Explanatory & Coherence guided LLM Reasoning

Student ID: 10458052

An extended research project report
submitted to the University of Manchester
for the degree of MSc Data Science in the Faculty of Humanities

School of Social Sciences

2024

Table of Contents

Total word count: 7,094.

List of Illustrations

Abstract

Large Language Models have advanced in generating text-based responses across various domains but continue to struggle with producing coherent explanations, particularly in complex reasoning tasks. This report addresses this by integrating two epistemic concepts—Explanatory Power and Coherence—into LLM architectures, particularly within a Self-Discover framework. These modules aim to enhance the factual accuracy and logical consistency of LLM-generated explanations.

The research focuses on how Explanatory Power and Coherence affect explanation quality, evidence utilisation, and factual alignment in LLMs across evidence scenarios, including full, missing, wrong, and mixed evidence. It also examines how these modules impact tasks requiring the selection and assignment of evidence to competing claims.

The experimental evaluation uses two data sources: the CIViC database, offering structured, expert-curated evidence on genetic variants in cancer, and the Right for Right Reasons (R4C) Reading Comprehension dataset, which includes noisy, unstructured data. Tests involve generating explanations under various evidence scenarios and evaluating models using metrics such as BERTScore, Fluency, Semantic Similarity, Natural Language Inference Score, and Claim Support.

Results show that incorporating Explanatory Power and Coherence modules improves metrics like Coherence, Claim Support, and Fact Verification. However, these gains often come at the cost of other metrics. The larger GPT-4o model consistently outperforms GPT-4o Mini across most metrics, especially in complex reasoning tasks. While the epistemic modules improve reasoning depth and factual alignment, trade-offs in naturalness and explanation completeness are noted, particularly with competing claims and evidence.

These findings contribute to improving LLM reasoning, highlighting the need for a balanced approach that enhances explanation accuracy and coherence while maintaining fluency. Future research could refine these modules to optimise performance in real-world decision-making where explanation quality is critical.

Acknowledgements

First and foremost, I would like to express my gratitude towards Dr André Freitas for providing the opportunity of undertaking this project, as well as for continuous support, professional advice, and guidance throughout the process.

I would also like to extend my appreciation to Kristaps Stolarovs, Marek Michalowski, Diana Akolzina, and Ethan Fowler, as their advice and suggestions helped to shape the scope and body of this project in all of its stages.

The input of these people, as well as of other friends and family, has been invaluable – thank you.

Declaration

No portion of the work referred to in this extended research project report has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Intellectual Property Statement

i. The author of this extended research project report (including any appendices and/or schedules to this report) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this report, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks, and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the report, for example graphs and tables (“Reproductions”), which may be described in this report, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this report, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see https://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant dissertation restriction declarations deposited in the University Library, The University Library’s regulations (see https://www.library.manchester.ac.uk/about/regulations/) and in The University’s Guidance for the Presentation of dissertations.

Chapter 1

Introduction

Large Language Models (LLMs) have transformed artificial intelligence, excelling in tasks from text generation to reasoning. However, limitations persist in their ability to generate meaningful explanations for complex inference. This report addresses two critical epistemic concepts—Explanatory Power (EXP) and Coherence (COH)—to improve LLMs' reasoning capabilities, particularly in generating plausible, well-supported explanations. Integrating these concepts aims to enhance LLM reasoning quality, increasing effectiveness and trustworthiness in decision-making processes requiring high-level reasoning and accurate evidence interpretation.

1.1 Background

LLM use has grown exponentially, with models like OpenAI's GPT or Google's Gemini pushing boundaries. As LLMs integrate into healthcare, law, and policy-making, the need for robust, reliable, and logically coherent explanations intensifies. Despite proficiency in surface-level tasks, many LLMs struggle with generating deep, logically consistent answers for complex reasoning tasks. Traditional approaches often overlook reasoning depth, resulting in explanations that may be factually correct but lack coherence or fail to meaningfully connect evidence and conclusions.

Research in improving LLM reasoning has explored methods like CoT, ReAct, or ReWOO frameworks. These aim to address challenges by introducing structured reasoning processes or combining reasoning with external actions. While promising, significant challenges remain in multi-step reasoning, coherence, and handling contradictory or incomplete evidence.

Explanatory Power is the ability to effectively account for evidence with clarity and simplicity. This is critical in domains like medical diagnostics or scientific research, where explanations must be accurate and meaningful. Coherence ensures all elements of an explanation align logically, preventing contradictions and disjointed reasoning. These concepts offer a framework for improving LLMs' reasoning capabilities, enabling generation of explanations that are accurate, contextually appropriate, and logically sound.

1.2 Scope of the Study

This report addresses limitations in LLM-generated explanations by integrating EXP and COH modules into state-of-the-art natural language models. It explores how these epistemic concepts influence reasoning, particularly in scenarios with varying evidence conditions (full, missing, or wrong). By incorporating EXP and COH, the aim is to improve factual accuracy, coherence, relevance, and overall quality of LLM-generated explanations. This study serves as a proof of concept and basis for future research rather than a final framework.

The research aims to answer the following key questions:

(1) How does the inclusion of Explanatory Power and Coherence modules affect the quality of explanations generated by Large Language Models across different evidence scenarios?

(2) In what ways do Explanatory Power and Coherence notions affect the evidence selection, gathering, and utilisation processes in large language models during explanation generation?

(3) What are the relative contributions of Explanatory Power and Coherence to improving the factual accuracy and alignment with evidence in explanations generated by large language models?

(4) How does the presence of Explanatory Power and Coherence modules affect the ability of large language models to differentiate and select appropriate evidence for multiple claims in combined evidence scenarios?

1.3 Structure of the Report

The report is structured as follows:

- **Chapter 2: Literature Review** - Overviews relevant literature on LLM reasoning capabilities, integration of Explanatory Power and Coherence modules, and identifies key research gaps.

- **Chapter 3: Methodology** - Details experimental setup, including data sources, test designs, core architectures, and evaluation metrics.

- **Chapter 4: Results and Analysis** - Presents and interprets experiment results, including metrics for various evidence scenarios, and compares models' performance.

- **Chapter 5: Discussion** - Discusses broader implications, highlighting trade-offs between metrics like Coherence, Claim Support, and Explanation Accuracy.

- **Chapter 6: Conclusions, Limitations, and Recommendations** - Summarises key conclusions, discusses study limitations, and suggests future research directions.

Chapter 2

Literature Review

2.1 Introduction

Intelligent reasoning has become a critical area of AI research, particularly in LLMs, where generating factual explanations is essential. As AI becomes integrated into decision-making, improving LLM reasoning is crucial for producing trustworthy outputs with minimal supervision.

This literature review focuses on two key epistemic concepts—Explanatory Power and Coherence—and their role in LLM-generated explanations. By assessing these concepts in current reasoning frameworks, this chapter lays the groundwork for evaluating the state of the art in LLM reasoning and identifying key challenges.

It also highlights deficiencies in LLM-generated explanations, emphasizing the need to focus not just on accuracy but on the reasoning process itself. These epistemic notions are positioned as valuable tools for enhancing the design and evaluation of future LLM systems.

2.2 Explanatory Power and Coherence

The epistemic notions of Explanatory Power and Coherence can significantly enhance our understanding of LLM reasoning and explanation generation. Explanatory Power refers to an explanation's ability to account for evidence with clarity, simplicity, and minimal assumptions (Schupbach and Sprenger, 2011). In LLMs, dealing with vast data, this ensures explanations are factually accurate, relevant, concise, and meaningful. For instance, an LLM providing a scientific explanation should not only state facts but also link them in a way that maximises relevance to the question.

Explanatory Power is crucial when LLMs autonomously generate plausible explanations, especially in high-stakes areas like scientific research or medical diagnostics, where there is often a trade-off between hypothesis complexity and evidence (Glass, 2023).

Coherence focuses on the logical consistency of an explanation and its alignment with known facts and context. It ensures that each element supports the others, preventing contradictions. Coherence is crucial in evaluating complex, multi-step, or competing explanations (Glass, 2007). For instance, in legal reasoning, an LLM must ensure its arguments follow a logical progression without contradicting established knowledge or previous claims (Amaya, 2007). A coherent explanation strengthens trust in the model's outputs, particularly in automated decision-making processes requiring minimal oversight.

Together, these notions provide an efficient framework for evaluating and improving LLM reasoning, as both play crucial roles in how humans process new information (Douven and Schupbach, 2015), and recent work shows that formal models of Explanatory Power and Coherence can be empirically tested to refine our understanding of human cognition (Schupbach and Sprenger, 2011). While critiques exist regarding their relationship with truth conduciveness (Hansson and E. J. Olsson, 1999; E. Olsson, 2023), these notions push models beyond surface-level correctness toward comprehensive, consistent, and contextually appropriate explanations—essential for trustworthy AI systems.

2.3 Capabilities and Challenges of LLM Reasoning

LLMs leverage vast data and machine learning algorithms to generate explanations, predictions, and human-like reasoning (Morishita et al., 2023). Their reasoning arises from analysing patterns through transformer architectures, generating contextually relevant outputs based on probabilistic word relationships (Y. Zhang et al., 2024). When generating reasoning chains, LLMs use pattern matching and statistical inference to predict explanations from a given set of evidence (Jhamtani and Clark, 2020), excelling in summarisation, question answering, and constrained response generation tasks.

Despite their strengths, LLMs struggle with tasks requiring deeper understanding or logical inference due to their reliance on statistical patterns from training data (Kunz and Kuhlmann, 2024; Ajwani et al., 2024), rather than true comprehension. This is evident when tasked with causality or multi-step reasoning (Jhamtani and Clark, 2020; Dalvi et al., 2022). For instance, LLMs may generate superficially plausible explanations that miss underlying causal relationships (Morishita et al., 2023; Dalvi et al., 2022).

They also struggle with coherence in long-form explanations, leading to contradictions or disjointed reasoning (Wang, Yue, and Sun, 2023).

Finally, LLMs are susceptible to biases in training data, leading to explanations that reinforce stereotypes or favour common patterns over more plausible or contextually appropriate explanations (Y. Zhang et al., 2024; Kunz and Kuhlmann, 2024). Addressing these challenges is crucial for enhancing the reliability and trustworthiness of LLM-generated answers in real-world decision-making tasks.

2.4 Current Approaches

Advancements in frameworks for improving LLM reasoning have progressed significantly, offering methods to enhance explanation quality and reasoning chains. These range from simple prompting to advanced systems incorporating symbolic reasoning, modularity, and verbal reinforcement learning, all aimed at improving natural-language reasoning and decision-making.

One early method, Zero-shot Chain-of-Thought (CoT) prompting, guides models to tackle tasks step by step without task-specific examples. While successful in some tasks as a lightweight, easily implementable solution, it struggles with more complex reasoning requiring deeper understanding or logical inference (Kojima et al., 2023).

The Self-Discover framework enables LLMs to autonomously compose reasoning structures by dynamically selecting modules, improving accuracy on complex benchmarks like BigBench-Hard by up to 32%. Though it enhances reasoning depth, challenges remain in generalising to more diverse or complex tasks (Zhou et al., 2024).

The ReAct framework integrates reasoning and acting, allowing models to engage in dynamic decision-making by combining internal reasoning with external actions. This reduces issues like hallucination and error propagation, though synchronising reasoning and acting remains a source of error (Yao et al., 2023).

In contrast, ReWOO introduces a modular approach by decoupling reasoning from external observations, reducing token consumption and improving computational efficiency. This allows smaller models to perform comparably to larger ones. While scalable, optimising its modular components in real-world settings remains complex (Xu et al., 2023).

The Reflexion framework, focused on verbal reinforcement learning, improves reasoning through self-reflection and feedback integration, mimicking human learning. This enhances adaptability and decision-making without extensive fine-tuning. However, its effectiveness depends heavily on the quality and structure of feedback loops, limiting its use in less controlled environments (Shinn et al., 2023).

In conclusion, while each approach improves LLM reasoning, none provide a complete solution for enhancing explanation generation. Ongoing development must address gaps in multi-step reasoning, robustness, and justification quality.

2.5 The Research Gap

Current approaches to improving LLM outputs often prioritise correctness over reasoning depth and quality. This focus on accuracy neglects whether explanations are meaningful, relevant, or coherent. Consequently, even accurate outputs may lack insightful explanations for complex problems, undermining trustworthiness and utility in decision-making tasks.

To address this, it is essential to improve not just end-results, but also the quality of reasoning behind them. Introducing Explanatory Power and Coherence enables a more holistic approach, emphasising the synthesis of relevant information and logical connections. Incorporating these notions can lead to more robust models capable of producing trustworthy, meaningful explanations.

This research aims to bridge the gap by integrating Explanatory Power and Coherence into LLM reasoning. Focusing on these notions, the project seeks to develop more reliable and contextually appropriate reasoning frameworks, improving performance in real-world applications.

Chapter 3

Methodology

3.1 General Overview

This study aims to explore how incorporating Explanatory Power and Coherence into a SOTA natural-language framework enhances reasoning, particularly in generating plausible explanations from evidence and claims. A series of tests were devised using two main data sources to evaluate a prompt-based framework supported by retrieval-augmented generation under varying evidence accuracy, completeness, and relevance. The following sections outline the data, test definitions, core architectures, and evaluation metrics used.

3.2 Data and Test Definitions

3.2.1 Data Sources and Preparation

Clinical Interpretation of Variants in Cancer (CIViC)

This database was selected for its structured, expert-curated data, offering a unique challenge for reasoning in complex medical scenarios. It contains concise, dense information about genetic variants, therapies, and molecular profiles related to cancer and its treatment, providing a suitable base for testing inference capabilities under difficult conditions (Good et al., 2014).

The data originated from two primary sources: assertion data, summarising evidence of clinical information for variants in specific cancer contexts, and molecular profiles, which combine multiple CIViC variants across genes.

Dataset preparation began with creating a baseline from accepted evidence and assertion data from the August CIViC database release. This data, containing molecular profiles, evidence items, and associated assertions, was formatted into a consistent JSON structure. The process involved cleaning and aligning assertion claims with their evidence and context. Relevant fields were transformed into claims, explanations, and supporting evidence. Exact file structure can be seen in Appendix E. This dataset formed the basis for various evidence scenarios tested in later sections.

The exact mapping between CIViC and JSON files is shown in Appendix E, with an example in Figure 3.1. Preparation primarily involved joining relevant fields, as the CIViC database is constantly curated by medical experts, ensuring quality and completeness. Missing entries were appropriately labelled.

Example Full Evidence Entry

Claim:
HER2 amplification predicts sensitivity to Trastuzumab.

Explanation:
HER2 amplification (...) standard of care for HER2-positive breast cancer.

Evidence:

- **EID 1122:** HERA was a Phase III trial (...) trastuzumab treatment for one year.
- **EID 528:** A randomized clinical trial (...) trastuzumab with chemotherapy.
- **...**

Context:

- **Molecular Profile:** ERBB2 Amplification
- **Molecular Profile Summary:** HER2 (ERBB2) amplifications (...) targeted in neoadjuvant breast cancer treatment.
- **Disease:** HER2-receptor Positive Breast Cancer
- **Therapies:** Trastuzumab
- **Phenotypes:** None specified

Figure 3.1: Example of extracted Textual Data from CIViC database.

An additional knowledge base of unique evidence items aggregated across all assertions was created for use in later testing scenarios. Figure 3.2 provides an example.

Evidence Items

Evidence Items:

- **EID 1193:** In a set of 31 patients with NSCLC, RT-PCR was performed to test for known EML4-ALK variants. Five patients (...) as other EML4-ALK NSCLC cell lines.
- **EID 1356:** In Ba/F3 cells, crizotinib sensitivty of the NPM-ALK variant was compared to crizotinib sensitivity (...) for NPM-ALK in Ba/F3 cells.
- **EID 1197:** Ba/F3 cells transduced with EML4-ALK variants v1 (E13;A20), v2 (E20;A20), v3a (E6;A20) and (...) some impact on patient outcome in targeted therapy.
- ...

Figure 3.2: Example of Evidence Items within generated knowledgebase.

Right Reasons Reading Comprehension (R4C)

This dataset, built on HotpotQA, presents a different challenge than CIViC, containing long, unstructured, and sometimes irrelevant information. It tests a model's ability to filter noise and distil critical facts. R4C complements CIViC's structured data by evaluating the capacity to navigate complex inputs and generate coherent, accurate explanations (Inoue, Stenetorp, and Inui, 2020).

Data preparation involved extracting and processing raw files from R4C and using LLMs to generate plausible evidence-based explanations. To reduce model-specific bias, OpenAI's **GPT-4o** and Anthropic's **Claude 3.5 Sonnet** were used interchangeably (OpenAI, 2024; Anthropic, 2024), both configured with a temperature of 0.7 to produce human-like explanations.

The process began by retrieving relevant Question-Answer pairs from HotpotQA (Yang et al., 2018). These were input to an LLM to generate claims, which served as foundational statements for subsequent evidence and explanations.

For evidence generation, annotated facts from R4C were processed through an LLM to remove duplicates, producing a set of *golden evidence* to ensure accuracy in subsequent explanations.

The generated claims and golden evidence were then used by LLMs to produce explanations supporting the claims. Claims, evidence, and explanations were saved as JSON files. Figure 3.3 shows the exact workflow, with prompts in Appendix E.

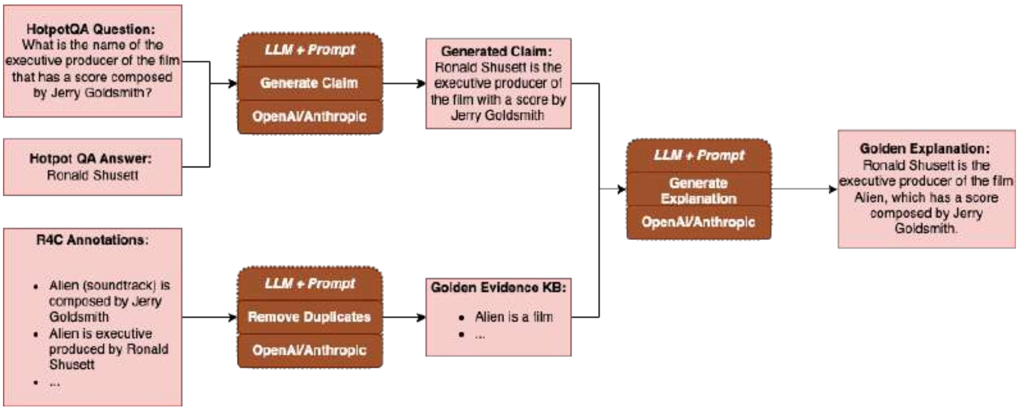
HotpotQA Question:
What is the name of the executive producer of the film that has a score composed by Jerry Goldsmith?
Hotpot QA Answer:
Ronald Shusett
R4C Annotations:
Alien (soundtrack) is composed by Jerry Goldsmith
Alien is executive produced by Ronald Shusett
...
LLM + Prompt
Generate Claim
OpenAI/Anthropic
LLM + Prompt
Remove Duplicates
OpenAI/Anthropic
Generated Claim:
Ronald Shusett is the executive producer of the film with a score by Jerry Goldsmith
Golden Evidence KB:
Alien is a film
...
LLM + Prompt
Generate Explanation
OpenAI/Anthropic
Golden Explanation:
Ronald Shusett is the executive producer of the film Alien, which has a score composed by Jerry Goldsmith.

Figure 3.3: R4C Data Generation Process.

Example of such generated and extracted data can be seen in Figure 3.4.

Example Full Evidence Entry

Claim:
Eenasul Fateh, known by his stage name Aladin, worked as a consultant to help organizations improve their performance.

Explanation:
Eenasul Fateh, known as Aladin, was a former international management consultant (...) to assist organizations in enhancing their performance.

Golden Evidence:

- **EID 11:** Management consulting is the practice of helping organizations to improve performance.
- **EID 14:** Eenasul Fateh was known as Aladin.
- ...

Evidence:

- **EID 1:** Sylvester Samuels, known by his stage name Lil Ru (...) is an American rapper.
- **EID 10:** Mick Batyske, known by his stage name Mick (...) is an American DJ and entrepreneur.
- ...

Figure 3.4: Example of processed Textual Data from R4C dataset.

Note that the *Golden Evidence* refers to R4C annotated facts, where *Evidence* refers to the original HotpotQA paragraphs.

3.2.2 Test Definitions

Three tests were designed to measure the reasoning capabilities of models under different scenarios:

1. **Explanation Test:** This test focuses on the generation of plausible explanations when given a set of evidence and a corresponding claim. Four variations were designed:

    1.1. **Full Evidence:** All relevant evidence is presented.

    1.2. **Missing Evidence:** At least one piece of evidence is kept, the rest are skipped at random.

    1.3. **Wrong Evidence:** From one to four irrelevant pieces of evidence are added.

    1.4. **Mixed Evidence:** A combination of missing and incorrect information based on rules specified above.

2. **Selection Test:** Model selects and explains the claim most supported by given evidence.

3. **Assignment Test:** Model assigns from three evidence sets to two claims and generates explanations.

CIViC data was used across all tests, while R4C supported only the *Full Evidence* scenario to minimize API costs. Examples are shown in Figures 3.1 and 3.4, with additional examples in Appendix E. Models were provided only relevant evidence, the claim, and context (if applicable). For R4C, models received *Evidence* but not *Golden Evidence*, though evaluated against the latter where applicable. Models never saw the associated explanation but were evaluated against it. Exact test prompts are in Appendix E.

3.3 Core Architectures

3.3.1 Self-Discover Agent

The Self-Discover framework is this project's primary framework (Zhou et al., 2024). It uses multi-layered reasoning to process input and generate evidence-based explanations. Self-Discover was chosen for its adaptability in reasoning tasks, ease of integrating epistemic concepts like Explanatory Power or Coherence, and SOTA results across several tasks, making it an excellent comparison baseline.

At the core of the Self-Discover framework are several reasoning modules, each with a specific role in guiding the model's reasoning process, which for this project are:

- **Baseline Modules**: Standard reasoning strategies from the original Self-Discover paper, serving as control mechanisms for performance comparisons.
- **Explanatory Modules**: Guide the model to produce detailed cause-effect explanations, maximizing explanatory power.
- **Coherence Modules**: Ensure logical consistency in reasoning across multiple steps and selected evidence.

To maintain simplicity and align with the original approach, the Explanatory and Coherence modules were adapted to fit the framework's structure, improving complex task handling with minimal added complexity. The exact module formulations are in Appendix B.

The Self-Discover architecture follows the following process regarding task solving:

1. **Task Contextualization**: Introduces Explanatory and/or Coherence concepts if selected, aligning the task towards epistemic concepts. Definitions are in Appendix B.
2. **Module Selection**: Framework selects most suitable reasoning modules based on full task description.
3. **Module Adaptation**: Selected modules are adapted to align with the specific task.
4. **Structured Reasoning**: Adapted modules are incorporated into a step-by-step reasoning structure.
5. **Task Execution**: Framework uses structured reasoning steps to solve the task and produce output.

This process is illustrated in Figure 3.5, with step prompts in Appendix B. Task *Contextualization* was introduced in this project to help LLMs handle complex epistemic notions, unlike the original Self-Discover paper. EXP and COH modules were also added. All LLMs used had their temperature set to 0 for reproducibility.

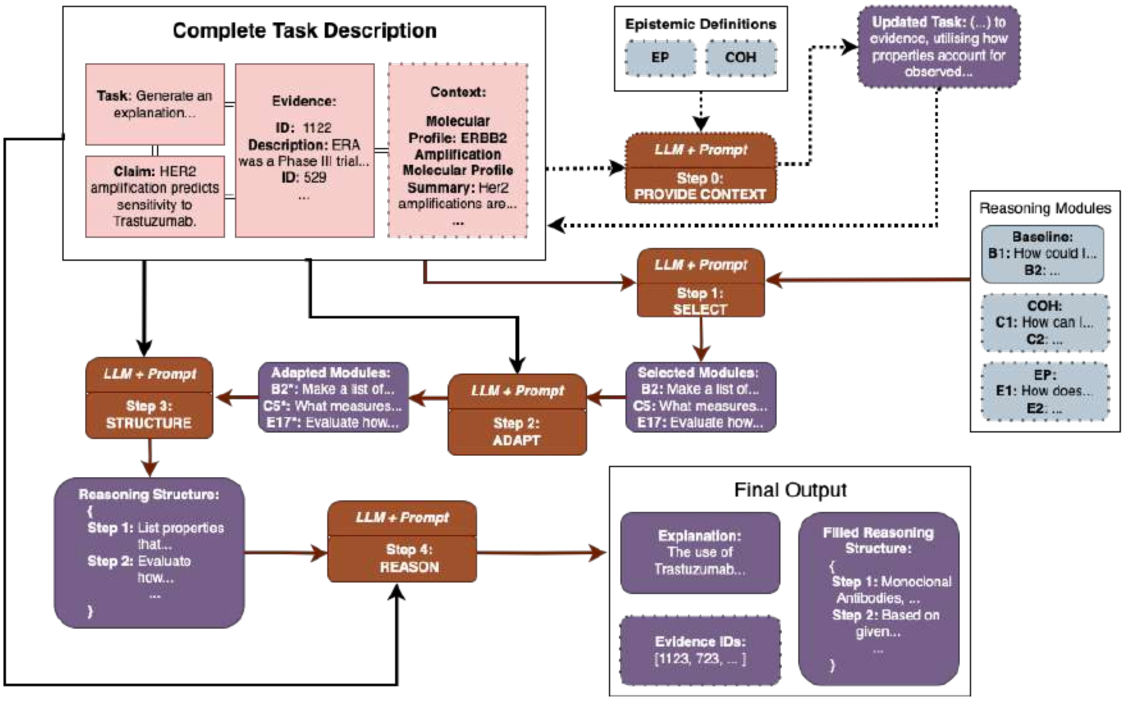

Figure 3.5: Adapted Self-Discover workflow (Zhou et al., 2024).

Note that dashed arrows and shapes indicate optional steps or modules.

3.3.2 Self-RAG Agent

For tasks with partially missing evidence, a Self-RAG agent was implemented (Asai et al., 2023). It combines LLMs with document retrieval to generate contextually relevant responses. Self-RAG integrates retrieval and generation workflows, enabling continuous improvement through self-assessment and iterative refinement. In this project, it served as a supporting agent, allowing core models to search and incorporate evidence from a general knowledgebase. This approach addresses real-world scenarios with omitted crucial evidence and evaluates models' retrieval capabilities in specific scenarios.

Self-RAG architecture follows the following process for evidence retrieval:

1. **Document Retrieval**: System retrieves top 5 relevant documents using vectorized embeddings based on semantic similarity to query.

2. **Relevance Check**: System evaluates document relevance. If none found, query is transformed and retrieval repeats.

3. **Answer Generation**: LLM generates answer based on retrieved documents, cross-referencing with provided evidence.

4. **Hallucination Detection**: System checks for portions of answer not based on retrieved documents. Revises if hallucinations detected.

5. **Usefulness Evaluation**: System assesses if answer sufficiently addresses query. If not useful, query is transformed and process restarts (limited to one transformation).

This process can be further examined in Figure 3.6, and exact step prompts in Appendix C. Similarly to Self-Discover, the LLM used in this framework was set with a temperature of 0.

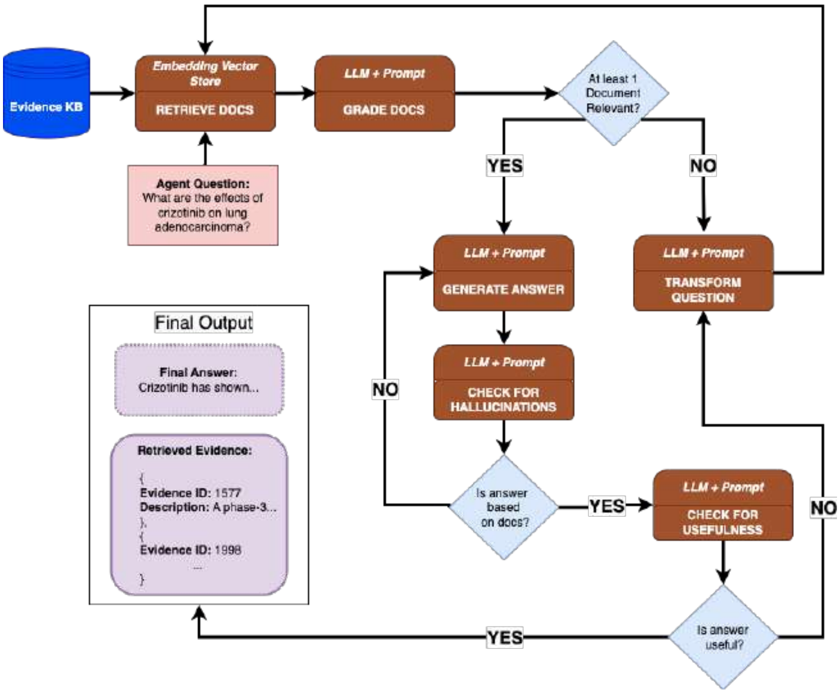

Figure 3.6: Self-RAG workflow (Asai et al., 2023).

3.4 Evaluation Metrics

Due to the fact that both reference and generated data is in the form of text, all defined and used metrics for this project are LM-based, with the models used shown in Table 3.7.

Metric	**Model Used**
Semantic Similarity	sentence-transformers/all-mpnet-base-v2
Natural Language Inference (NLI)	tasksource/deberta-small-long-nli
Fluency	gpt2-medium
BERTScore	facebook/bart-large-mnli

Table 3.7: Metrics their corresponding models, accessed via HuggingFace (Wolf et al., 2020).

All other metrics are based on the above. While these model-based metrics inherently carry model-specific biases, general trends across tasks should remain consistent even with different models. The metrics used in this project are:

3.4.1 BERTScore

$$\text{Precision} = \frac{1}{|C|} \sum_{c \in C} \max_{r \in R} \text{sim}(c, r) \tag{3.1}$$

$$\text{Recall} = \frac{1}{|R|} \sum_{r \in R} \max_{c \in C} \text{sim}(r, c) \tag{3.2}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

where C is the set of candidate tokens, R is the set of reference tokens, and $\text{sim}(c, r)$ is the cosine similarity between token embeddings c and r.

This evaluates the token-level similarity between generated and reference explanations by measuring precision, recall, and F1 score based on token embeddings (T. Zhang et al., 2020).

3.4.2 Fluency

$$\text{Fluency} = \max\left(0, 1 - \frac{\text{Perplexity}}{100}\right) \tag{3.4}$$

where Perplexity is calculated as the exponentiation of the negative log-likelihood loss of the language model on the tokenized generated text.

This measures the naturalness of the generated text. In other words, it evaluates how well the explanation aligns with typical language patterns.

3.4.3 Semantic Similarity

$$\text{Semantic Similarity}(T_1, T_2) = \frac{\mathbf{E}_1 \cdot \mathbf{E}_2}{\|\mathbf{E}_1\| \|\mathbf{E}_2\|} \tag{3.5}$$

where $\mathbf{E}_1$ and $\mathbf{E}_2$ are the mean-pooled embeddings of texts T_1 and T_2, $\cdot$ represents the dot product, and $\| \cdot \|$ is the L2 norm.

This evaluates how semantically close the generated text is to the reference text. Used as a construction metric.

3.4.4 NLI score

$$\text{NLI Score}(P, H) = \begin{cases} 2.5 \times \text{score}, & \text{if entailment} \\ 1 \times \text{score}, & \text{if neutral} \\ -5 \times \text{score}, & \text{if contradiction} \end{cases} \tag{3.6}$$

where P is the premise text, H is the hypothesis text, and score is the confidence score assigned by the NLI model for the predicted relation (entailment, neutral, or contradiction).

This evaluates the relationship between the premise and hypothesis, assigning different weightings depending on whether the hypothesis entails, is neutral to, or contradicts the premise. Also used as a construction metric.

3.4.5 Explanation Accuracy

$$\text{Explanation Accuracy}(G, R) = \text{Semantic Similarity}(G, R) \tag{3.7}$$

where G is the generated explanation, and R is the reference explanation.

This evaluates how semantically close the generated explanation is to the reference explanation.

3.4.6 Claim Accuracy

$$\text{Claim Accuracy}(C, G) = \text{Semantic Similarity}(C, G) \tag{3.8}$$

where C is the claim text, and G is the generated explanation.

This measures how closely the generated explanation reflects the semantics of the claim.

3.4.7 Claim Support

$$\text{Claim Support}(C, G) = \text{NLI Score}(G, C) \tag{3.9}$$

where C is the claim, and G is the generated explanation.

This assesses how well the generated explanation supports the claim.

3.4.8 Coherence

$$\text{Coherence}(G) = \frac{1}{n-1} \sum_{i=1}^{n-1} \text{NLI Score}(S_i, S_{i+1}) \tag{3.10}$$

where G is the generated explanation, S_i is the i-th sentence in the explanation, and n is the number of sentences in the explanation.

This measures the logical consistency between consecutive sentences in the generated explanation.

3.4.9 Fact Verification

$$\text{Fact Verification}(G, E) = \frac{1}{|E|} \sum_{e \in E} \max_{s \in \text{sent}(e)} \text{NLI Score}(s, G) \tag{3.11}$$

where G is the generated explanation, E is the set of evidence texts, and $\text{sent}(e)$ refers to the sentences in the evidence text e.

This evaluates how well the generated explanation aligns with all relevant evidence.

3.4.10 Explanation Completeness

$$\text{Explanation Completeness}(G, R) = \frac{1}{|R|} \sum_{r \in \text{sent}(R)} \max_{g \in \text{sent}(G)} \text{Semantic Similarity}(r, g) \tag{3.12}$$

where G is the generated explanation, R is the reference explanation, $\text{sent}(R)$ and $\text{sent}(G)$ are the sentences in the reference and generated explanations, respectively, and r and g are sentences in R and G.

This measures how much of the reference explanation is covered by the generated explanation.

3.5 Experimental Setup

The experimental setup evaluated LLMs' reasoning capabilities in explanation-based tasks, using consistent agent frameworks for task-specific outputs. The following subsections detail the chosen models, test workflow, and overall experiment structure.

3.5.1 Selected LLMs

The following Table 3.8 summarises the LLMs employed in the experiments, along with the agent architectures applied and the reasoning behind their selection:

Model	Agent(s)	Reason for Selection
GPT-4o	Self-Discover	Selected for its SOTA performance across multiple LLM tasks, including reasoning.
GPT-4o-mini	Self-Discover, Self-RAG	Selected to evaluate the impact of reduced model size on performance within the framework.

Table 3.8: LLMs and agents used in the experiments.

Note that the same Self-RAG agent was used across all relevant experiments for consistency.

GPT-3.5 Turbo and Claude 3 Haiku were initially included but later excluded from further analysis due to frequent hallucinations and inconsistent outputs. Results for these models are available in the project's code repository (Appendix A).

3.5.2 Testing Workflow

The testing process for each model follows a consistent three-step workflow:

1. **Additional Evidence Retrieval (Optional)**: For incomplete/mixed evidence tasks, self-RAG retrieved evidence for two Self-Discover agent questions, adding unique entries to the existing set.

2. **Answer Generation**: Models were given the task prompt for evidence selection and explanation generation. The solving process was consistent across models, focusing on reasoning through available evidence.

3. **Output Extraction**: Generated explanations were extracted using regex-assisted methods and stored in JSON format. Each scenario used a fixed pool of 30 test cases for consistency.

The evidence retrieval step is shown in Figure 3.10, and the testing workflow in Figure 3.9. Tested reasoning module combinations included Baseline, Explanatory, Coherence, and hybrid configurations (Baseline with Explanatory, Baseline with both Explanatory and Coherence), all integrated into a Self-Discover agent. Exact prompts are in Appendices E, C, and B. Agent architectures and testing workflows were implemented using LangChain (Chase, 2024), with original framework prompts used where appropriate (Zhou et al., 2024; Asai et al., 2023).

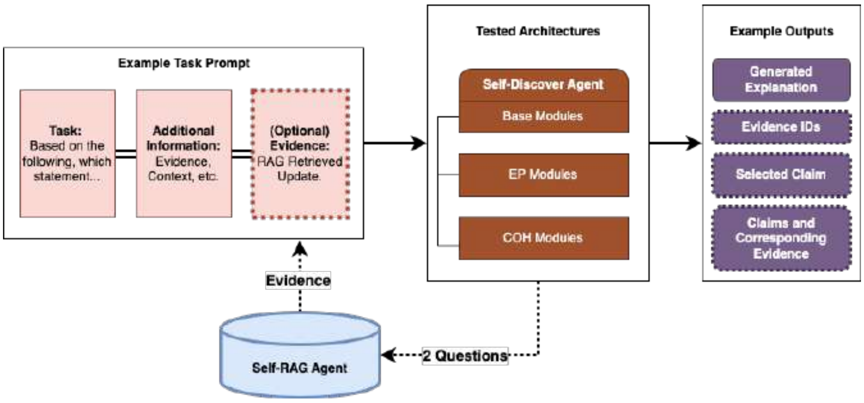

Figure 3.9: General Testing Workflow procedure.

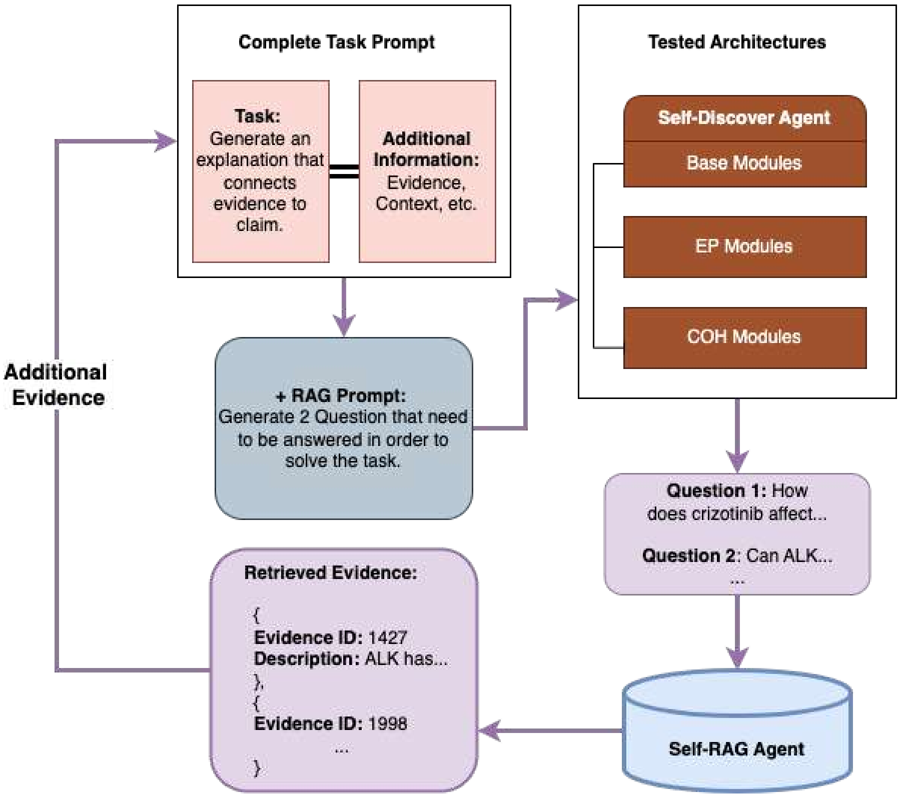
Complete Task Prompt
Task: Generate an explanation that connects evidence to claim.
Additional Information: Evidence, Context, etc.
Tested Architectures
Self-Discover Agent
Base Modules
EP Modules
COH Modules
Additional Evidence
+ RAG Prompt: Generate 2 Question that need to be answered in order to solve the task.
Question 1: How does crizotinib affect...
Question 2: Can ALK...
...
Retrieved Evidence:
{
Evidence ID: 1427
Description: ALK has...
},
{
Evidence ID: 1998
...
}
Self-RAG Agent

Figure 3.10: Evidence Retrieval process during tests.

Chapter 4

Results and Analysis

4.1 General Overview

This section presents relevant comparison tables of metric averages for the defined test scenarios. The best-scoring agent's value is emboldened, with relative gain compared to the baseline agent shown. A brief analysis highlights key findings, expanded upon in the next chapter. Supporting density plots are in Appendix D.

4.2 Explanation Test Results

4.2.1 Full Evidence

In the Full Evidence CIViC scenario, incorporating additional modules significantly improves *Coherence* and *Claim Support*. However, improvements in *BERTScore* and *Fact Verification* are marginal, with a slight decrease in the latter when adding multiple modules (Table 4.1).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.5899	0.5893	**0.5950 (+0.87%)**	0.5877	0.5899
BERTScore (Precision)	0.5740	**0.5836 (+1.67%)**	0.5836	0.5775	0.5737
BERTScore (Recall)	0.6093	0.5984	0.6099	0.6015	**0.6101 (+0.12%)**
Fluency	0.8077	0.7955	0.7979	0.7885	**0.8126 (+0.61%)**
Coherence	0.9005	0.7330	0.8655	**0.9503 (+5.54%)**	0.8948
Explanation Accuracy	0.8356	0.8223	0.8387	**0.8391 (+0.42%)**	0.8326
Explanation Completeness	0.6669	0.6488	0.6656	0.6653	**0.6724 (+0.82%)**
Claim Accuracy	0.8515	0.8506	0.8464	0.8515	**0.8580 (+0.76%)**
Claim Support	1.7955	1.8075	1.9917	**2.0633 (+14.91%)**	1.7745
Fact Verification	**1.0707**	1.0268	1.0518	1.0133	1.0140

Table 4.1: Full Evidence Results (CIViC) - GPT-4o.

For GPT-4o Mini, improvements are noticeable in *Coherence* and *Claim Accuracy* when incorporating epistemic modules. However, metrics such as *Fluency* and *Fact Verification* show a slight decline, indicating a trade-off when adding modules (Table 4.2).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	**0.5874**	0.5790	0.5865	0.5772	0.5843
BERTScore (Precision)	0.5771	0.5794	**0.5865 (+1.65%)**	0.5758	0.5828
BERTScore (Recall)	**0.6013**	0.5806	0.5888	0.5814	0.5887
Fluency	**0.8043**	0.7308	0.7796	0.7306	0.7704
Coherence	0.8369	**0.9129 (+14.37%)**	0.8692	0.9572	0.8945
Explanation Accuracy	0.8280	0.8183	0.8212	0.8299	**0.8359 (+0.95%)**
Explanation Completeness	0.6428	0.6388	0.6399	0.6411	**0.6437 (+0.15%)**
Claim Accuracy	0.8500	0.8611	0.8402	0.8629	**0.8678 (+2.10%)**
Claim Support	**2.0171**	1.8286	1.6119	1.8616	2.0065
Fact Verification	**1.0116**	0.9648	0.9769	0.9564	0.9536

Table 4.2: Full Evidence Results (CIViC) - GPT-4o Mini.

. The GPT-4o results on the R4C test show notable improvements in *Claim Support* and *Fact Verification* when incorporating epistemic modules. *BERTScores* improve, especially in *Recall*, but *Fluency* slightly decreases, and *Coherence* gains are modest (Table 4.3).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.6150	0.6110	**0.6307 (+2.55%)**	0.6020	0.6100
BERTScore (Precision)	0.5340	0.5254	**0.5438 (+1.85%)**	0.5207	0.5278
BERTScore (Recall)	0.7302	0.7349	**0.7553 (+3.44%)**	0.7185	0.7263
Fluency	**0.7899**	0.7631	0.7787	0.7873	0.7700
Coherence	0.9045	0.7179	0.7417	0.8967	**0.9177 (+1.46%)**
Explanation Accuracy	0.8065	0.7981	**0.8217 (+1.89%)**	0.8048	0.8144
Explanation Completeness	0.7628	0.7593	**0.7955 (+4.29%)**	0.7746	0.7833
Claim Accuracy	0.7975	0.7896	**0.8095 (+1.50%)**	0.7922	0.8016
Claim Support	0.6528	0.5373	0.6371	**0.9701 (+48.61%)**	0.7327
Fact Verification	1.0806	1.2344	**1.2667 (+17.22%)**	0.9650	1.2528

Table 4.3: Full Evidence Results (R4C) - GPT-4o.

For GPT-4o Mini, significant improvements are observed in *Claim Support* and *Fact Verification* with additional modules. While *BERTScore* and *Explanation Completeness* show slight gains, *Explanation Accuracy* decreases, indicating potential trade-offs (Table 4.4).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.6260	0.6081	**0.6346 (+1.36%)**	0.6063	0.6117
BERTScore (Precision)	0.5423	0.5218	**0.5491 (+1.26%)**	0.5163	0.5240
BERTScore (Recall)	0.7456	0.7362	**0.7557 (+1.35%)**	0.7381	0.7399
Fluency	0.7505	0.7575	0.7587	**0.7805 (+4.00%)**	0.7654
Coherence	**1.0079**	0.9378	0.9720	0.8780	0.9140
Explanation Accuracy	**0.8091**	0.7777	0.8063	0.7858	0.7569
Explanation Completeness	0.7690	**0.7803 (+1.56%)**	0.7769	0.7810	0.7631
Claim Accuracy	0.8032	0.7740	**0.8118 (+1.08%)**	0.7872	0.7690
Claim Support	0.7453	0.7999	0.6668	0.9056	**0.9281 (+24.53%)**
Fact Verification	0.9380	1.1104	**1.1477 (+22.35%)**	0.9332	0.9920

Table 4.4: Full Evidence Results (R4C) - GPT-4o Mini.

4.2.2 Missing Evidence

In the Missing Evidence scenario for GPT-4o, improvements are seen in *Fluency*, *Explanation Completeness*, and *Fact Verification* with additional modules. *BERTScore* and *Coherence* show minimal changes, suggesting better factual alignment but limited gains in coherence (Table 4.5).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	**0.5946**	0.5886	0.5946	0.5878	0.5917
BERTScore (Precision)	0.5792	0.5767	**0.5861 (+1.19%)**	0.5722	0.5807
BERTScore (Recall)	**0.6131**	0.6037	0.6059	0.6076	0.6072
Fluency	0.7983	0.8066	0.7996	**0.8189 (+2.58%)**	0.7740
Coherence	**0.8515**	0.8253	0.7953	0.7868	0.8089
Explanation Accuracy	0.8313	0.8308	0.8324	0.8373	**0.8405 (+1.11%)**
Explanation Completeness	0.6622	0.6634	0.6602	0.6671	**0.6771 (+2.25%)**
Claim Accuracy	0.8483	0.8477	**0.8510 (+0.31%)**	0.8434	0.8490
Claim Support	1.8774	1.7140	1.8164	1.7415	**2.0106 (+7.09%)**
Fact Verification	1.0035	1.0278	0.9918	**1.0439 (+4.02%)**	1.0425

Table 4.5: Missing Evidence Results (CIViC) - GPT-4o.

For GPT-4o Mini, *Coherence* shows a significant improvement, while *BERTScore Precision* increases slightly with additional modules. However, *Fluency* and *Explanation Completeness* decline slightly, indicating improvements in some metrics come at a cost to others (Table 4.6).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.5857	0.5830	**0.5875 (+0.31%)**	0.5780	0.5852
BERTScore (Precision)	0.5774	0.5815	0.5864	0.5734	**0.5870 (+1.66%)**
BERTScore (Recall)	**0.5971**	0.5877	0.5913	0.5855	0.5862
Fluency	**0.7905**	0.7344	0.7899	0.7845	0.7561
Coherence	0.8609	**0.9147 (+9.71%)**	0.8015	0.9445	0.8993
Explanation Accuracy	0.8269	0.8200	0.8276	0.8235	**0.8320 (+0.62%)**
Explanation Completeness	**0.6583**	0.6426	0.6415	0.6463	0.6516
Claim Accuracy	**0.8623**	0.8513	0.8515	0.8599	0.8509
Claim Support	**1.9640**	1.6375	1.6291	1.8742	1.9158
Fact Verification	1.0111	0.9825	0.9957	0.9836	**1.0146 (+0.35%)**

Table 4.6: Missing Evidence Results (CIViC) - GPT-4o Mini.

4.2.3 Wrong Evidence

In the Wrong Evidence scenario for GPT-4o, improvements are notable in *Claim Support* and *Fact Verification* with the *COH* module, enhancing alignment despite incorrect information. However, *Coherence* declines slightly, indicating challenges in maintaining consistency with incorrect evidence (Table 4.7).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.5933	0.5872	0.5920	0.5904	**0.5948 (+0.24%)**
BERTScore (Precision)	**0.5847**	0.5769	0.5797	0.5820	0.5775
BERTScore (Recall)	0.6060	0.6008	0.6077	0.6020	**0.6153 (+1.54%)**
Fluency	0.7825	0.8039	0.7961	**0.8116 (+3.73%)**	0.8072
Coherence	**0.9073**	0.7132	0.7590	0.8973	0.7814
Explanation Accuracy	0.8319	**0.8401 (+1.17%)**	0.8276	0.8290	0.8416
Explanation Completeness	0.6588	**0.6696 (+1.63%)**	0.6631	0.6614	0.6695
Claim Accuracy	0.8499	0.8531	0.8521	**0.8599 (+1.18%)**	0.8509
Claim Support	1.7982	1.6856	**2.1460 (+19.34%)**	1.8550	1.6562
Fact Verification	0.9942	1.0248	**1.0348 (+4.08%)**	1.0135	1.0174

Table 4.7: Wrong Evidence Results (CIViC) - GPT-4o.

For GPT-4o Mini, *Coherence* improves with the *Baseline+EXP+COH* combination, while *BERTScore Precision* rises slightly, reflecting better token-level alignment. However, *Fluency* and *Fact Verification* decline slightly, indicating trade-offs in generating natural and factually accurate explanations with additional modules (Table 4.8).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.5853	0.5779	**0.5864 (+0.19%)**	0.5825	0.5844
BERTScore (Precision)	0.5775	0.5789	**0.5892 (+2.02%)**	0.5820	0.5850
BERTScore (Recall)	**0.5965**	0.5787	0.5866	0.5855	0.5864
Fluency	**0.7828**	0.7652	0.7730	0.7551	0.7709
Coherence	0.8988	0.8890	0.9150	0.8574	**0.9289 (+3.35%)**
Explanation Accuracy	0.8171	0.8180	0.8230	**0.8255 (+1.03%)**	0.8211
Explanation Completeness	0.6436	0.6377	**0.6525 (+1.39%)**	0.6444	0.6340
Claim Accuracy	0.8478	0.8582	0.8409	**0.8677 (+2.35%)**	0.8405
Claim Support	**2.0892**	1.7562	1.5331	1.9979	1.7848
Fact Verification	**1.0169**	0.9651	0.9684	0.9669	0.9587

Table 4.8: Wrong Evidence Results (CIViC) - GPT-4o Mini.

4.2.4 Mixed Evidence

In the Mixed Evidence scenario for GPT-4o, *Coherence* improves significantly with the *COH* module, while *Fact Verification* also increases noticeably. *Fluency* and *Explanation Completeness* show slight gains, indicating more natural and complete explanations (Table 4.9).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.5903	0.5866	**0.5949 (+0.77%)**	0.5813	0.5894
BERTScore (Precision)	0.5727	0.5730	**0.5872 (+2.54%)**	0.5684	0.5759
BERTScore (Recall)	**0.6113**	0.6036	0.6057	0.5967	0.6072
Fluency	0.8042	0.7971	0.7890	0.7988	**0.8120 (+0.98%)**
Coherence	0.7362	0.7257	**0.8873 (+20.52%)**	0.7138	0.8148
Explanation Accuracy	0.8308	0.8347	**0.8406 (+1.18%)**	0.8184	0.8304
Explanation Completeness	0.6645	0.6714	0.6668	0.6544	**0.6754 (+1.65%)**
Claim Accuracy	0.8497	0.8555	**0.8562 (+0.77%)**	0.8548	0.8533
Claim Support	1.7643	**2.0178 (+14.37%)**	1.9900	1.7381	1.9203
Fact Verification	1.0103	1.0520	**1.0848 (+7.37%)**	1.0093	1.0207

Table 4.9: Mixed Evidence Results (CIViC) - GPT-4o.

For GPT-4o Mini, *BERTScore Precision* improves slightly, indicating better token alignment, and *Claim Accuracy* increases by a few percent. However, *Fluency* declines, and *Coherence* remains high but drops slightly from the *Baseline* (Table 4.10).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	0.5841	0.5803	0.5866	0.5734	**0.5899 (+1.00%)**
BERTScore (Precision)	0.5752	0.5856	0.5903	0.5708	**0.5908 (+2.71%)**
BERTScore (Recall)	**0.5962**	0.5778	0.5857	0.5780	0.5910
Fluency	**0.8032**	0.7330	0.7627	0.7602	0.7713
Coherence	**0.9437**	0.9087	0.9002	0.8626	0.8814
Explanation Accuracy	0.8262	0.8200	0.8253	0.8141	**0.8351 (+1.08%)**
Explanation Completeness	0.6444	0.6397	**0.6486 (+0.64%)**	0.6360	0.6440
Claim Accuracy	0.8470	0.8558	0.8516	0.8443	**0.8641 (+2.02%)**
Claim Support	**2.0385**	1.6233	1.4283	1.6386	1.6760
Fact Verification	**1.0230**	0.9646	0.9978	0.9802	0.9877

Table 4.10: Mixed Evidence Results (CIViC) - GPT-4o Mini.

4.3 Selection Test Results

In the Selection scenario for GPT-4o, *Coherence* improves slightly, and *Fact Verification* sees a small gain. However, *Fluency* and other metrics drop noticeably, indicating limited benefit from additional modules (Table 4.11).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	**0.6019**	0.5826	0.5864	0.5789	0.5921
BERTScore (Precision)	**0.5985**	0.5779	0.5910	0.5775	0.5912
BERTScore (Recall)	**0.6087**	0.5894	0.5839	0.5822	0.5947
Fluency	**0.7336**	0.6420	0.6870	0.6736	0.6177
Coherence	0.9258	0.8155	0.8635	0.8117	**0.9306 (+0.53%)**
Explanation Accuracy	**0.8330**	0.8096	0.7925	0.7913	0.8041
Explanation Completeness	**0.6473**	0.6235	0.6275	0.6206	0.6172
Claim Accuracy	**0.8360**	0.7947	0.7930	0.7896	0.7911
Claim Support	**1.8905**	1.7875	1.6589	1.4952	1.7969
Fact Verification	0.9516	0.9359	0.9254	0.9415	**0.9583 (+0.70%)**

Table 4.11: Selection Results (CIViC) - GPT-4o.

For GPT-4o Mini, the addition of modules decreases performance across all metrics, indicating difficulties in handling tasks with higher contextual complexity and reasoning demands (Table 4.12).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	**0.5919**	0.5784	0.5750	0.5708	0.5846
BERTScore (Precision)	**0.5842**	0.5761	0.5752	0.5654	0.5803
BERTScore (Recall)	**0.6024**	0.5833	0.5773	0.5788	0.5919
Fluency	**0.7405**	0.6013	0.5950	0.6050	0.6424
Coherence	**0.8960**	0.8722	0.7344	0.8189	0.8376
Explanation Accuracy	**0.8207**	0.7830	0.7849	0.7603	0.7648
Explanation Completeness	**0.6336**	0.6187	0.5973	0.6029	0.6027
Claim Accuracy	**0.8217**	0.7934	0.8100	0.7705	0.7736
Claim Support	**1.6221**	1.5669	1.5957	1.6005	1.2304
Fact Verification	**0.9582**	0.9162	0.9006	0.9059	0.9305

Table 4.12: Selection Results (CIViC) - GPT-4o Mini.

4.4 Assignment Test Results

In the Assignment scenario for GPT-4o, *Fluency* improves slightly with the *COH* module, and *Claim Support* rises slightly. *Coherence* remains strong but decreases with additional modules, showing limited benefits overall (Table 4.13).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	**0.6231**	0.6035	0.6135	0.5950	0.6099
BERTScore (Precision)	**0.6288**	0.6145	0.6272	0.6074	0.6155
BERTScore (Recall)	**0.6195**	0.5947	0.6022	0.5863	0.6058
Fluency	0.7833	0.7888	**0.8010 (+2.27%)**	0.7562	0.7951
Coherence	**0.9543**	0.9099	0.8588	0.9320	0.8357
Explanation Accuracy	**0.8579**	0.8495	0.8515	0.8244	0.8527
Explanation Completeness	**0.6152**	0.6041	0.6050	0.5928	0.6123
Claim Accuracy	**0.8815**	0.8815	0.8722	0.8429	0.8658
Claim Support	2.1974	2.1209	**2.2451 (+2.17%)**	2.0689	2.1627
Fact Verification	**1.0498**	0.9904	0.9870	0.9805	0.9604

Table 4.13: Assignment Results (CIViC) - GPT-4o.

For GPT-4o Mini, *Coherence* improves significantly with the *EXP* module, and *Fluency* rises by a few percent. However, *Explanation Completeness* and *Accuracy* decline significantly with the *Baseline+EXP+COH* combination, indicating module-dependent trade-offs (Table 4.14).

Metrics	Modules				
	Baseline	EXP	COH	Baseline+EXP	Baseline+EXP+COH
BERTScore (F1)	**0.6133**	0.5912	0.5801	0.5888	0.5477
BERTScore (Precision)	**0.6182**	0.6023	0.6022	0.6005	0.5662
BERTScore (Recall)	**0.6108**	0.5820	0.5616	0.5792	0.5327
Fluency	0.7650	**0.7830 (2.35%)**	0.7791	0.7675	0.7640
Coherence	0.8143	**0.9259 (13.70%)**	0.8692	0.9173	0.8356
Explanation Accuracy	**0.8532**	0.8415	0.7740	0.8331	0.6154
Explanation Completeness	**0.6000**	0.5919	0.5534	0.5864	0.4369
Claim Accuracy	0.8679	**0.8756 (0.89%)**	0.8102	0.8673	0.6351
Claim Support	**2.0953**	1.9363	1.8560	2.0712	1.6396
Fact Verification	**1.0758**	0.9875	0.9400	0.9568	0.9315

Table 4.14: Assignment Results (CIViC) - GPT-4o Mini.

Chapter 5

Discussion

5.1 Impact of Epistemic Modules on Explanation Quality

This study examined integrating Explanatory Power and Coherence modules into LLMs across various evidence scenarios. These modules generally improved explanation quality, especially in coherence and claim alignment, though extent varied by task and metric. In the Full Evidence scenario with CIViC data, GPT-4o saw notable gains in Coherence and Claim Support, suggesting modules helped produce more structured and aligned explanations. However, BERTScore and Explanation Accuracy showed marginal improvements, while Fact Verification slightly declined with additional modules.

For GPT-4o Mini, similar trends emerged, though improvements were less pronounced due to its size. Coherence and Claim Accuracy improved with epistemic modules, but Fluency and Fact Verification often declined, indicating a trade-off between explanation structure and text naturalness, especially in smaller models.

In the R4C variation, both GPT-4o and GPT-4o Mini showed stronger improvements across most metrics, particularly Claim Support and Fact Verification. This suggests the modules helped distil key information from complex, noisy statements, improving explanation quality and coherence, especially with the COH module. These results indicate the modules excel in both specialised medical contexts and general reasoning tasks.

In challenging scenarios like Missing and Wrong Evidence, modules improved Explanation Completeness, Fact Verification, and Claim Support, particularly for GPT-4o. However, other metrics saw limited change, highlighting selective effectiveness. While enhancing coherence and explanatory quality, modules may struggle with complex multi-step reasoning when evidence is incomplete or incorrect.

5.2 Effect on Evidence Selection and Utilisation

The EXP and COH modules aimed to enhance LLMs' evidence handling during explanation generation. In the Wrong Evidence scenario, modules significantly boosted evidence-related metrics, enabling better identification, retrieval, and alignment of relevant evidence with claims, even when incorrect information was present. This indicates modules help prioritise relevant over irrelevant evidence, though at the expense of Fluency and overall Coherence.

In the Mixed Evidence scenario, GPT-4o's most significant improvements were in Coherence and Fact Verification, confirming epistemic modules help filter and effectively use correct evidence despite irrelevant data. GPT-4o Mini showed modest gains, though its size limited ability to handle complex evidence scenarios. Modules provided clearer benefits in scenarios requiring higher reasoning complexity, highlighting their role in improving evidence selection under challenging conditions.

5.3 Trade-offs Between Metrics

A clear theme is the trade-off between metrics when incorporating epistemic modules. When linguistic metrics improved, explanatory metrics often declined, and vice versa. This suggests EXP and COH modules inherently favour one aspect over another, challenging simultaneous optimization of both within the same framework.

The interaction between baseline and epistemic modules is noteworthy. The Baseline+EXP+COH combination generally provided best results for most metrics. However, individual modules often performed best in several scenarios, indicating performance and utilization are heavily task-dependent. In some tasks, these modules performed worse than baseline, showcasing that even SOTA LLMs can be overwhelmed by specialized concepts compared to general approaches.

5.4 Performance in Selection and Assignment Tasks

In contextually difficult tasks like Selection and Assignment Tests, epistemic modules showed limited benefit. In the Selection scenario, Coherence and Fact Verification improved slightly, but other metrics declined, suggesting modules struggle with nuanced evidence differentiation. For GPT-4o Mini, performance decreased across all metrics with additional modules, highlighting its limitations in complex reasoning tasks. This indicates epistemic modules are helpful in standard reasoning tasks, but more contextually and information-demanding tests may exceed their current capabilities, particularly for smaller models.

In the Assignment Test, modules slightly improved few metrics, but baseline agent mostly outperformed them. This underscores the difficulty in maintaining high-quality explanations when assigning multiple evidence sets to competing claims, especially for smaller models. It suggests epistemic modules may require additional tuning or architectural changes to effectively handle more complex multi-step reasoning tasks.

5.5 Model Size and Sequential Models

Model size significantly affected epistemic modules' effectiveness. GPT-4o consistently outperformed GPT-4o Mini across most scenarios, especially in complex reasoning tasks. GPT-4o Mini struggled to maintain acceptable quality in more complex tasks, suggesting model capacity limits integration and benefits from additional reasoning modules.

This indicates epistemic modules can enhance reasoning in both small and large models, but their effectiveness is proportional to model size. Larger models can handle additional complexity introduced by these modules, whereas smaller models may need simpler or more efficient implementations to avoid degradation. The one-shot testing scenario highlights potential for sequential models to ease these issues. A sequential model could refine reasoning iteratively, potentially improving explanation quality without sacrificing efficiency. An example of such implementation is included in Appendix E.

5.6 General Summary

This chapter examined effects of adding Explanatory Power and Coherence modules to LLMs across various evidence scenarios. These modules generally enhance explanation quality, particularly in generic reasoning cases with full evidence. However, improvements in some metrics, like BERTScore, were often limited, with visible trade-offs between explanatory and coherence-based metrics.

Epistemic modules' impact was more pronounced in complex scenarios like Mixed and Wrong Evidence, with notable improvements in explanation-based metrics. However, the smaller model showed reduced benefits, highlighting trade-offs between model size and performance.

The study underscores challenges in balancing metric improvements when adding epistemic modules. Their advantages in reasoning quality are task-dependent and may decrease linguistic quality, especially in smaller models.

Chapter 6

Conclusions, Limitations, and Recommendations

6.1 Conclusions

This study shows that integrating Explanatory Power and Coherence modules into LLMs enhances reasoning and explanation quality, especially in complex evidence scenarios. Improvements in Coherence, Claim Support, and Fact Verification are notable, but impact depends on specific context and task. For example, GPT-4o improved in coherence and claim alignment in one scenario, but had only marginal gains in BERTScore or Fact Verification.

A key insight is the selective effectiveness of these epistemic modules. They can significantly improve logical structure and relevance of explanations, but often introduce trade-offs in fluency or factual consistency. This is particularly evident in smaller models like GPT-4o Mini, where gains in reasoning quality were often offset by declines in other metrics, reflecting limitations of reduced model size and capacity. These results indicate that while epistemic modules can powerfully improve explanation generation, their benefits must be weighed against risks of diminished performance in other areas.

Enhancing reasoning and explanatory capacity requires a balanced approach addressing both reasoning depth and surface-level fluency and accuracy. Task complexity crucially determines the modules' efficacy, with demanding tasks like evidence differentiation and multi-step reasoning in Selection and Assignment tests highlighting areas for improvement.

6.2 Limitations

API costs and time constraints limited the study's scope, restricting the number of tests and LLMs evaluated. This reduced scenario and model diversity. A more comprehensive experiment set, including a broader range of models, would have allowed for more robust analysis of epistemic modules' impact across different architectures and task types.

The research's novelty necessitated developing study-specific scoring methods due to lack of established frameworks for evaluating LLM-generated explanations semantically. While sufficient for this analysis, these methods need further refinement and validation for broader applicability. Task complexity required larger models like GPT-4o for reliable results, as smaller models were prone to hallucinations and inconsistencies. This reliance on larger models limits findings' generalisability to smaller, resource-constrained systems.

Due to limited epistemology expertise, the resultant modules were relatively simple, inspired by epistemic concepts rather than representing their optimal forms. This simplification may have influenced the discussed results.

6.3 Recommendations

Building on these findings, future research should explore:

1. Expanding beyond the medical context to test epistemic modules in areas like legal reasoning, education, and scientific tasks, providing insights into generalisability.

2. Incorporating more SOTA LLMs to determine how different architectures respond to these modules.

More varied test cases should be explored. Future work should examine variations like gradual evidence introduction, larger knowledge bases, or shorter evidence statements. Testing across different evidence conditions could reveal additional insights. Sequential models, allowing iterative refinement of reasoning and explanations, should be employed to potentially mitigate observed trade-offs between metrics. Multi-agent systems, with specialised agents collaborating or supervising, could enhance explanation quality by distributing reasoning tasks. Other reasoning frameworks besides Self-Discover should also be adapted and investigated to increase research depth.

Optimising these modules for smaller models should be prioritised. Simplified implementations could allow smaller models to achieve better performance without sacrificing explanation quality, extending applicability to resource-constrained environments. A more extensive and rigorous implementation based on expert opinion would also ensure optimal explanatory quality.

In conclusion, while this study highlights epistemic modules' potential to enhance LLM reasoning in complex tasks, further refinement is necessary to balance improvements across all metrics. Future research should explore new contexts, test cases, and model architectures to understand system trade-offs. This can lead to more robust, reliable reasoning frameworks, improving LLMs' trustworthiness and utility in decision-making scenarios.

Bibliography

Ajwani, Rohan et al. (2024). *LLM-Generated Black-box Explanations Can Be Adversarially Helpful*. arXiv: 2405.06800 [cs.CL]. URL: https://arxiv.org/abs/2405.06800.

Amaya, Amalia (2007). "Formal Models of Coherence and Legal Epistemology". In: *Artificial Intelligence and Law* 15.4, pp. 429–447. DOI: 10.1007/s10506-007-9050-4.

Anthropic (2024). *Claude [Large language model]*. URL: https://www.anthropic.com/claude.

Asai, Akari et al. (2023). *Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection*. arXiv: 2310.11511 [cs.CL]. URL: https://arxiv.org/abs/2310.11511.

Chase, Harrison (2024). *LangChain*. Version released on 2024-09-09. URL: https://github.com/langchain-ai/langchain.

Dalvi, Bhavana et al. (2022). *Explaining Answers with Entailment Trees*. arXiv: 2104.08661 [cs.CL]. URL: https://arxiv.org/abs/2104.08661.

Douven, Igor and Jonah N. Schupbach (2015). "The Role of Explanatory Considerations in Updating". In: *Cognition* 142.C, pp. 299–311. DOI: 10.1016/j.cognition.2015.04.017.

Glass, David H. (2007). "Coherence Measures and Inference to the Best Explanation". In: *Synthese* 157.3, pp. 275–296. DOI: 10.1007/s11229-006-9055-7.

— (2023). "How Good is an Explanation?" In: *Synthese* 201.2, pp. 1–26. DOI: 10.1007/s11229-022-04025-x.

Good, Benjamin et al. (Aug. 2014). "Organizing knowledge to enable personalization of medicine in cancer". In: *Genome biology* 15, p. 438. DOI: 10.1186/s13059-014-0438-7.

Hansson, Sven Ove and Erik J. Olsson (1999). "Providing Foundations for Coherentism". In: *Erkenntnis* 51.2-3, pp. 243–265. DOI: 10.1023/a:1005510414170.

Inoue, Naoya, Pontus Stenetorp, and Kentaro Inui (July 2020). “R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 6740–6750. DOI: 10.18653/v1/2020.acl-main.602. URL: https://aclanthology.org/2020.acl-main.602.

Jhamtani, Harsh and Peter Clark (2020). *Learning to Explain: Datasets and Models for Identifying Valid Reasoning Chains in Multihop Question-Answering*. arXiv: 2010.03274 [cs.CL]. URL: https://arxiv.org/abs/2010.03274.

Kojima, Takeshi et al. (2023). *Large Language Models are Zero-Shot Reasoners*. arXiv: 2205.11916 [cs.CL]. URL: https://arxiv.org/abs/2205.11916.

Kunz, Jenny and Marco Kuhlmann (2024). *Properties and Challenges of LLM-Generated Explanations*. arXiv: 2402.10532 [cs.CL]. URL: https://arxiv.org/abs/2402.10532.

Morishita, Terufumi et al. (2023). *Learning Deductive Reasoning from Synthetic Corpus based on Formal Logic*. arXiv: 2308.07336 [cs.AI]. URL: https://arxiv.org/abs/2308.07336.

Olsson, Erik (2023). “Coherentist Theories of Epistemic Justification”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University.

OpenAI (2024). *GPT-4 [Large language model]*. URL: https://openai.com/research/gpt-4.

Schupbach, Jonah N. and Jan Sprenger (2011). “The Logic of Explanatory Power*”. In: *Philosophy of Science* 78.1, pp. 105–127. ISSN: 00318248, 1539767X. URL: https://www.jstor.org/stable/10.1086/658111 (visited on 09/09/2024).

Shinn, Noah et al. (2023). *Reflexion: Language Agents with Verbal Reinforcement Learning*. arXiv: 2303.11366 [cs.AI]. URL: https://arxiv.org/abs/2303.11366.

Wang, Boshi, Xiang Yue, and Huan Sun (2023). *Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate*. arXiv: 2305.13160 [cs.CL]. URL: https://arxiv.org/abs/2305.13160.

Wolf, Thomas et al. (2020). *HuggingFace’s Transformers: State-of-the-art Natural Language Processing*. arXiv: 1910.03771 [cs.CL]. URL: https://arxiv.org/abs/1910.03771.

Xu, Binfeng et al. (2023). *ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models.* arXiv: 2305.18323 [cs.CL]. URL: https://arxiv.org/abs/2305.18323.

Yang, Zhilin et al. (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Yao, Shunyu et al. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models.* arXiv: 2210.03629 [cs.CL]. URL: https://arxiv.org/abs/2210.03629.

Zhang, Tianyi et al. (2020). *BERTScore: Evaluating Text Generation with BERT.* arXiv: 1904.09675 [cs.CL]. URL: https://arxiv.org/abs/1904.09675.

Zhang, Yadong et al. (2024). *LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models.* arXiv: 2404.01230 [cs.CL]. URL: https://arxiv.org/abs/2404.01230.

Zhou, Pei et al. (2024). *Self-Discover: Large Language Models Self-Compose Reasoning Structures.* arXiv: 2402.03620 [cs.AI]. URL: https://arxiv.org/abs/2402.03620.

Appendix A

Code Repository

A.1 Code Repository

The project repository, available at `https://github.com/rk-izak/DATA72000-IBE/tree/main`, contains the source code, data handling and plotting scripts, experimental setups, prompts, and reasoning modules used within this project. For more information, or to reproduce the results, the reader is advised to access the repository first.

A.1.1 Code Attribution

This core libraries and frameworks used to enable interactions with language models include:

- **LangChain AI**: Powered most architectures/agents and workflows via the LangChain framework `https://www.langchain.com/` and LangGraph `https://langchain-ai.github.io/langgraph/`.

- **Self-Discover**: Adapted from LangChain's Self-Discover Tutorial `https://langchain-ai.github.io/langgraph/tutorials/self-discover/self-discover/`, inspired by `https://github.com/catid/self-discover/tree/main?tab=readme-ov-file`. See Self-Discover Paper (Zhou et al., 2024).

- **Self-RAG**: Version based on the LangChain Tutorial `https://langchain-ai.github.io/langgraph/tutorials/rag/langgraph_self_rag/`. Refer to the Self-RAG paper (Asai et al., 2023).

- **OpenAI/Anthropic LLMs**: Models used: `GPT-4o`, `GPT-4o-mini`, `GPT-3.5-Turbo`, and `Claude3 Haiku`. See relevant table below for details.

As such, the relevant prompts used for either Self-Discover or Self-RAG agent workflows have not been modified from their original sources. These can be seen in later appendices.

Additional libraries supporting the project include:

- Transformers (Huggingface)
- PyTorch
- Pandas
- Matplotlib
- Seaborn
- NLTK
- Scikit-learn
- SciPy

For an exhaustive list, refer to the `requirements.txt` in the repository or the relevant `README.md` files.

The exact LLM snapshots used for the project are as follows:

Model Snapshot	**Company**
`gpt-4o-2024-08-06`	OpenAI
`gpt-4o-mini-2024-07-18`	OpenAI
`gpt-3.5-turbo-0125`	OpenAI
`claude-3-haiku-20240307`	Anthropic

Table A.1: LLM snapshots.

A.1.2 Data Sources and Availability

The main data sources for this project include publicly available knowledge bases and benchmarks:

- CIViC (Clinical Interpretation of Variants in Cancer) `https://civicdb.org/`
- R4C (Right for the Right Reason RC) `https://naoya-i.github.io/r4c/`
- HotpotQA (Explainable Multi-hop QA) `https://hotpotqa.github.io/`

Appendix B

Self-Discover Agent Materials

B.1 Reasoning Modules

Baseline Reasoning Modules (Part 1)

- **B1**: How could I devise an experiment to help solve that problem?
- **B2**: Make a list of ideas for solving this problem, and apply them one by one to the problem to see if any progress can be made.
- **B3**: How could I measure progress on this problem?
- **B4**: How can I simplify the problem so that it is easier to solve?
- **B5**: What are the key assumptions underlying this problem?
- **B7**: What are the alternative perspectives or viewpoints on this problem?
- **B8**: What are the long-term implications of this problem and its solutions?
- **B10**: **Critical Thinking**: This style involves analyzing the problem from different perspectives, questioning assumptions, and evaluating the evidence or information available. It focuses on logical reasoning, evidence-based decision-making, and identifying potential biases or flaws in thinking.
- **B13**: **Use Systems Thinking**: Consider the problem as part of a larger system, understanding the interconnectedness of various elements. Focus on identifying the underlying causes, feedback loops, and interdependencies that influence the problem, and developing holistic solutions that address the system as a whole.
- **B15**: **Use Reflective Thinking**: Step back from the problem, take time for introspection and self-reflection. Examine personal biases, assumptions, and mental models that may influence problem-solving, and being open to learning from past experiences to improve future approaches.
- **B16**: What is the core issue or problem that needs to be addressed?
- **B17**: What are the underlying causes or factors contributing to the problem?
- **B18**: Are there any potential solutions or strategies that have been tried before? If yes, what were the outcomes and lessons learned?

Figure B.1: Chosen Baseline Reasoning Modules (Part 1).

Baseline Reasoning Modules (Part 2)

- **B19**: What are the potential obstacles or challenges that might arise in solving this problem?
- **B20**: Are there any relevant data or information that can provide insights into the problem? If yes, what data sources are available, and how can they be analyzed?
- **B21**: Are there any stakeholders or individuals who are directly affected by the problem? What are their perspectives and needs?
- **B23**: How can progress or success in solving the problem be measured or evaluated?
- **B24**: What indicators or metrics can be used?
- **B25**: Is the problem a technical or practical one that requires a specific expertise or skill set? Or is it more of a conceptual or theoretical problem?
- **B27**: Is the problem related to human behavior, such as a social, cultural, or psychological issue?
- **B28**: Does the problem involve decision-making or planning, where choices need to be made under uncertainty or with competing objectives?
- **B29**: Is the problem an analytical one that requires data analysis, modeling, or optimization techniques?
- **B32**: Is the problem time-sensitive or urgent, requiring immediate attention and action?
- **B33**: What kinds of solutions typically are produced for this kind of problem specification?
- **B34**: Given the problem specification and the current best solution, have a guess about other possible solutions.
- **B35**: Let's imagine the current best solution is totally wrong. What other ways are there to think about the problem specification?
- **B36**: What is the best way to modify the current best solution, given what you know about these kinds of problem specifications?
- **B37**: Ignoring the current best solution, create an entirely new solution to the problem.
- **B38**: Let's think step by step.
- **B39**: Let's make a step-by-step plan and implement it with clear reasoning and explanation.

Figure B.2: Chosen Baseline Reasoning Modules (Part 2).

Explanatory Power Reasoning Modules

- **E1**: Compare the explanatory power of multiple hypotheses for the same set of evidence. Which one is superior?
- **E2**: Does the hypothesis reduce the surprise associated with the evidence?
- **E3**: What are the key factors that determine the explanatory power of a hypothesis?
- **E4**: How would introducing an alternative hypothesis affect the explanatory power of the current explanation?
- **E5**: Assess the simplicity of the hypothesis and its relationship to explanatory power.
- **E6**: What is the role of background knowledge in determining the explanatory power of a hypothesis?
- **E7**: What are the minimal assumptions required for this hypothesis to have appropriate explanatory power?
- **E8**: Determine if the hypothesis is the best explanation based on its ability to explain the evidence better than alternatives.
- **E9**: Does the hypothesis have strong explanatory power, or are there significant gaps in the explanation?
- **E10**: What additional evidence would increase the explanatory power of the hypothesis?
- **E11**: Does the hypothesis provide a more powerful explanation compared to others? Why or why not?
- **E12**: Evaluate how well the hypothesis explains both the evidence and excludes alternative explanations.
- **E13**: How can the explanatory power of a hypothesis be tested empirically?
- **E14**: Identify any biases that may affect the perceived explanatory power of the hypothesis.
- **E15**: Are there any ways to enhance the explanatory power of the current hypothesis?
- **E16**: Analyze the trade-offs between explanatory power and the simplicity of a hypothesis.
- **E17**: What is the significance of the explanatory power of a hypothesis in making predictive inferences?
- **E18**: Does this hypothesis explain the evidence better than a probabilistic or causal model?
- **E19**: Consider how the explanatory power of a hypothesis might change with new or additional evidence.
- **E20**: What are the implications of the hypothesis if it provides high explanatory power?
- **E21**: How does explanatory power affect the overall credibility of a hypothesis?
- **E22**: What conditions must be met for a hypothesis to have maximum explanatory power?
- **E23**: Assess whether the hypothesis is comprehensive enough to cover all aspects of the evidence.
- **E24**: Determine the robustness of the explanatory power when applied to varied evidence.
- **E25**: Is there a simpler hypothesis that provides similar explanatory power? If so, what are its implications?
- **E26**: How does the prior probability of a hypothesis affect its explanatory power?
- **E27**: How does the explanatory power of the hypothesis relate to its explanatory scope?
- **E28**: What role does coherence play in determining the explanatory power of a hypothesis?
- **E29**: Evaluate the hypothesis in terms of its explanatory depth and breadth.
- **E30**: Is the explanatory power of the hypothesis sufficient to warrant belief in its truth?

Figure B.3: Explanatory Power Reasoning Modules.

Coherence Reasoning Modules

- **C1**: How can I determine if a set of beliefs or propositions are coherent with each other?
- **C2**: Identify the logical, explanatory, and probabilistic relationships between the beliefs in this set.
- **C3**: Evaluate whether the addition of new evidence affects the coherence of existing hypotheses.
- **C4**: What measures can be used to quantify the coherence between different beliefs?
- **C5**: Consider whether all beliefs in the system mutually support each other.
- **C6**: How can the coherence of this explanation be improved?
- **C7**: Assess if the hypothesis explains the evidence in a way that fits well with other accepted beliefs.
- **C8**: Use a probabilistic approach to calculate how new evidence changes the coherence of current beliefs.
- **C9**: How can I identify the weakest link in the coherence of a set of beliefs?
- **C10**: Determine whether there are any contradictory beliefs within the system.
- **C11**: How do different evidence pieces contribute to the overall coherence of an explanation?
- **C12**: Analyze whether this set of beliefs forms a coherent network or if it requires restructuring.
- **C13**: What is the role of coherence in justifying a belief or hypothesis?
- **C14**: Evaluate whether coherence alone is sufficient to indicate truth in this context.
- **C15**: How can inconsistencies be resolved to improve coherence?
- **C16**: Determine if there are alternative hypotheses that offer greater coherence with the available evidence.
- **C17**: How does the coherence of this explanation compare with competing explanations?
- **C18**: How can coherence-based reasoning help in making decisions under uncertainty?
- **C19**: How does coherence interact with other epistemic virtues like simplicity and explanatory depth?
- **C20**: Assess if the coherence of a belief set is disrupted by external, new, or conflicting evidence.
- **C21**: Analyze the trade-offs between coherence and other epistemic measures for this context.
- **C22**: How does coherence relate to Bayesian measures of confirmation?
- **C23**: Evaluate whether a lack of coherence suggests the need for hypothesis revision.
- **C24**: Develop a method to rank hypotheses based on their coherence with the evidence.
- **C25**: What are the limitations of using coherence as the sole criterion for belief justification?
- **C26**: How can coherence help in choosing the best explanation among several competing ones?
- **C27**: Determine if coherence supports the inference to the best explanation in this case.
- **C28**: What are the implications of coherence for scientific theory acceptance?
- **C29**: Develop strategies to enhance coherence in collaborative knowledge-building efforts.
- **C30**: Evaluate the impact of coherence on the robustness of a scientific theory or model.

Figure B.4: Coherence Reasoning Modules.

B.2 Epistemic Definitions Used

Explanatory Power Definition

Explanatory Power:

Explanatory power refers to the ability of a hypothesis (H) to account for or make probable a set of observed facts (E) given background knowledge (K). High explanatory power indicates that H makes E significantly more expected than it would be without H.

Figure B.5: Applied definition of Explanatory Power.

Coherence Definition

Coherence:

Coherence is the extent to which a set of propositions, such as hypotheses and observations, fit together in a mutually supportive way within a given body of background knowledge (K). Coherence is enhanced when elements of a hypothesis explain and support each other, forming a consistent and unified whole.

Figure B.6: Applied definition of Coherence.

B.3 Workflow Step Prompts

Select Reasoning Modules

Prompt:

Select several reasoning modules that are crucial to utilize in order to solve the given task.

All reasoning module descriptions:

reasoning_modules

Task:

task_description

Select several modules that are crucial for solving the task above.

Figure B.7: Select Modules Step Prompt.

Adapt Reasoning Modules

Prompt:

Rephrase and specify each reasoning module so that it better helps solving the task.

SELECTED module descriptions:

selected_modules

Task:

task_description

Adapt each reasoning module description to better solve the task.

Figure B.8: Adapt Reasoning Modules Step Prompt.

Structure Reasoning Plan

Prompt:

Operationalize the reasoning modules into a step-by-step reasoning plan in JSON format.

Example Task:

If you follow these instructions, do you return to the starting point? Always face forward. Take 1 step backward. Take 9 steps left. Take 2 steps backward. Take 6 steps forward. Take 4 steps forward. Take 4 steps backward. Take 3 steps right.

Example Reasoning Structure:

```
{
    "Position after instruction 1":
    "Position after instruction 2":
    "Position after instruction n":
    "Is final position the same as starting position":
}
```

Adapted module description:

adapted_modules

Task:

task_description

Implement a reasoning structure for solvers to follow step-by-step and arrive at the correct answer.

Figure B.9: Structure Reasoning Step Prompt.

Apply Reasoning Plan

Prompt:

Follow the step-by-step reasoning plan in JSON to correctly solve the task. Fill in the values following the keys by reasoning specifically about the task.

Reasoning Structure:

reasoning_structure

Task:

task_description

Figure B.10: Apply Reasoning Step Prompt.

Update Task Description with Context

Prompt:

Given the following context and task description, update the task description to incorporate the relevant aspects of the context.

Context:

context

Original Task Description:

task_description

Updated Task Description:

Figure B.11: Contextualise Task Step Prompt.

Appendix C

Self-RAG Materials

C.1 Workflow Step Prompts

Document Relevance Check

Prompt:

You are an expert grader assessing the relevance of retrieved documents to a user question. Your goal is to identify documents that contain information directly related to answering the question.

Consider both explicit keyword matches and implicit semantic relevance.

Task:

Provide a binary score ’yes’ or ’no’ and explain your reasoning.

Figure C.1: Document Relevance Check Prompt.

Answer Generation

Prompt:

You are an assistant tasked with answering questions based on the provided context. Ensure your response is directly relevant to the question and grounded in the given information.

If you don’t know the answer, just say that you don’t know. Use five sentences maximum and keep the answer concise.

Figure C.2: Answer Generation Prompt.

Query Transformation

Prompt:

You are an expert at reformulating questions to improve information retrieval. Analyze the input question and create a version that is more likely to match relevant documents.

Consider expanding abbreviations, including synonyms, and clarifying ambiguous terms.

Figure C.3: Query Transformation Prompt.

Hallucination Grader

Prompt:

You are an expert fact-checker assessing whether an AI-generated answer is grounded in the provided documents. Carefully compare the answer to the information in the documents.

Task:

Provide a binary score ’yes’ or ’no’ and explain your reasoning.

Figure C.4: Hallucination Grader Prompt.

Answer Relevance Grader

Prompt:

You are an expert evaluator assessing whether an answer fully addresses and resolves a given question. Consider completeness, relevance, and clarity of the answer.

Task:

Provide a binary score ’yes’ or ’no’ and explain your reasoning.

Figure C.5: Answer Relevance Grader Prompt.

Appendix D

Additional Figures

D.1 Supporting Density Plots

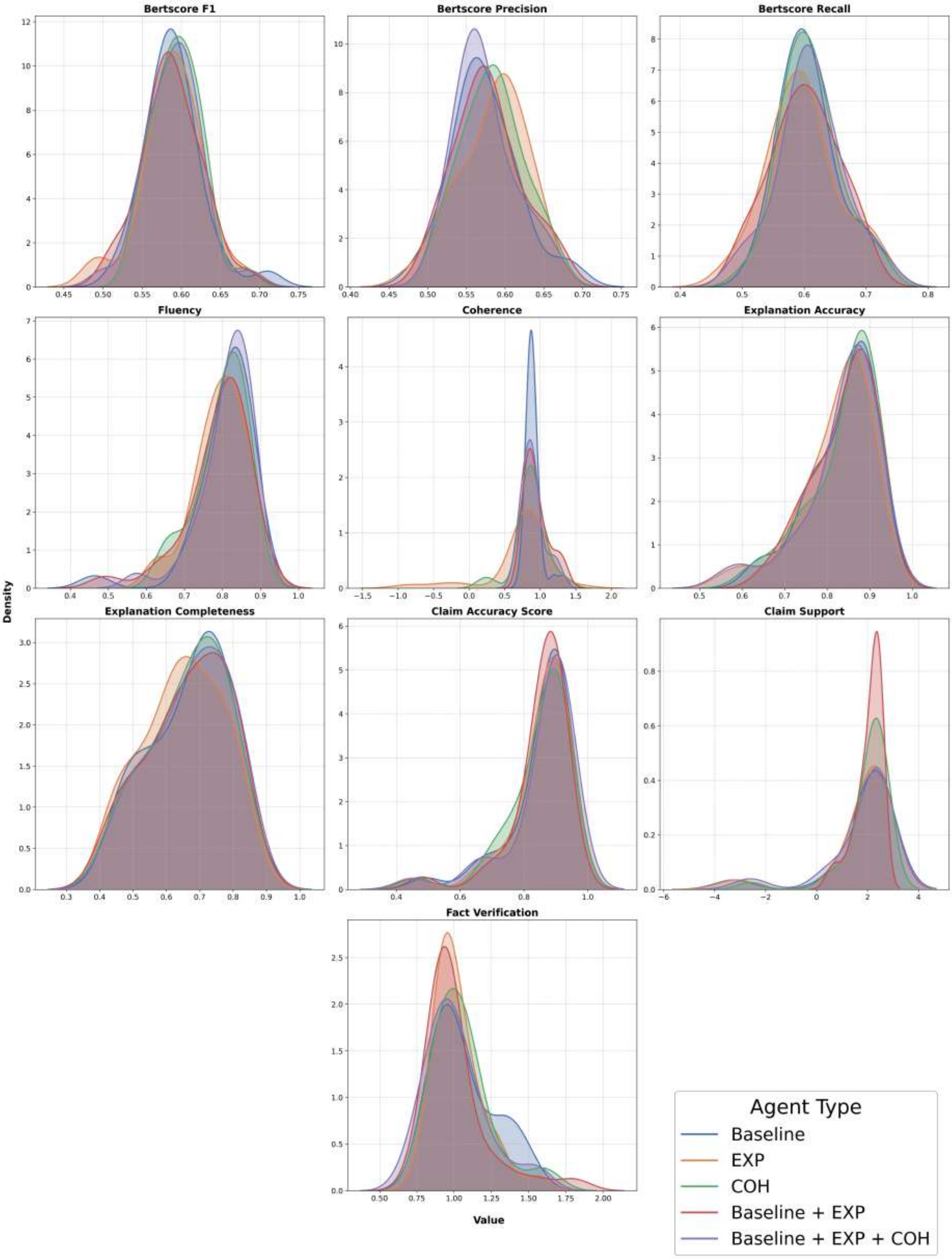

Figure D.1: Full Evidence Densities (CIViC) - GPT-4o

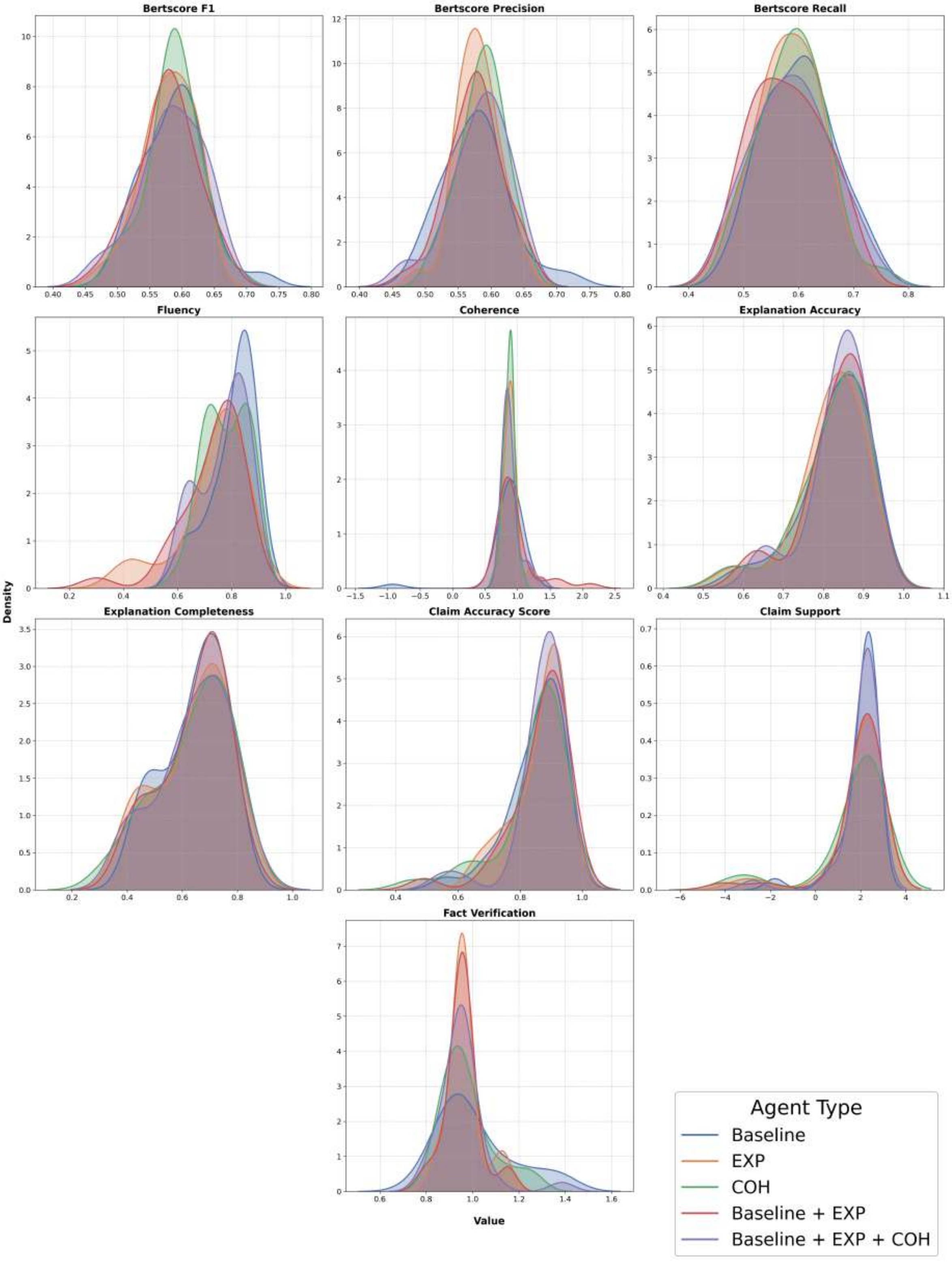

Figure D.2: Full Evidence Densities (CIViC) - GPT-4o Mini

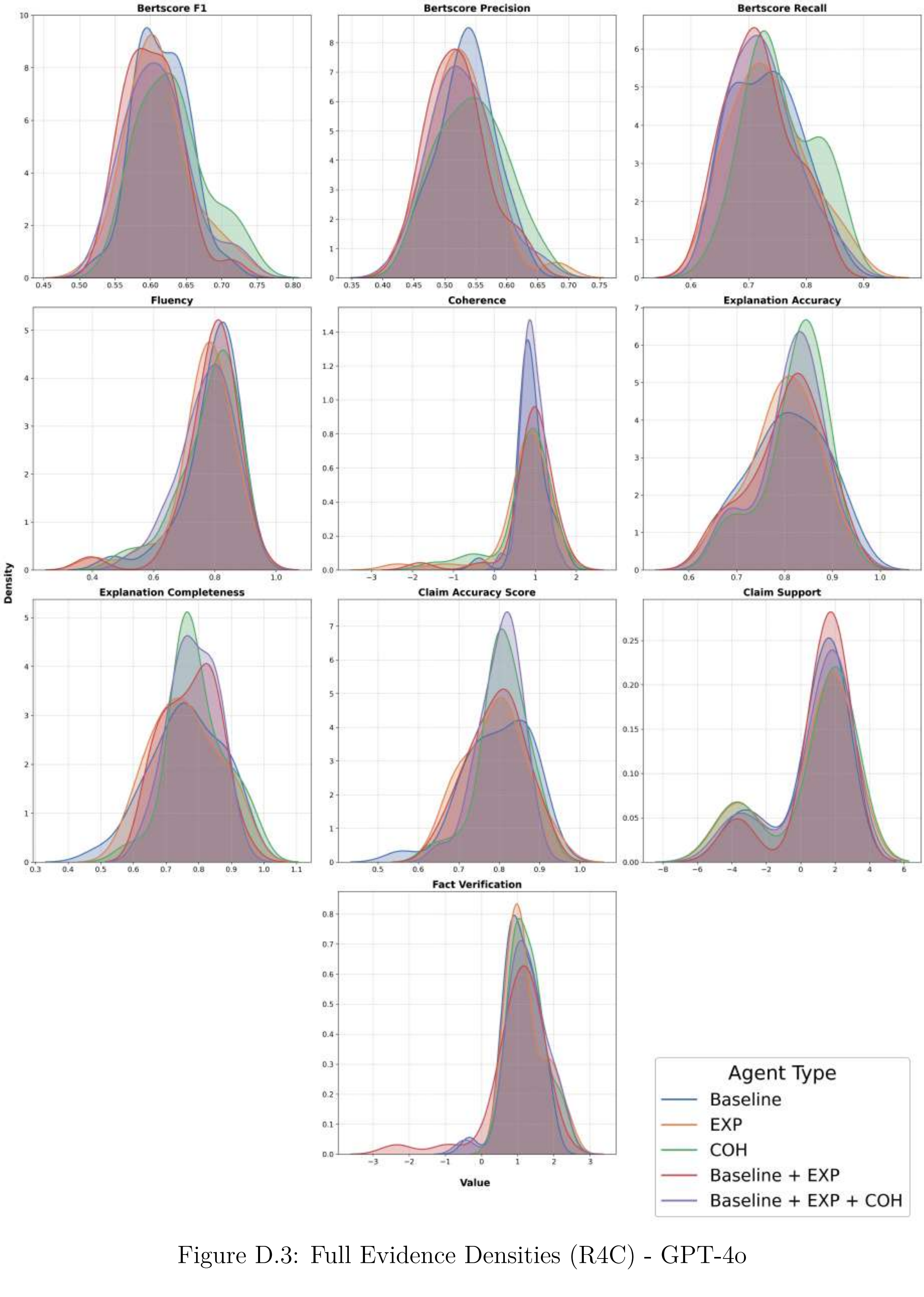

Figure D.3: Full Evidence Densities (R4C) - GPT-4o

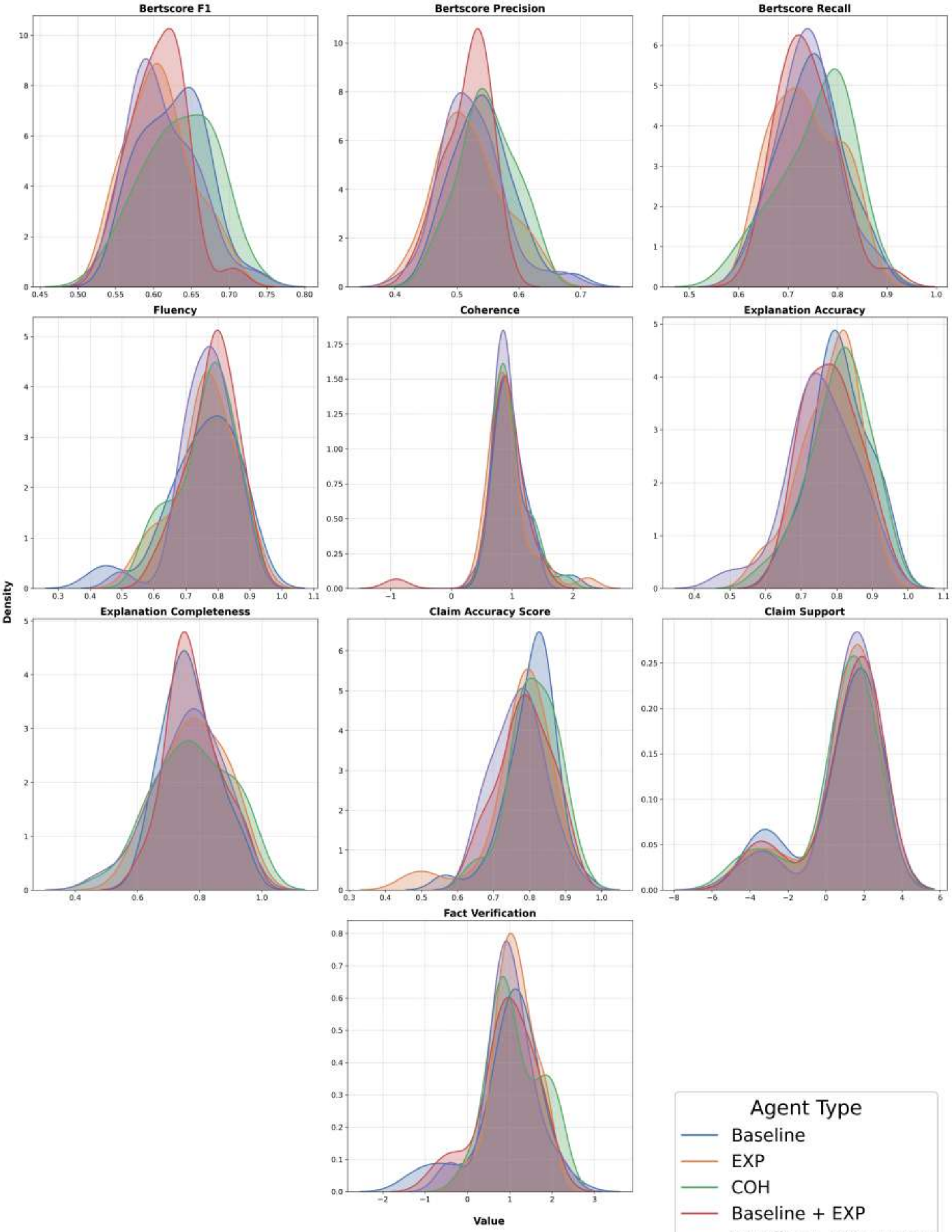

Figure D.4: Full Evidence Densities (R4C) - GPT-4o Mini

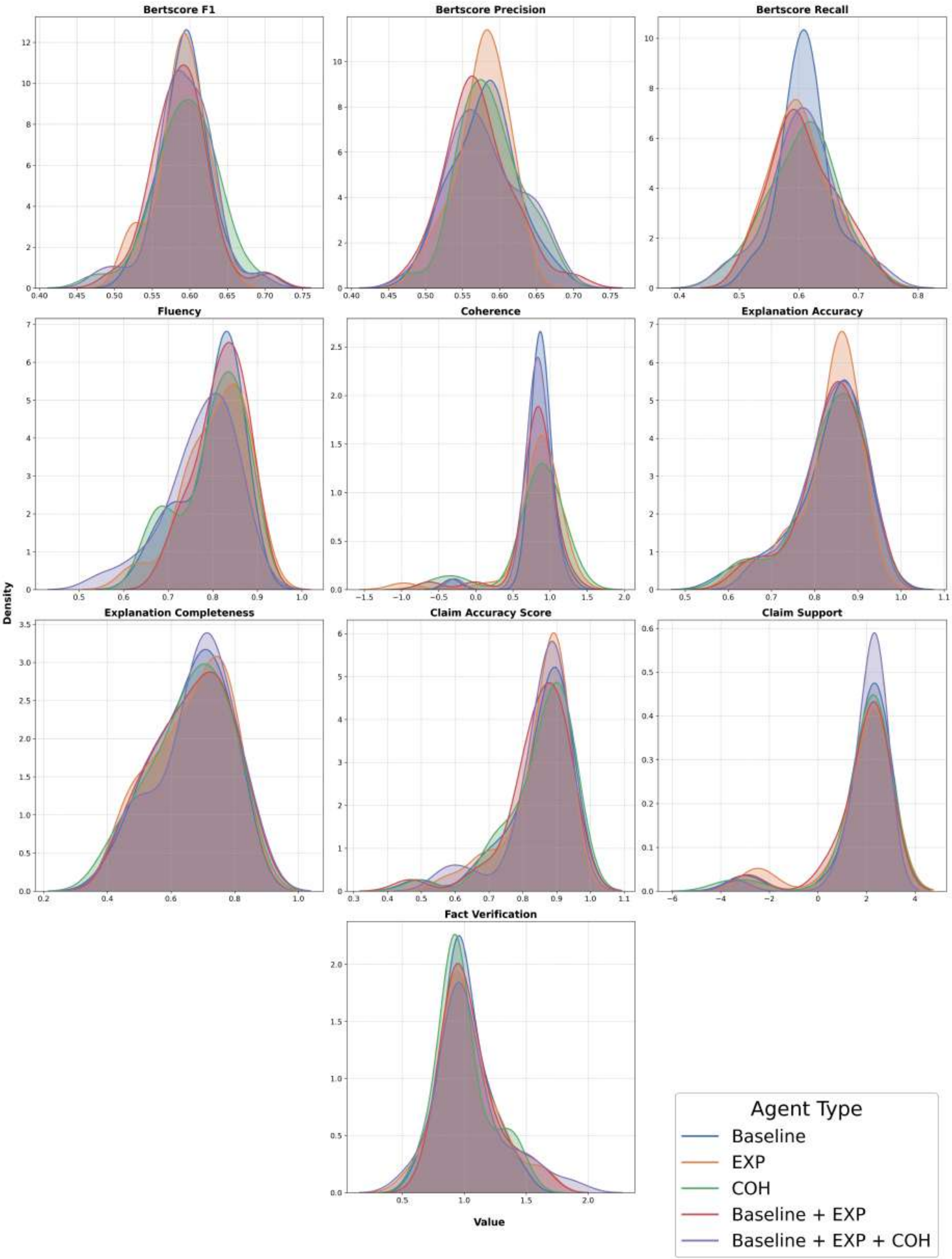

Figure D.5: Missing Evidence Densities (CIViC) - GPT-4o

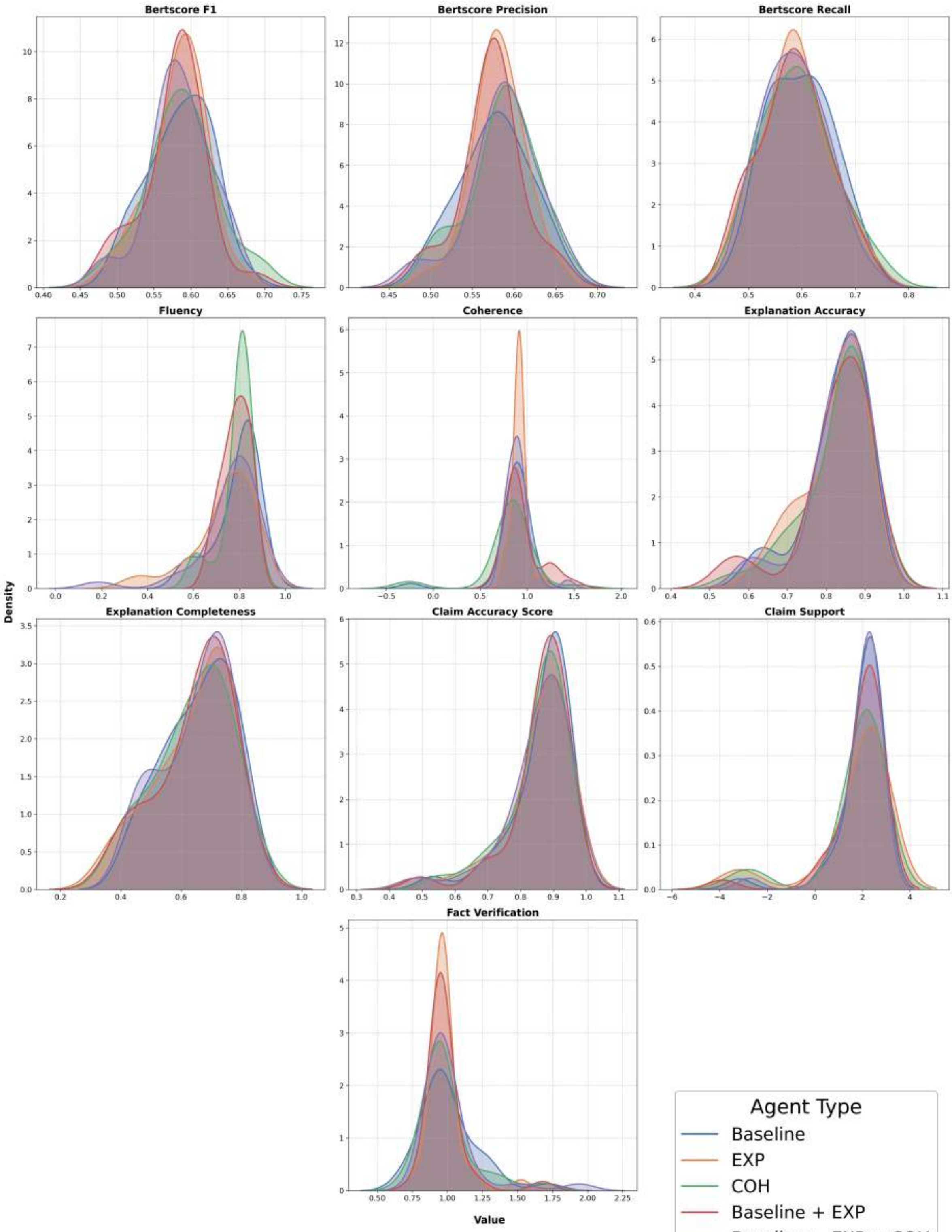

Figure D.6: Missing Evidence Densities (CIViC) - GPT-4o Mini

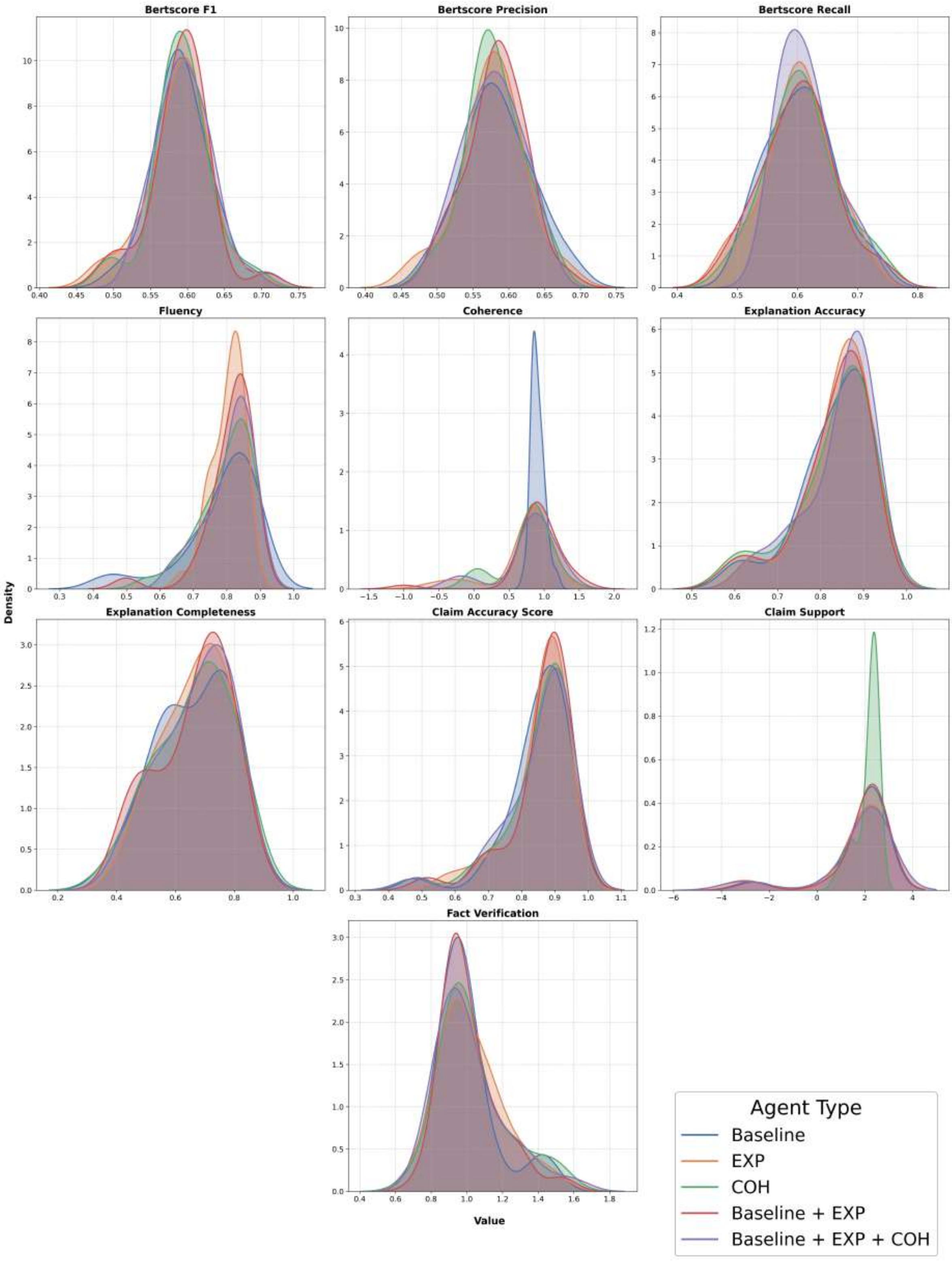

Figure D.7: Wrong Evidence Densities (CIViC) - GPT-4o

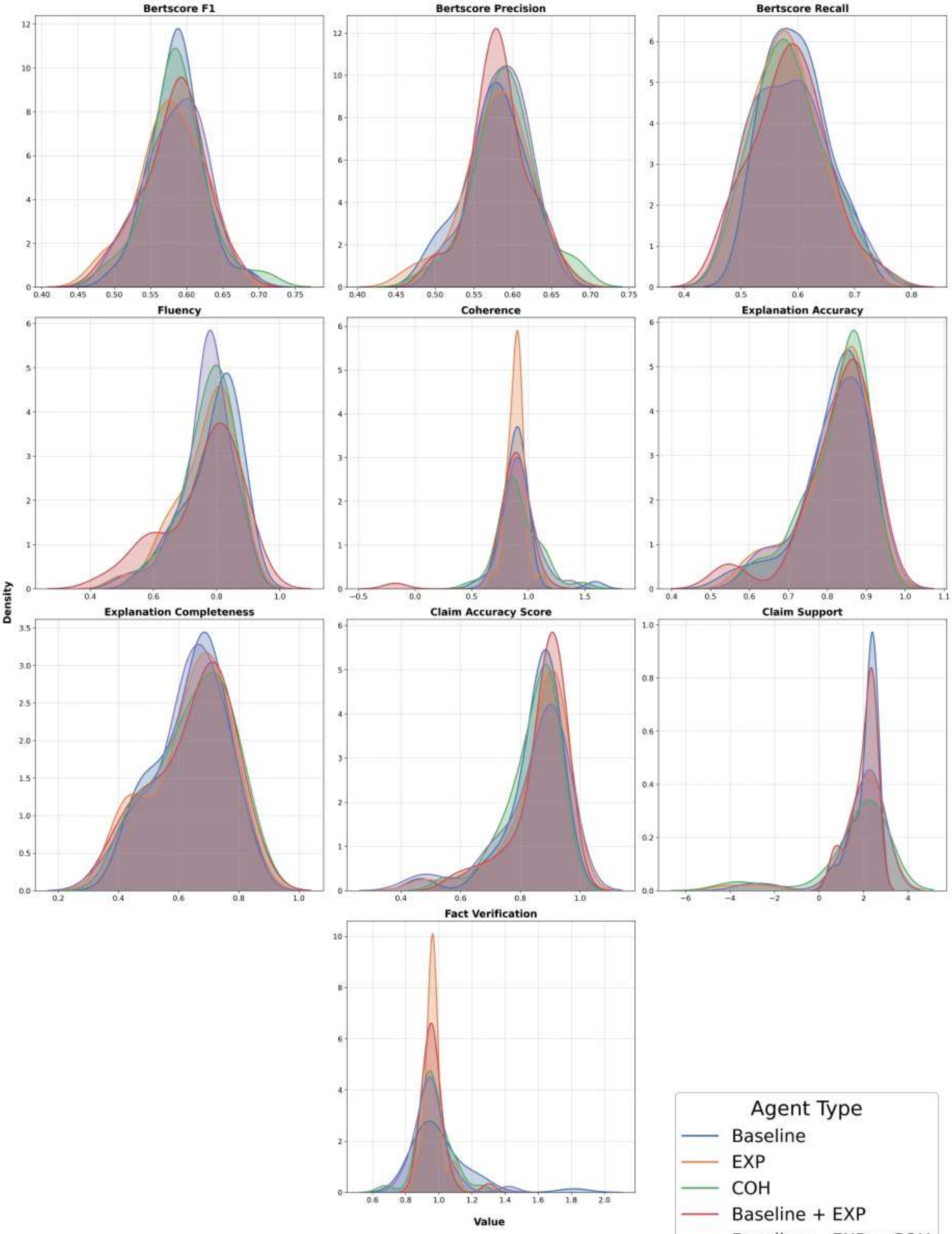

Figure D.8: Wrong Evidence Densities (CIViC) - GPT-4o Mini

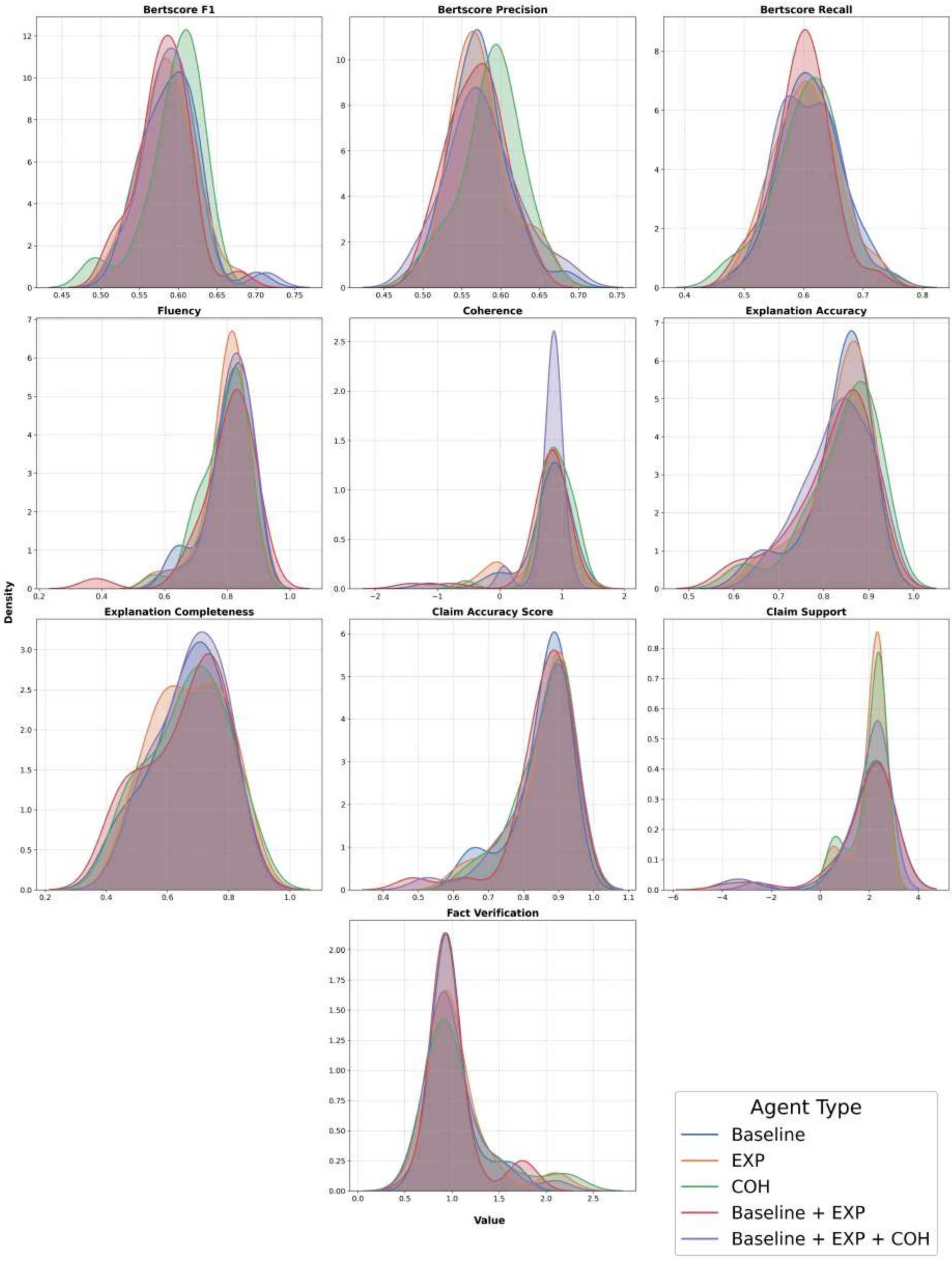

Figure D.9: Mixed Evidence Densities (CIViC) - GPT-4o

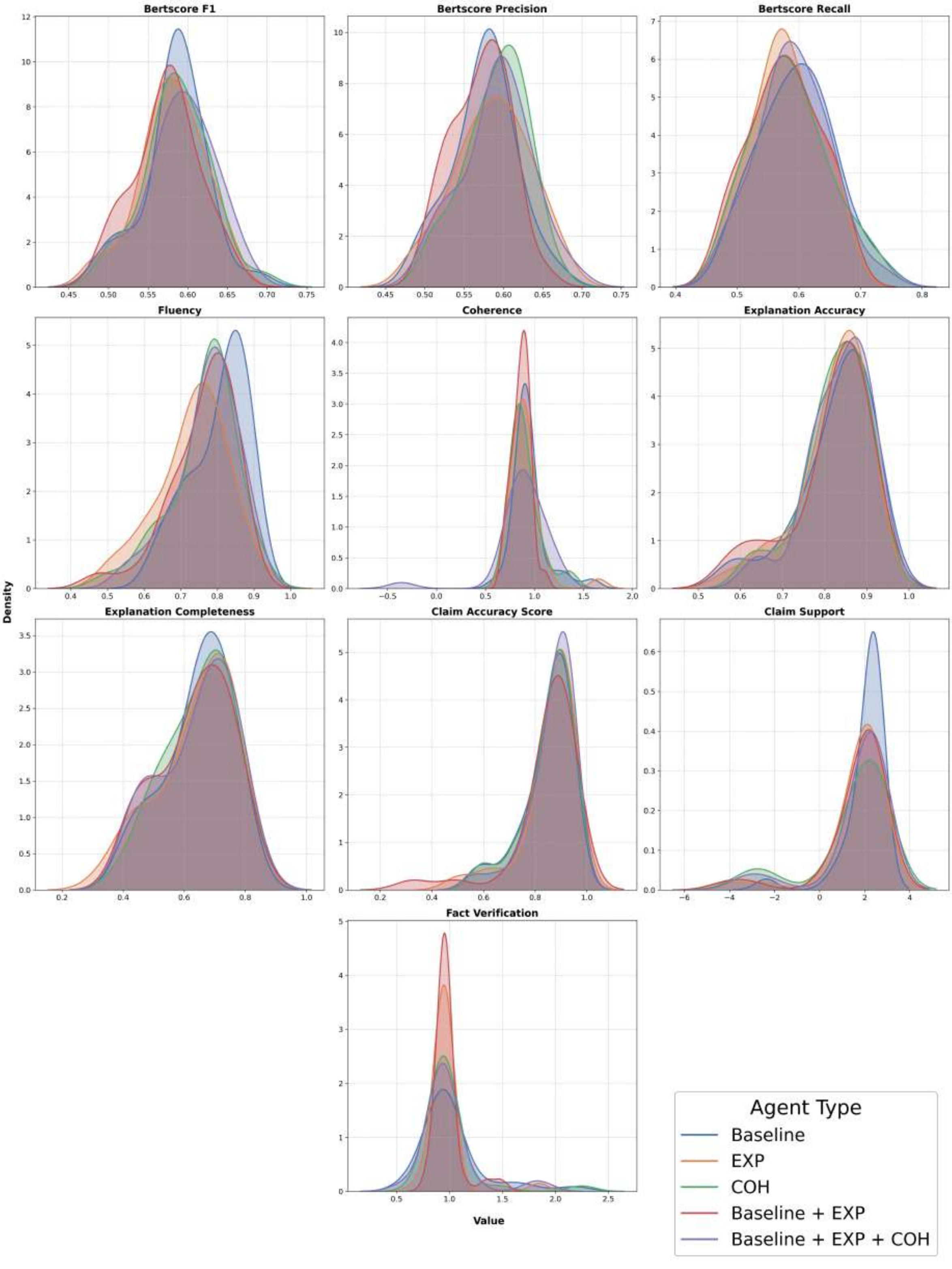

Figure D.10: Mixed Evidence Densities (CIViC) - GPT-4o Mini

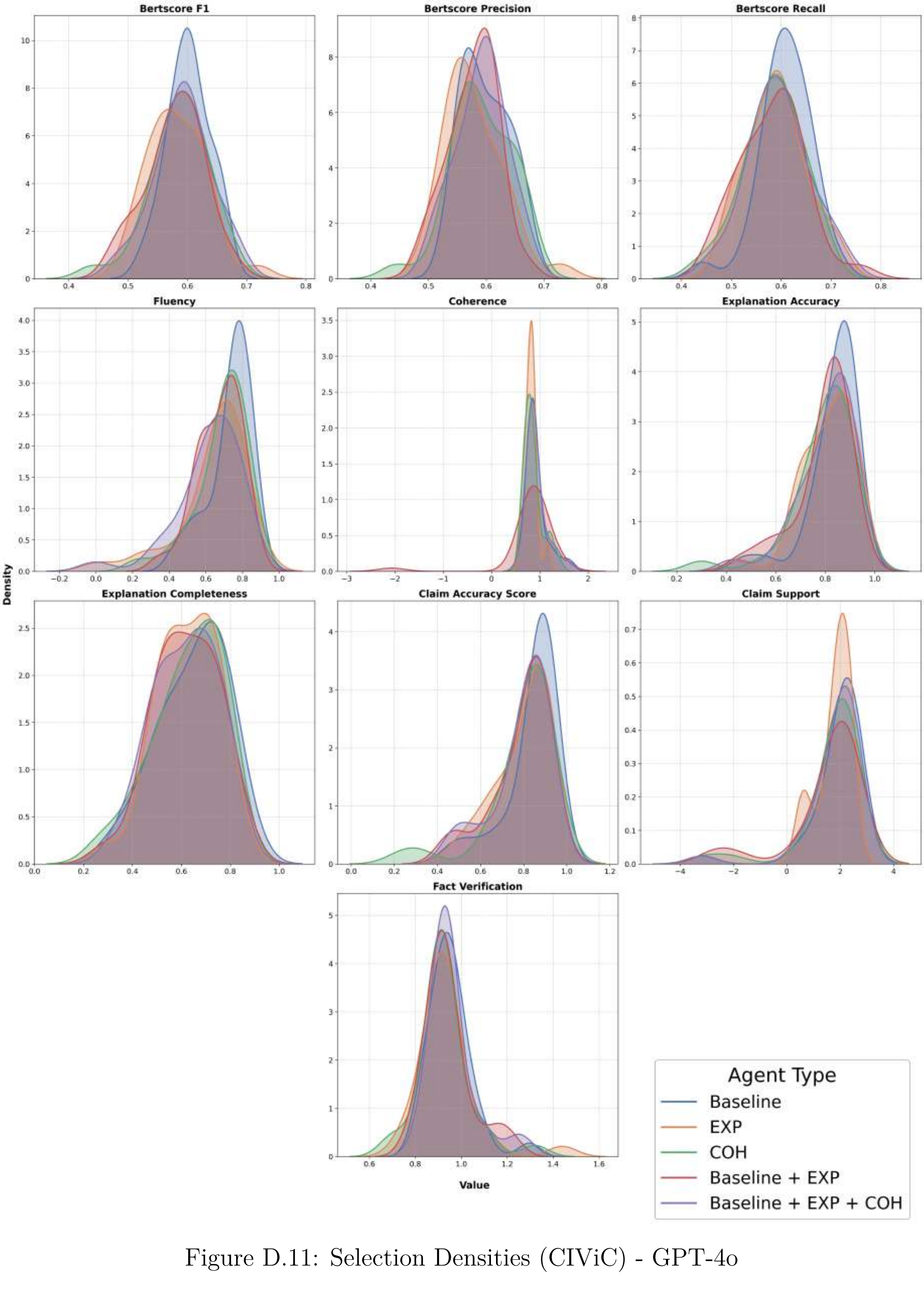

Figure D.11: Selection Densities (CIViC) - GPT-4o

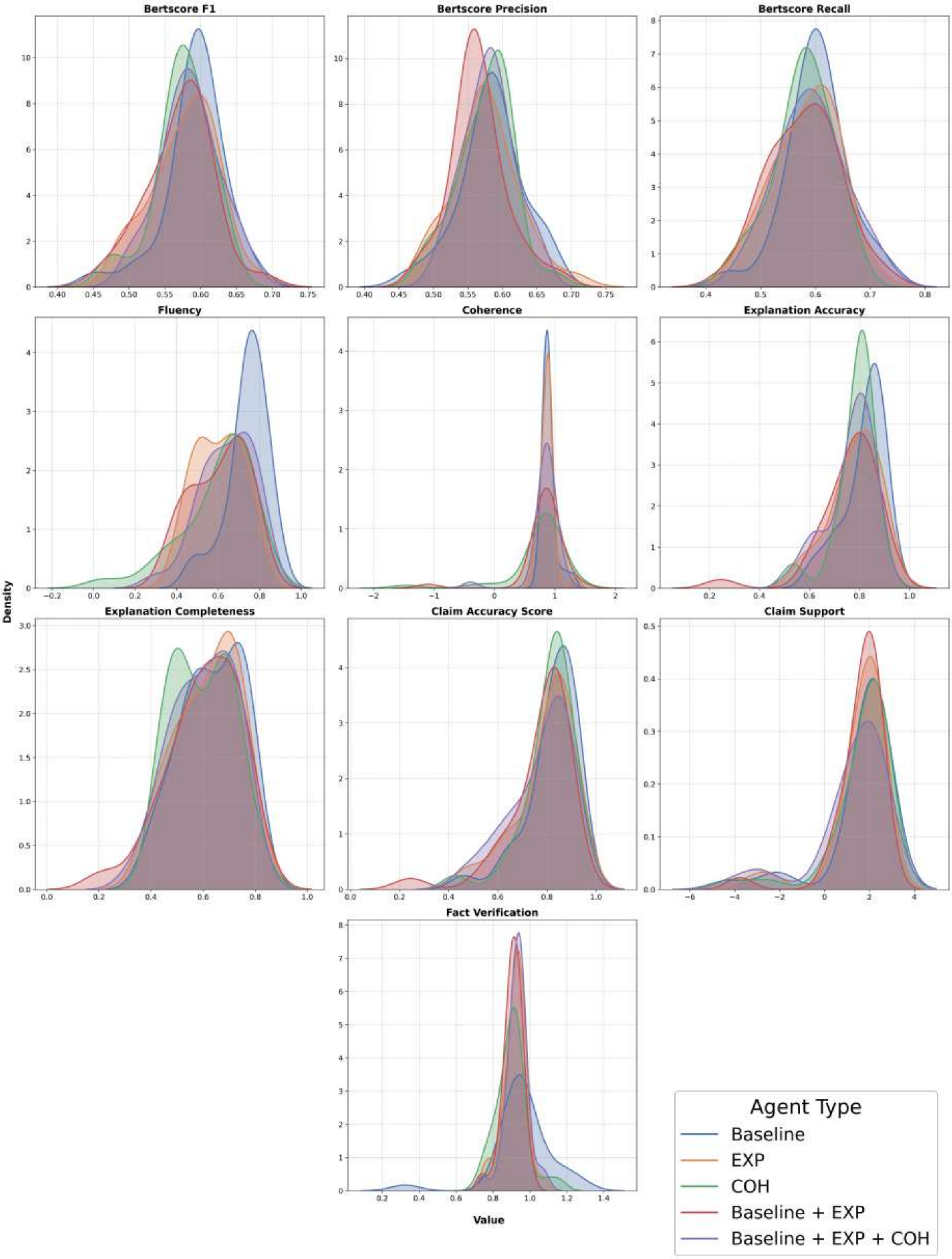

Figure D.12: Selection Densities (CIViC) - GPT-4o Mini

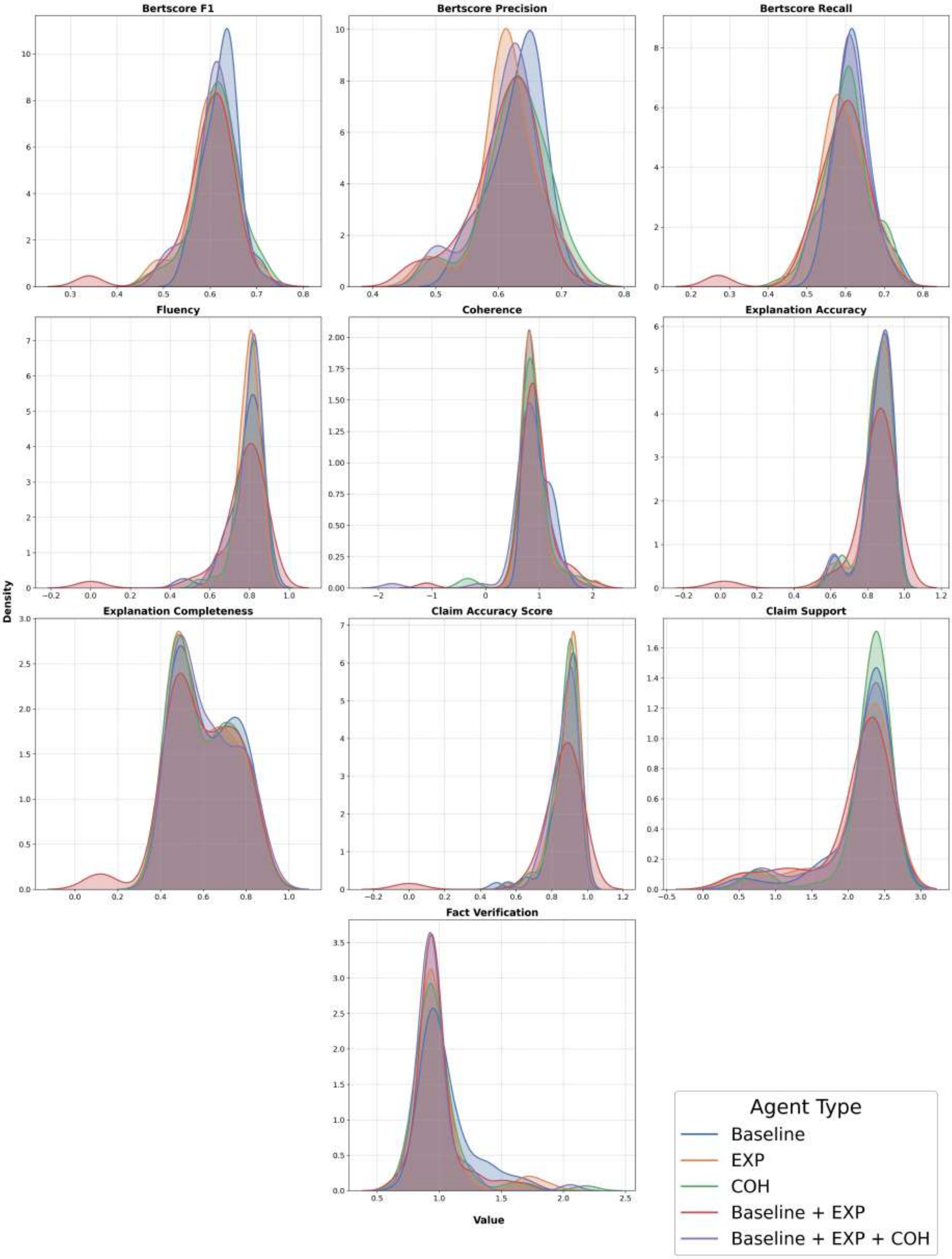

Figure D.13: Assignment Densities (CIViC) - GPT-4o

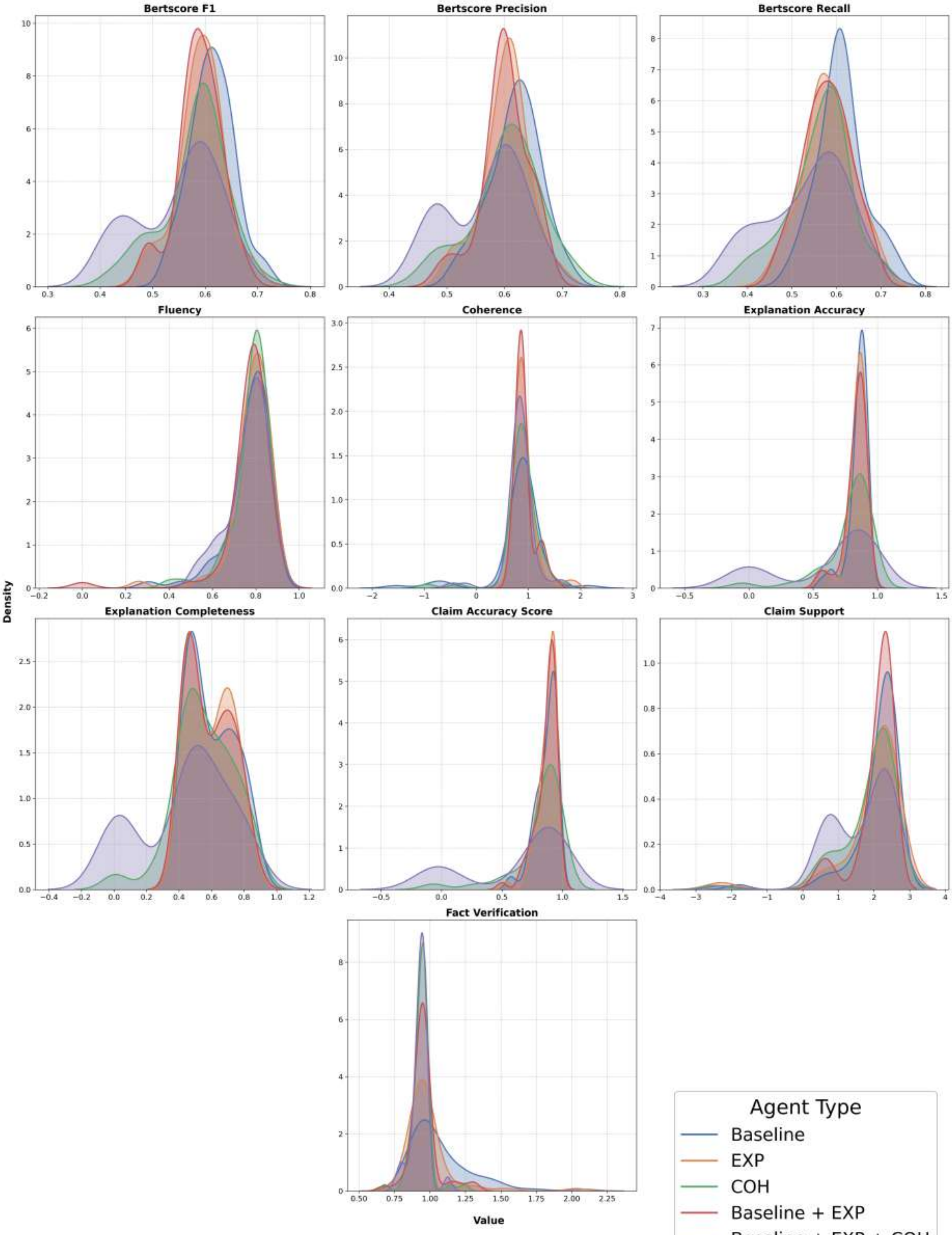

Figure D.14: Assignment Densities (CIViC) - GPT-4o Mini

Appendix E

Other Supplementary Materials

E.1 Additional Test Examples

Example Missing Evidence Entry

Claim:

HER2 amplification predicts sensitivity to Trastuzumab.

Explanation:

HER2 amplification defines a (...) standard of care for HER2-positive breast cancer patients.

Evidence:

- **EID 529:** A randomized clinical trial of 186 patients (...) trastuzumab in addition to chemotherapy.
- **EID 528:** A randomized clinical trial of 469 patients (...) trastuzumab in addition to chemotherapy.
- **...**

Missing Evidence:

- **EID 1122:** HERA was a Phase III trial (...) standard of care.
- **...**

Context:

- **Molecular Profile:** ERBB2 Amplification
- **Molecular Profile Summary:** Her2 (ERBB2) amplifications are seen (...) targeted in neoadjuvant breast cancer treatment.
- **Disease:** Her2-receptor Positive Breast Cancer
- **Therapies:** Trastuzumab
- **Phenotypes:** None specified

Figure E.1: Example of Missing Evidence Textual Data.

Note that the *Missing Evidence* entity is hidden from the model.

Example Wrong Evidence Entry

Claim:

HER2 amplification predicts sensitivity to Trastuzumab.

Explanation:

HER2 amplification defines a (...) standard of care for HER2-positive breast cancer patients.

Evidence:

- **EID 1122:** HERA was a Phase III trial (...) standard of care.
- **EID 528:** A randomized clinical trial of 469 patients (...) trastuzumab in addition to chemotherapy.
- **...**

Wrong Evidence:

- **EID 11774:** Copy-number analysis was performed on 44 (...) MYB family in the biology of low-grade gliomas.
- **...**

Context:

- **Molecular Profile:** ERBB2 Amplification
- **Molecular Profile Summary:** Her2 (ERBB2) amplifications are seen (...) targeted in neoadjuvant breast cancer treatment.
- **Disease:** Her2-receptor Positive Breast Cancer
- **Therapies:** Trastuzumab
- **Phenotypes:** None specified

Figure E.2: Example of Wrong Evidence Textual Data.

Note that the *Wrong Evidence* entity is seen as normal evidence by the model.

Example Mixed Evidence Entry

Claim:

HER2 amplification predicts sensitivity to Trastuzumab.

Explanation:

HER2 amplification defines a (...) standard of care for HER2-positive breast cancer patients.

Evidence:

- **EID 1122:** HERA was a Phase III trial (...) standard of care.
- **EID 529:** A randomized clinical trial of 186 patients (...) trastuzumab in addition to chemotherapy.
- ...

Missing Evidence:

- **EID 528:** A randomized clinical trial of 469 patients (...) trastuzumab in addition to chemotherapy.
- ...

Wrong Evidence:

- **EID 1643:** A DNAJB1:PRKACA fusion transcript was detected (...) in all primary and metastatic samples.
- ...

Context:

- **Molecular Profile:** ERBB2 Amplification
- **Molecular Profile Summary:** Her2 (ERBB2) amplifications are seen (...) targeted in neoadjuvant breast cancer treatment.
- **Disease:** Her2-receptor Positive Breast Cancer
- **Therapies:** Trastuzumab
- **Phenotypes:** None specified

Figure E.3: Example of Mixed Evidence Textual Data. The same rules apply as above.

Example Selection Test Entry

Claim A:
PAX5 p.P80R as essential diagnostic criteria of the provisional B lymphoblastic leukaemia with PAX5 p.P80R subtype.

Explanation A:
PAX5 missense variant p.Pro80Arg defines a genetic subtype of B-lymphoblastic leukemia (...) recognized in the WHO and ICC classification.

Evidence:

- **EID 11519:** RNA-seq transcriptome analysis (...) identified a unique PAX5 mutation, p.P80R, in all cases.
- **EID 7290:** In a large-scale international study (...) found P80R mutation in the PAX5 gene.
- **...**

Context A:

- **Molecular Profile:** PAX5 P80R
- **Disease:** B-lymphoblastic Leukemia/lymphoma With PAX5 P80R
- **Therapies:** None specified
- **Phenotypes:** None specified

Claim B:
ETV6::NTRK3-positive infantile fibrosarcoma tumors are sensitive to larotrectinib.

Context B:

- **Molecular Profile:** NTRK3 ETV6::NTRK3
- **Disease:** Congenital Fibrosarcoma
- **Therapies:** Larotrectinib
- **Phenotypes:** Pediatric onset

Figure E.4: Example of Selection Test Textual Data.

Example Assignment Test Entry

Claim A:

BCOR ITD is a desirable diagnostic criteria for clear cell sarcoma of kidney.

Explanation A:

Clear cell sarcoma of the kidney (...) associated with BCOR overexpression.

Evidence:

- **EID 11424:** Whole transcriptome sequencing (...) identified BCOR ITD and overexpression of BCOR in all samples.
- **EID 11423:** High BCOR expression (...) associated with an internal tandem duplication (ITD) of BCOR.
- **EID 7496:** In this phase 1 dose-escalation study (...) treatment with larotrectinib reduced tumour burden.
- **EID 6099:** Four out of six patients (...) had a response with larotrectinib.
- **...**

Context A:

- **Molecular Profile:** BCOR ITD
- **Disease:** Kidney Clear Cell Sarcoma
- **Phenotypes:** Pediatric onset, Childhood onset, Infantile onset

Claim B:

ETV6-NTRK3–positive B-cell lymphoblastic leukemia patients can be sensitive to larotrectinib.

Explanation B:

An ETV6-NTRK3 gene fusion (...) may respond to larotrectinib.

Context B:

- **Molecular Profile:** NTRK3 ETV6::NTRK3
- **Disease:** B-lymphoblastic Leukemia/lymphoma
- **Therapies:** Larotrectinib

Figure E.5: Example of Assignment Test Textual Data.

E.2 Defined Task Prompts

Explanation Task

Generate a 3-5 sentence long explanation of the claim using relevant evidence. Critically address how evidence supports or contradicts the claim. Be cautious of incomplete or incorrect evidence. Incorporate additional information if provided and useful. Format output as:

- Generated Explanation: [explanation]

Figure E.6: Explanation Task Definition.

Selection Task

Select appropriate claim (A or B) based on provided evidence. Generate a 3-5 sentence long explanation of chosen claim using evidence. Critically address how evidence supports or contradicts the claim. Be cautious of incomplete or incorrect evidence. Incorporate additional context if provided and useful. Format output as:

- Selected Claim: [A or B]
- Generated Explanation: [explanation]

Figure E.7: Selection Task Definition.

Assignment Task

Generate 3-5 sentence long explanations for claims A and B using relevant evidence. Identify supporting or contradicting evidence for each claim. Be cautious of unrelated, incomplete, or incorrect evidence. Incorporate additional context if provided and useful. Format output as:

- Claim A:
  - Generated Explanation: [explanation for A]
- Claim B:
  - Generated Explanation: [explanation for B]

Figure E.8: Assignment Task Definition.

E.3 R4C Data Extraction-Generation Prompts

Generate Claim

Convert the following question and answer into a single and concise sentence claim. Do NOT introduce any new information, just paraphrase in simple language. Do NOT return ANY other additional text, just the claim:
Question: {question}
Answer: {answer}
Claim:

Figure E.9: R4C Claim Generation Prompt.

Generate Explanation

Using the following evidence, create a short and concise explanation (1-3 sentences) for the claim. Use simple and plain language. Do NOT introduce any new information, just combine what is given. Keep paraphrasing to a minimum. Do NOT return ANY other additional text, just the explanation:
Evidence: {evidence}
Claim: {claim}
Explanation:

Figure E.10: R4C Explanation Generation Prompt.

Remove Duplicate Information

Remove any informationally duplicate entries from the following list, including paraphrases or similar information. Do NOT change the content under any circumstances, only remove. Return only the unique informational content without any leading dashes or bullet points. Do NOT return ANY other additional text, just the evidence:
{evidence}
Unique information:

Figure E.11: R4C Remove Duplicate Information Prompt.

E.4 RAG Task Prompt

Question Generation Task

Generate {n} useful questions to solve this task: {task}

Focus on missing information, not solution methods. Ensure diversity of information and avoid repetition.

Format:

Question 1: ...

Question 2: ...

Continue until {n} questions are generated.

Figure E.12: RAG Evidence Retrieval Prompt.

E.5 Example Sequential System

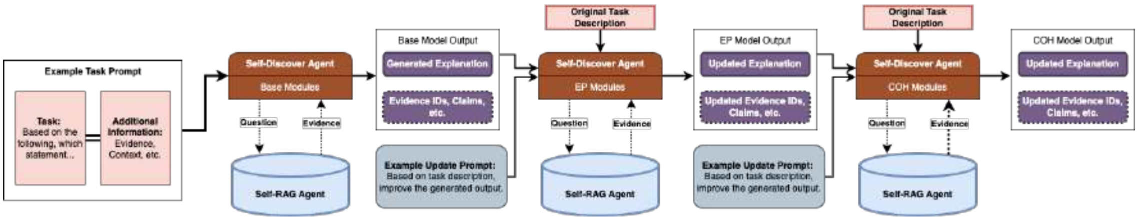
Example Task Prompt
Task: Based on the following, which statement...
Additional Information: Evidence, Context, etc.
Self-Discover Agent
Base Modules
Question
Evidence
Self-RAG Agent
Base Model Output
Generated Explanation
Evidence IDs, Claims, etc.
Example Update Prompt: Based on task description, improve the generated output.
Original Task Description
Self-Discover Agent
EP Modules
Question
Evidence
Self-RAG Agent
EP Model Output
Updated Explanation
Updated Evidence IDs, Claims, etc.
Example Update Prompt: Based on task description, improve the generated output.
Original Task Description
Self-Discover Agent
COH Modules
Question
Evidence
Self-RAG Agent
COH Model Output
Updated Explanation
Updated Evidence IDs, Claims, etc.

Figure E.13: Example Sequential Agent System.

E.6 CIVIC Schema

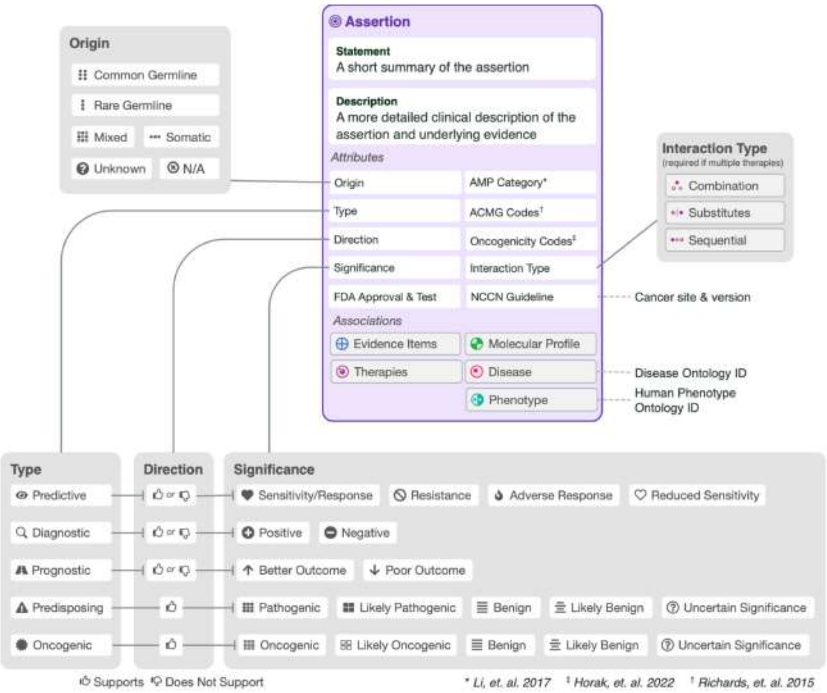

Figure E.14: Assertion Attributes and Associations (Good et al., 2014).

JSON Field	CIViC Field
claim	Summary
explanation	Description
evidence	
evidence_id	EID
description	EID:Statement
context	
Molecular Profile	Molecular Profile Name
Molecular Profile Summary	Molecular Profile Name:Description
Disease	disease
Therapies	therapies
Phenotypes	phenotypes

Table E.15: Mapping of JSON Fields to CIViC Fields.