källa 3]. Frågan om hur dessa mål ska nås är central då det krävs genomgripande system- och teknikskiften under en kort tidsperiod [Källa 19].

Analys
Strategin för negativa utsläpp vilar på en kombination av tekniska lösningar och naturliga sänkor. Bio-CCS (BECCS) har identifierats som en nyckelkomponent för att nå både nettonoll 2045 och långsiktigt negativa utsläpp, då tekniken möjliggör permanenta negativa utsläpp [Källa 1]. Regeringen har operationaliserat detta genom Industriklivet samt införandet av driftstöd via omvänd auktionering från 2022 [Källa 1]. Som alternativ och komplement diskuteras förstärkta kolsänkor i landskapet, inklusive ökat

upptag i skog och mark samt kolinbindning på åkermark och återföring av dikad organogen mark till våtmark [Källa 2, Källa 10]. Medan kolsänkor i skog och mark kan täcka upp till hälften av de kompletterande åtgärderna fram till 2045, bedöms deras effekt avklinga över tid [Källa 2, Källa 6]. Därför ses bio-CCS som den mer skalbara lösningen för framtiden [Källa 2]. Andra diskuterade alternativ inkluderar verifierade utsläppsminskningar i andra länder [Källa 3, Källa 15] samt DACCS (direktinfångning från atmosfären), även om det senare ännu saknar aktuella investeringar i Sverige [Källa 1].

Implikationer
För svensk klimatpolitik innebär detta att man inte kan förlita sig enbart på naturliga kolsänkor, då dessa är temporära [Källa 2, Källa 6]. Det finns ett kritiskt beroende av att skala upp bio-CCS för att hantera utsläpp från jordbruket och basindustrin [Källa 1, Källa 3]. Vidare innebär strategin att kompletterande åtgärder får utgöra maximalt 15 procent av måluppfyllelsen år 2045, vilket tvingar transport- och energisektorer att i princip nå nollutsläpp utan kompensation [Källa 1]. Detta ställer höga krav på både teknikutveckling och politisk styrning för att skapa tillräckliga incitament för industrins omställning [Källa 18].

Rekommendationer
För att säkra måluppfyllelsen bör följande policyåtgärder prioriteras: 1) Fortsatt och utökat stöd till bio-CCS genom både investeringsstöd (Industriklivet) och driftstöd för att accelerera kommersiell implementering [Källa 1]. 2) Utveckling av en långsiktig strategi för att bygga upp volymen kompletterande åtgärder, med mål om minst 10,7 miljoner ton koldioxid per år till 2045 [Källa 5]. 3) Implementering av kraftfulla och långsiktiga styrmedel, såsom skatter och avgifter på utsläpp, för att driva den tekniska omställningen i industrin [Källa 14]. 4) Identifiering och uppskalning av långlivade kolsänkor för att komplettera de mer kortlivade biologiska sänkorna [Källa 7].

D.1 Citation Traceability

The Single-RAG answer above was generated with 20 labelled chunks in its context and cites 12 of them. Because the analyst model may cite only by the integer labels it is shown, every marker can be verified against the saved record; here every citation resolves to a provided chunk, giving a traceability ratio of 1.00 for this cell (consistent with Table 3.5). The cited chunks are listed below under the labels presented to the model, which are the labels used in the answer. They come from nine documents dated between 2013 and 2022, and the full text of each chunk is reproduced exactly as presented to the model, with its leading document breadcrumb removed for readability.

`[Källa 1]` **Miljöprövningsutredningen** *(SOU, 2020)*

> Om prövning och omprövning > Siktet är netto noll och negativt därefter Sverige ska nå nettonollutsläpp senast 2045 för att därefter nå negativa utsläpp. Prognoser visar att det utrymme om maximalt 15 procent som tillåts kompenseras med så kallade kompletterande åtgärder behövs framför allt för de utsläpp som är svårast att minska, t.ex. utsläpp från jordbrukssektorn och vissa

diffusa utsläpp inom industrin. Det innebär att utsläppen i övriga sektorer, bl.a. transportsektorn, el- och värmesektorn och även industrin, i princip behöver vara noll senast 2045. Negativa utsläpp kan åstadkommas genom ökat upptag av koldioxid i skog och mark samt genom bio-CCS eller direktinfångning och lagring av koldioxid i atmosfären (DACCS). I detta sammanhang ingår däremot inte upptag i skog och mark utan det som ingår är permanenta negativa utsläpp i form av bio-CCS och DACCS. Investeringar i bio-CCS är på gång i Sverige. Genom styrmedlet Industriklivet ger regeringen stöd till forskning, utveckling, test, demonstration och investeringar av bio-CCS. Från och med 2022 har regeringen även infört ett driftstöd till bio-CCS i form av så kallad omvänd auktionering. Vad gäller DACCS finns inga aktuella investeringar på gång i Sverige och en storskalig tillämpning av tekniken ligger längre fram i framtiden.

`[Källa 2]` **Med de nya svenska klimatmålen i sikte** *(DiVA, 2017)*

Med de nya svenska klimatmålen i sikte > 6.7.3 Åtgärder internationellt kan också bidra 109 FN:s klimatpanels riktlinjer för nationell växthusgasinventering från 2006 111 http://unfccc.int/focus/long-term_strategies/items/9971.php [2017-10-19] - Sammanfattningsvis skulle ungefär upp till hälften av de kompletterande åtgärderna 2045 kunna täckas av förstärkta kolsänkor i landskapet medan resten skulle kunna täckas av bio-CCS. - Eftersom effekten av förstärkta kolsänkor avklingar över tid är det andra alternativet bio-CCS som kan utvecklas och ta allt större omfattning, särskilt med tanke på behovet av netto-negativa utsläpp efter 2045. - Oavsett dessa negativa utsläpp skulle internationellt samarbete kunna bidra till att snabbare att minska utsläppen av växthusgaser i världens länder, vid sidan av utsläppsminskningen i Sverige.

`[Källa 3]` **Styrmedel för att öka kolsänkor i skogssektorn** *(DiVA, 2022)*

Styrmedel för att öka kolsänkor i skogssektorn > 1. Sammanfattning Sverige har satt upp målet att ha noll nettoutsläpp redan år 2045, och att därefter uppnå negativa utsläpp. För att nå nettonollutsläpp får kompletterande åtgärder tillgodoräknas inom givna begränsningar. Sådana kompletterande åtgärder kan vara ökat nettoupptag i skog och mark, verifierade utsläppsminskningar i andra länder samt avskiljning och lagring av biogen koldioxid (bio-CCS). Kompletterande åtgärder behövs för att klara klimatmålet om nettonollutsläpp, detta eftersom alla bruttoutsläpp av växthusgaser inte kommer att kunna reduceras till noll till år 2045. Framförallt är det tydligt i jordbrukssektorn där framförallt djurhållning bidrar med utsläpp. Men även inom viss basindustri kan det bli svårt att få ner utsläppen till noll.

`[Källa 4]` **Fördjupad analys av den svenska klimatomställningen 2021** *(DiVA, 2021)*

Fördjupad analys av den svenska klimatomställningen 2021 > 1.1. Sverige ska nå netto-noll utsläpp senast 2045 Målet om noll nettoutsläpp innebär att de territoriella utsläppen ska vara minst 85 procent lägre senast år 2045 än vad utsläppen var år 1990. De kvarvarande utsläppen ned till noll kan kompenseras genom så kallade kompletterande åtgärder. Som kompletterande åtgärder räknas upptag av koldioxid i skog och mark till följd av ytterligare åtgärder, utsläppsminskningar genomförda utanför Sveriges gränser, samt avskiljning och lagring av koldioxid från förbränning av biobränslen, så kallad bio-CCS. På motsvarande sätt som för det långsiktiga målet finns även möjlighet att nå delar av målen för den icke-handlande sektorn till år 2030 och 2040 genom kompletterande åtgärder med högst 8 respektive 2 procentenheter av utsläppsminskningsmålen år 2030

och 2040. Dessutom finns ett etappmål som säger att utsläppen från inrikes transporter, utom koldioxidutsläpp från inrikes flyg, ska minska med minst 70 procent senast år 2030

`[Källa 5]` **Fördjupad analys av den svenska klimatomställningen 2021** *(DiVA, 2021)*

Fördjupad analys av den svenska klimatomställningen 2021 > 2.1. Sammanfattning Arbetet med att successivt bygga upp en allt större mängd kompletterande åtgärder, främst i form av åtgärder som kan bidra till negativa utsläpp i LULUCF-sektorn och i form av bio-CCS har inletts i Sverige. Den klimatpolitiska vägvalsutredningen föreslog i betänkandet Vägen till en klimatpositiv framtid 13 att Sverige bör anta en strategi för att successivt bygga upp volymen kompletterande åtgärder. Strategin föreslogs ges en inriktning mot att Sverige ska åstadkomma kompletterande åtgärder som minst motsvarar 3,7 miljoner ton 2030 för att därefter successivt öka och senast 2045 uppgå till minst 10,7 miljoner ton koldioxid per år. Figur 9: Uppdaterade målscenarier 2021 och en föreslagen bana för kompletterande åtgärder jämförs med Miljömålsberedningens tidigare målscenario från 2016. Källa: Naturvårdsverket, 2021b med ytterligare bearbetning av Naturvårdsverket. Senast år 2045 ska växthusgasutsläppen i Sverige nå netto noll, och dessförinnan bör etappmål till år 2030 och 2040 klaras. För att ta oss till målen är det viktigt att åskådliggöra möjliga vägar till måluppfyllelse, det vill säga visa vilka åtgärder som utifrån dagens kunskap kan vara möjliga att genomföra, när i tid och hur snabbt de kan genomföras, samt vilka effekter åtgärderna kan bidra med.

`[Källa 6]` **Med de nya svenska klimatmålen i sikte** *(DiVA, 2017)*

Med de nya svenska klimatmålen i sikte > Industri- och energianläggningar, verksamheter som ingår i EU:s handel med utsläppsrätter, spelar en avgörande roll för att nå nettonollutsläpp år 2045. Målstyrd infrastrukturplanering innebär att säkerställa att dagens planering av transportsystemets infrastruktur, t ex genom nationella och regionala planerna för transportinfrastruktur, sker med klimatmålen som utgångspunkt. Inte minst är det av största vikt att investeringar i transportinfrastruktur som motverkar klimatmålen inte genomförs. markanvändning i form av kolsänkor. Kompletterande åtgärder får användas för att nå etappmålen. 2030 tillåts de stå för 8 procentenheter av målluppfyllelsen och 2040 för 2 procentenheter. För att nå nettonollutsläpp 2045 får kompletterande åtgärder tillgodoräknas med upp till 15 procentenheter. Vi bedömer att ungefär hälften av de kompletterande åtgärderna kan komma från kolsänkor. Dessa upptag är av en mer temporär karaktär eftersom skogen vid en viss ålder blir för gammal för att ta upp mer koldioxid. För att få till nettonegativa utsläpp i framtiden kan så kallad bio-CCS (från våra biobränsleeldade energianläggningar) också komma att behövas.

`[Källa 7]` **Potentiella kolsänkor i Malmö stad** *(DiVA, 2022)*

Den första januari 2018 trädde Sveriges nya klimatlag i kraft. Den specificerar bland annat att regeringen regelbundet ska ta fram en handlingsplan för hur klimatmålen ska uppnås. Målet är att senast år 2045 ska Sverige inte ha några nettoutsläpp av växthusgaser till atmosfären, och därefter uppnå negativa utsläpp. Negativa utsläpp innebär att utsläppen av växthusgaser från verksamheter i Sverige är mindre än till exempel den mängd koldioxid som tas upp av naturen som en del av kretsloppet, eller mindre än de utsläpp Sverige bidrar till att minska utomlands genom att investera i olika klimatprojekt. För att uppnå detta mål ska utsläppen vara 85 % lägre än referensåret 1990

1 . Resterande utsläpp behöver tas bort från atmosfären med kompletterande åtgärder. För att nå målet behöver kolsänkor 2 identifieras och skalas upp, och det finns många olika lagringsmetoder. Vissa metoder ger en kortlivad kolsänka och andra metoder ger upphov till betydligt längre tidshorisont för kolsänkan. Långlivade kolsänkor har en inlagring på flera hundra eller tusen år som till exempel BECCS. BECCS är koldioxidavskiljning från rökgaser efter förbränning av biobränslen som komprimeras till flytande form och lagras i underjordiska bergsformationer. Ursprungligen från engelska Bio Energy with Carbon Capture and Storage.

`[Källa 10]` **Fördjupad analys av den svenska klimatomställningen 2021** *(DiVA, 2021)*

Fördjupad analys av den svenska klimatomställningen 2021 > Målscenariot som nådde närmast nettonollutsläpp 2050 uppdaterades av Miljömålsberedningen och målåret för nettonollutsläpp sattes till 2045 Dessutom antogs även vissa åtgärder för ökad kolsänka genomföras genom ökad kolinbindning på åkermark (på mineraljord), fler buskar och träd på betesmark samt koldioxideffekter genom att dikad organogen mark återfördes till våtmark. De sistnämnda åtgärdstyperna sänkte dock inte utsläppen från jordbrukssektorn i målscenarierna eftersom de räknas till sektorn markanvändning, förändrad markanvändning och skogsbruk, LULUCF. Utsläppen 2030 härrörde främst från industriprocesser, jordbrukssektorn, viss avfalls-, kol, torv- och naturgasförbränning inom el- och fjärrvärmetillförsel, samt kvarvarande användning av fossila drivmedel i transportsektorn och i arbetsmaskiner. Kvarvarande utsläpp 2050 härrörde främst från jordbrukssektorn, processutsläpp från industrin (CCS- teknik på befintliga anläggningar bedömdes inte kunna ta bort utsläppen helt), metan- och lustgasutsläpp från förbränningsprocesser samt från övrig hantering av organiska substanser i olika delar av samhället (avloppsreningsverk, rötning, lösningsmedel) men även från förbränning av (plast)avfall. Eftersom införandet av CCS-teknik även delvis antogs omfatta förbränning av biobränslen, bio-CCS, kompenserades kvarvarande utsläpp delvis i scenariot.

`[Källa 14]` **2050 Ett koldioxidneutralt Sverige** *(DiVA, 2013)*

2050 Ett koldioxidneutralt Sverige > utsläppsminskninGar i andra länder FÖr att nå nollresultat Miljoner ton CO2-ekv. Miljoner ton CO2-ekv. Miljoner ton CO2-ekv. Totala och sektorsvisa utsläppsminskande banor mot 2050 med högre elbehov och utan koldioxidinfångning jämfört med målscenario 1. I industrin ingår åtgärder som minskar förbränningsutsläpp och alternativ teknik i järn och stålindustrin. Nettoupptag för skogsbruk till år 2100 för fyra alternativa nationella utvecklingsscenarier. Figuren redovisar endast skogsbruk (inte hela Skogs- och markanvändningssektorn) och att inlagring av kol i olika produktkategorier inte är inkluderad. Klimatpolitiken behöver kraftfulla och långsiktiga styrmedel för att nå målen. Skatter och avgifter på utsläpp hör till de mest effektiva styrmedlen. Handel med utsläppsrätter kan bli effektivt, om utsläppsrätterna blir en bristvara och priserna drivs upp. Att sätta pris på utsläppen är 'motorn', det som främst driver samhället mot att nå klimatmålen.

`[Källa 15]` **Fördjupad analys av svensk klimatstatistik 2017** *(DiVA, 2017)*

Fördjupad analys av svensk klimatstatistik 2017 > 2. Sveriges klimatmål är utmanande Att inte ha några nettoutsläpp av växthusgaser betyder i det här fallet att utsläppen av växthusgaser från verksamheter i Sverige ska vara minst 85 procent lägre år 2045 än utsläppen år 1990. De kvar-

varande utsläppen ned till noll kan kompenseras genom så kallade kompletterande åtgärder, vilka ger flexibilitet och kostnadseffektivitet för att nå målet. Dessa åtgärder kan även bidra till negativa nettoutsläpp efter 2045. Som kompletterande åtgärder räknas: - upptag av koldioxid i skog och mark till följd av ytterligare åtgärder (som är additionella, alltså utöver de åtgärder som redan genomförs), - utsläppsminskningar genomförda utanför Sveriges gränser, samt - negativa utsläpp och framförallt avskiljning och lagring av koldioxidutsläpp från biobränsleanvändning, så kallad bio-CCS. Etappmålen på väg mot det långsiktiga målet innebär att utsläppen i Sverige i de sektorer som inte omfattas av EU:s system för handel med utsläppsrätter (EU ETS)

[Källa 18] **Fördjupad analys av den svenska klimatomställningen 2019 Industrin i fokus** *(DiVA, 2019)*

Fördjupad analys av den svenska klimatomställningen 2019 > Industrin i fokus För att nå Sveriges långsiktiga mål - att nå nettonoll i utsläpp av växthusgaser senast 2045 - behövs en omfattande omställning inom industrin. Industrin står redo och det offentliga erbjuder betydande medel för ny teknik, men den politiska styrningen ger inte tillräckliga incitament för att genomföra omställningen. Rapporten analyserar framtidsscenarier för industrins klimatomställning, möjliga bidrag från resurseffektiviserande åtgärder, biomassans roll i omställningen samt hur incitamenten begränsas av gratis tilldelning av utsläppsrätter och vilka finansiella stöd som finns tillgängliga för omställningen. Naturvårdsverket har, på eget initiativ, tagit fram denna rapport för att följa upp utvecklingen mot de svenska klimatmålen och klimataspekter av Generationsmålet. Statistiken som publiceras i rapporten ligger till grund för uppföljningen av de nationella klimatmålen i klimatredovisningen som bilaga till budgetpropositionen, uppföljning av Sveriges miljömål samt internationell rapportering till EU och FN.

[Källa 19] **Begränsad klimatpåverkan** *(DiVA, 2019)*

Begränsad klimatpåverkan > I SVERIGE De centrala utmaningarna i Sverige på kort sikt för att nå etappmålen om att minska utsläppen inom Sveriges gränser handlar om att ställa om transportsystemet till ett transporteffektivt, energieffektivt och förnybart sådant. På längre sikt handlar det om att industrins process- och förbränningsutsläpp ska minska till nära noll genom att ny teknik utvecklas och införs. För att nå nettonollutsläpp krävs även kompletterande åtgärder som ökade kolförråd i skog och mark, koldioxidinfångning och -lagring i berggrunden (bio-CCS) och utsläppsminskande åtgärder i andra länder. Denna stora samhällsomställning med genomgripande system- och teknikskiften ska ske under en historiskt sett kort tidsperiod. Detta är en mycket stor utmaning då systemoch teknikskiftena kräver stora investeringar och har relativt lång tid för implementering. 31 Naturvårdsverket, (2018). Årlig uppföljning av Sveriges nationella miljömål 2018. (Samt motsvarande publikationer 2017 och 2016.)

Chapter E

Full Statistical Results

This appendix reports the complete per-model significance tests summarised in Section 3.2.3. The omnibus Friedman test (Table E.1) was computed for all seven analyst models; Bonferroni-corrected Wilcoxon post-hoc comparisons (Table E.2) were computed only for the four models whose omnibus test was significant. In every significant post-hoc comparison the direction favours no-RAG.

Table E.1: Friedman omnibus test per analyst model (composite score, four conditions as repeated measures over ten questions). Post-hoc pairwise comparisons (Table E.2) were computed only where the omnibus test was significant.

Model	χ^2	p	Kendall's W	Post-hoc?
gemma-4-e4b-it	19.07	< 0.001	0.636	yes
granite-4.1-8b	6.04	0.110	0.201	no
gemma-4-31b-it@4bit	15.87	0.001	0.529	yes
granite-4.1-30b	10.04	0.018	0.335	yes
qwen3.6-27b-ud-mlx	6.73	0.081	0.224	no
qwen3.6-35b-a3b-ud-mlx	9.19	0.027	0.306	yes
gemma-4-26b-a4b@4bit	4.33	0.228	0.160	no

The reliability diagnostics are as follows. Inter-judge agreement: Spearman $\rho = 0.37$ ($p < 10^{-10}$, $n = 279$). Length-bias check: Spearman $\rho = 0.04$ ($p = 0.52$, $n = 279$). Cross-model concordance on the condition ranking: Kendall's $W = 0.118$ across 7 models. Inter-dimension association between *argumentation* and *struktur*: Spearman $\rho = 0.47$, pooled over both judges.

On the composite across the 279 scored cells, Judge A (`microsoft/phi-4`) has mean 7.11 and standard deviation 1.59; Judge B (`nvidia/nemotron-3-nano-4b`) has mean 8.19 and standard deviation 0.97. Per-condition means by judge are given in Table E.3. Both judges independently rank no-RAG first, and the ranking is unchanged when each judge's scores are standardised to zero mean and unit variance before averaging (Section 3.7).

Table E.2: Complete Bonferroni-corrected Wilcoxon signed-rank pairwise comparisons for the four models with a significant Friedman test (composite score). Δ is the mean paired difference (first minus second condition); r_{rb} is the matched-pairs rank-biserial effect size; p_{adj} is Bonferroni-corrected across the six pairs. Bold marks $p_{\text{adj}} < 0.05$.

Model	Pair	Δ	r_{rb}	p_{adj}
gemma-4-e4b-it	no-RAG vs single	$+0.91$	$+1.00$	**0.012**
	no-RAG vs agentic	$+0.79$	$+1.00$	**0.012**
	no-RAG vs feedback	$+0.85$	$+1.00$	**0.012**
	single vs agentic	-0.12	$+0.16$	1.000
	single vs feedback	-0.06	$+0.07$	1.000
	agentic vs feedback	$+0.06$	$+0.29$	1.000
gemma-4-31b-it@4bit	no-RAG vs single	$+0.88$	$+0.91$	**0.047**
	no-RAG vs agentic	$+0.54$	$+0.82$	0.117
	no-RAG vs feedback	$+0.63$	$+1.00$	**0.012**
	single vs agentic	-0.34	-0.56	1.000
	single vs feedback	-0.25	-0.11	1.000
	agentic vs feedback	$+0.09$	$+0.33$	1.000
granite-4.1-30b	no-RAG vs single	$+0.13$	$+0.14$	1.000
	no-RAG vs agentic	$+0.31$	$+0.94$	0.094
	no-RAG vs feedback	$+0.42$	$+0.95$	**0.035**
	single vs agentic	$+0.18$	$+0.53$	0.961
	single vs feedback	$+0.29$	$+0.60$	0.621
	agentic vs feedback	$+0.11$	$+0.33$	1.000
qwen3.6-35b-a3b-ud-mlx	no-RAG vs single	$+1.11$	$+0.89$	0.059
	no-RAG vs agentic	$+0.54$	$+0.95$	**0.035**
	no-RAG vs feedback	$+0.29$	$+0.76$	0.281
	single vs agentic	-0.57	-0.22	1.000
	single vs feedback	-0.82	-0.58	0.703
	agentic vs feedback	-0.25	-0.50	1.000

Table E.3: Mean composite score per condition, by judge, across the 279 scored cells. The final column gives the condition mean after each judge's scores are standardised to zero mean and unit variance.

Condition	Judge A	Judge B	Composite	Standardised
No-RAG	7.15	8.37	7.76	$+0.11$
Feedback	7.14	8.21	7.67	$+0.02$
Single-RAG	7.12	8.06	7.59	-0.06
Agentic	7.01	8.13	7.57	-0.06

TRITA-ITM-EX 2026:591