*informatics*

MDPI

*Article*

# Multi-Model Dialectical Evaluation of LLM Reasoning Chains: A Structured Framework with Dual Scoring Agents

**Catalin Anghel [1,*], Andreea Alexandra Anghel [2], Emilia Pecheanu [1,*], Ioan Susnea [1], Adina Cocu [1] and Adrian Istrate [1]**

[1] Department of Computer Science and Information Technology, "Dunărea de Jos" University of Galati, Științei St. 2, 800146 Galati, Romania; ioan.susnea@ugal.ro (I.S.); adina.cocu@ugal.ro (A.C.); adrian.istrate@ugal.ro (A.I.)

[2] Faculty of Automation, Computer Science, Electrical and Electronic Engineering, "Dunărea de Jos" University of Galati, 800008 Galati, Romania; aa481@student.ugal.ro

* Correspondence: catalin.anghel@ugal.ro (C.A.); emilia.pecheanu@ugal.ro (E.P.); Tel.: +40-745802834 (C.A.); +40-722686122 (E.P.)

**Abstract**

(1) Background and objectives: Large language models (LLMs) such as GPT, Mistral, and LLaMA exhibit strong capabilities in text generation, yet assessing the quality of their reasoning—particularly in open-ended and argumentative contexts—remains a persistent challenge. This study introduces Dialectical Agent, an internally developed modular framework designed to evaluate reasoning through a structured three-stage process: opinion, counterargument, and synthesis. The framework enables transparent and comparative analysis of how different LLMs handle dialectical reasoning. (2) Methods: Each stage is executed by a single model, and final syntheses are scored via two independent LLM evaluators (LLaMA 3.1 and GPT-4o) based on a rubric with four dimensions: clarity, coherence, originality, and dialecticality. In parallel, a rule-based semantic analyzer detects rhetorical anomalies and ethical values. All outputs and metadata are stored in a Neo4j graph database for structured exploration. (3) Results: The system was applied to four open-weight models (Gemma 7B, Mistral 7B, Dolphin-Mistral, Zephyr 7B) across ten open-ended prompts on ethical, political, and technological topics. The results show consistent stylistic and semantic variation across models, with moderate inter-rater agreement. Semantic diagnostics revealed differences in value expression and rhetorical flaws not captured by rubric scores. (4) Originality: The framework is, to our knowledge, the first to integrate multi-stage reasoning, rubric-based and semantic evaluation, and graph-based storage into a single system. It enables replicable, interpretable, and multidimensional assessment of generative reasoning—supporting researchers, developers, and educators working with LLMs in high-stakes contexts.

**Keywords:** large language models (LLMs); dialectical reasoning; multi-agent evaluation; generative argumentation; prompt-based evaluation

check for updates

Academic Editor: Sotiris Kotsiantis

Received: 30 May 2025
Revised: 21 July 2025
Accepted: 23 July 2025
Published: 1 August 2025

**Citation:** Anghel, C.; Anghel, A.A.; Pecheanu, E.; Susnea, I.; Cocu, A.; Istrate, A. Multi-Model Dialectical Evaluation of LLM Reasoning Chains: A Structured Framework with Dual Scoring Agents. *Informatics* **2025**, *12*, 76. https://doi.org/10.3390/informatics12030076

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

### *1.1. Background and Objectives of the Study*

Recent progress in large language models (LLMs) such as GPT, Mistral, and LLaMA has significantly advanced natural language generation, enabling systems to produce coherent and contextually appropriate text for a wide range of tasks, including summarization,

dialogue generation, and question answering. As these models become increasingly embedded into high-stakes applications—such as education, governance, and decision support—the need for interpretable, structured, and evaluable reasoning grows correspondingly.

While many LLMs demonstrate impressive fluency, evaluating the depth, coherence, and argumentative structure of their output remains a fundamental challenge. Traditional automatic metrics such as BLEU [1], ROUGE [2], or exact match scores on datasets like SQuAD [3] are well-suited for factual and extractive tasks, but insufficient when applied to open-ended reasoning or discourse generation. These metrics emphasize surface similarity rather than logical structure, conceptual nuance, or dialectical integration.

Several prompting-based strategies have been proposed to foster more deliberative behavior in large language models. Chain-of-thought prompting [4] encourages the generation of intermediate reasoning steps, while iterative refinement approaches [5] enable models to critique and revise their initial outputs. Agent-based frameworks—such as Reflexion [6], which incorporates verbal self-feedback, and Generative Agents [7], which simulate human-like interactions through planning and memory—extend these mechanisms into multi-turn, context-aware settings. Nevertheless, the majority of these approaches rely on single-model architectures and lack structured, comparative evaluation protocols.

Dialectical reasoning—defined here as the capacity to synthesize divergent viewpoint perspectives into a coherent conclusion—remains underexplored. Multi-step reflection [4], internal critique [5], and simulated debate [6,7] offer promising heuristics, but rarely incorporate independent evaluation, role separation, or long-term representation of reasoning trajectories. Consequently, it remains unclear how different models behave when assigned dialectical roles under identical constraints [8].

In this context, the general objective of the present study is to develop and validate the Dialectical Agent framework—a modular system for comparative reasoning and hybrid evaluation of LLM outputs. The framework treats reasoning as a structured, three-stage process comprising the following: (1) stating an initial opinion, (2) generating counterarguments, and (3) producing a final synthesis that reconciles both perspectives. Each stage is executed via a distinct LLM, allowing for controlled role assignment and comparative experimentation.

The term "dialectical" is used in this study in its philosophical and argumentative sense, referring to structured reasoning through opposing viewpoints (opinion, counterargument, synthesis). It is unrelated to linguistic dialects or multimodal data. The entire framework operates in a monolingual, text-based setting.

Evaluation is conducted through two complementary mechanisms. First, a rubric-based scoring system assesses the final synthesis along four axes: clarity, coherence, originality, and dialecticality. Second, a semantic analyzer detects the presence of ethical values (e.g., empathy, fairness) and rhetorical anomalies (e.g., bias, superficiality). These signals contribute to a final qualitative label that summarizes each reasoning chain's strengths and weaknesses. All outputs are stored in a Neo4j graph database, supporting structured querying and visual inspection.

To guide system design and analysis, we formulated the following research questions:

RQ1. How can dialectical reasoning processes be operationalized through multi-stage prompting in large language models?

RQ2. How do LLMs handle successive dialectical roles (opinion, counterargument, synthesis) when operating within a structured reasoning pipeline?

RQ3. Can the combination of rubric-based scoring and semantic analysis provide a more comprehensive evaluation of generative reasoning than numeric metrics alone?

RQ4. How can reasoning outputs be structured and stored to enable long-term analysis and cross-model comparison?

All components of the Dialectical Agent framework were internally developed, including the reasoning pipeline, evaluator interface, and semantic analyzer. This design enables full control over role assignment, evaluation procedures, and structured output representation

### *1.2. Research Gap and Novelty*

Although significant progress has been made in encouraging deliberative behavior in LLMs [5,8–10], prior work tends to focus on single-model pipelines or internal critique loops. Most systems lack external evaluation mechanisms, modular generation pipelines, or persistent representations of reasoning structures.

This study fills these gaps through the following:

i. Introducing a modular, dialectically structured pipeline for multi-model reasoning;
ii. Combining rubric-based scoring with rule-based semantic analysis of ethical values and argumentative flaws;
iii. Supporting flexible model assignment and comparative evaluation of reasoning roles;
iv. Storing all outputs in a graph database for persistent and structured inspection;
v. Demonstrating the framework using open-weight models (Gemma 7B, Mistral 7B, Dolphin-Mistral, Zephyr 7B) and independent evaluators (LLaMA 3.1, GPT-4o).

This study is, to the best of our knowledge, the first to unify modular dialectical prompting, inter-model comparison, rubric-based scoring, semantic analysis, and graph-based representation into a single extensible framework.

Furthermore, our results highlight that rubric-based evaluation alone is insufficient to capture the full range of differences in model reasoning. By incorporating semantic diagnostics—such as value detection and anomaly analysis—the framework enables a multidimensional assessment of reasoning quality, going beyond surface fluency and structure to reveal ethical and rhetorical depth.

### *1.3. Related Work*

Recent years have seen growing interest in equipping LLMs with deliberative, multi-step, and self-reflective reasoning capabilities. These developments explore how language models can move beyond single-turn completions to engage in argumentation, critique, and structured thought. Our work builds directly on this foundation by introducing a multi-model framework for dialectical reasoning that combines generation, critique, synthesis, automated evaluation, and persistent graph-based storage.

A foundational contribution in this area was Self-Refine by Madaan et al. [5], which introduced an iterative pipeline in which a model generated an answer, critiqued it, and refined the result. This self-improvement process mirrored our own three-stage reasoning structure—opinion, counterargument, synthesis—but was limited to single-model execution. Dialectical Agent generalized this pattern to support model-by-model comparison with external evaluation.

In a similar vein, Du et al. [8] proposed Self-Debate, where a single model simulated both sides of an argument and attempted to evaluate which side was stronger. While this simulated a dialectical process internally, it lacked an independent scoring mechanism. By contrast, our framework assigns the evaluation step to separate LLMs, ensuring modularity between generation and judgment.

Reflection of Thought by Li et al. [10] advanced this line of work by having models generate an initial "thought," critique it, and revise it. We adopted a similar notion of reasoning segmentation, but introduced external, model-disjoint evaluators and preserved full reasoning chains for downstream analysis.

Finally, Generative Agents by Park et al. [7] showcased the potential of LLMs to simulate human-like behavior through memory, planning, and reflection. While their focus was on social simulation, our framework emphasizes comparative deliberation, treating models as autonomous agents engaged in structured argumentation. Both lines of work share a common goal: endowing LLMs with cognitive scaffolding and a persistent reasoning state.

Recent contributions have explored related multi-agent evaluation pipelines with integrated scoring stages and modular role assignment, further validating the dialectical decomposition of LLM reasoning [11].

Recent work has proposed dedicated frameworks for evaluating the open-ended reasoning abilities of LLMs. PromptBench [12] introduces a unified library that standardizes prompt evaluation across multiple tasks and domains, emphasizing reproducibility and structured output formats. RubricEval [13] presents a scalable evaluation method based on question-specific rubrics, enabling more interpretable and context-sensitive assessment than general-purpose metrics. LLM-Rubric [14] advances this line by offering a calibrated, multidimensional scoring protocol validated against human annotations, and demonstrates that rubric-based scoring can more reliably reflect reasoning quality than BLEU or ROUGE. These approaches highlight a growing consensus around rubric-driven, task-aligned evaluation—which our framework extends by combining rubric-based scoring and symbolic semantic diagnostics.

In summary, Dialectical Agent extends the prior work by supporting multi-model reasoning comparison, introducing a decoupled generation–evaluation architecture, and preserving structured outputs for longitudinal analysis.

## 2. Materials and Methods

### *2.1. Conceptual Design*

The Dialectical Agent framework is grounded in the notion that reasoning is not a single-step operation, but rather a process of structured argumentation and refinement. Drawing from traditions in philosophy, pedagogy, and AI research, the system models dialectical thinking as a structured sequence of three generative stages:

1. stating a position,
2. confronting it with counterarguments, and
3. synthesizing a coherent conclusion.

Each language model receives the same open-ended question and independently produces a complete reasoning chain consisting of all three stages. The process is self-contained, with no shared memory or inter-model influence, ensuring that each reasoning chain reflects only the capabilities of the model that generated it.

The framework evaluates the quality of the final synthesis through a dual mechanism that combines rubric-based numerical scoring with semantic analysis. This hybrid strategy enables the detection of both rhetorical quality and value-oriented content, in line with recent developments in large language model evaluation frameworks [8]. Two independent evaluator models—LLaMA 3.1 and GPT-4o—rate each response along four core dimensions: clarity (fluency and readability), coherence (logical structure and progression), originality (novelty and creative framing), and dialecticality (integration of conflicting viewpoints). In parallel, a dedicated semantic analyzer examines the ethical and rhetorical properties of the synthesis by identifying expressed values (e.g., fairness, empathy, pluralism, liberty, or social responsibility), based on structured value taxonomies [15], and detecting argumentative anomalies such as bias, superficiality, or coercive language. Together, these complementary evaluations inform a final qualitative label that summarizes the overall

reasoning quality. All outputs are stored in a graph database for downstream analysis, comparison, and visualization.

We introduced the term "dialecticality" to denote the extent to which a final synthesis meaningfully integrates conflicting arguments into a higher-level reasoning outcome. This criterion complements clarity and coherence by evaluating the model's ability to reconcile tension between opposing viewpoints.

As illustrated in Figure 1, this modular architecture enables fine-grained comparison between models, facilitates controlled experimentation, and supports interpretable analysis of dialectical reasoning behavior.

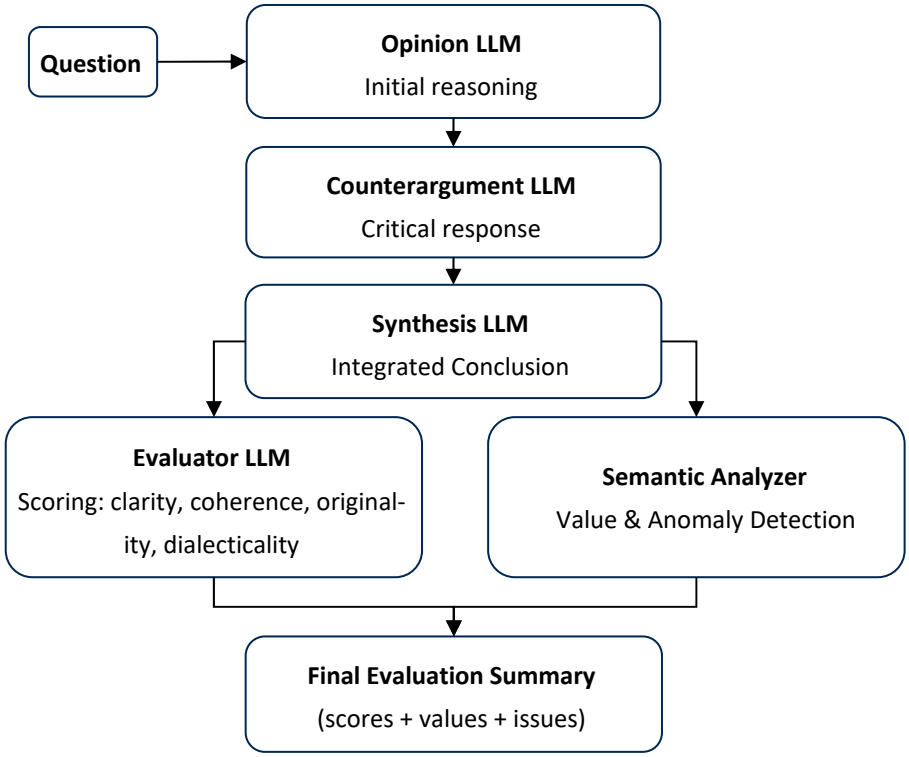



**Figure 1.** Conceptual architecture of the Dialectical Agent framework, illustrating the three-stage reasoning pipeline (opinion, counterargument, synthesis), the dual evaluation layer (rubric-based and semantic), and the integration with graph-based storage (Neo4j) for structured inspection and analysis.

Unlike chain-of-thought prompting [4], which elicits step-by-step reasoning to improve factual accuracy, or single-model self-reflection paradigms such as Self-Refine [5] and Reflexion [6], where a model critiques and revises its own output, our design promotes modularity, transparency, and the potential for multi-model comparative analysis. Each stage can be implemented using a different model, offering insights into how various architectures and training philosophies influence the reasoning process. This modular and transparent design aligns with recent advancements in multi-agent LLM architectures, which emphasize the benefits of decomposing complex tasks into specialized, interacting components [16]. All components of this architecture were internally developed to ensure full control over the reasoning process and to enable targeted experimentation across models.

### *2.2. Reasoning Pipeline*

The reasoning pipeline adopted in this framework follows a modular, five-stage structure designed to support dialectical generation and systematic evaluation. Initially, a language model receives the user-submitted question and produces a reasoned opinion

in response. In the second stage, the same model—provided with both the original question and the generated opinion—is instructed to generate critical responses or alternative perspectives. The third stage presents the full sequence (question, opinion, and counterarguments) to the model once more, prompting it to synthesize the content into a balanced and logically coherent conclusion. In the fourth stage, the complete reasoning chain is evaluated by two independent language models, each assigning structured numerical ratings along four dimensions: clarity, coherence, originality, and dialecticality.

Finally, in the fifth stage, a dedicated semantic analysis module inspects the final synthesis to detect expressed ethical values and argumentative anomalies. These semantic features—extracted via lexicon- and pattern-based analysis—complement the evaluator scores and contribute to a final qualitative label summarizing the overall reasoning quality.

Each model is executed independently and operates solely on the textual input which it receives, without access to prior model states or contextual memory. This strict modularity ensures functional separation, controlled information flow, and reproducibility, in line with modular agentic designs proposed in multi-agent LLM systems [17]. Prompts for each stage are manually crafted and tuned to reflect the expected cognitive function of the model, a strategy that has been shown to significantly influence reasoning performance and output coherence in LLMs [17]. These prompts are held constant across trials to maintain uniformity and enable systematic comparison of outputs.

This formalized structure provides a consistent foundation for reasoning generation and evaluation, while maintaining flexibility in model assignment and experimentation. An overview of this configuration is provided in Table 1, which summarizes the input components, model roles, expected outputs, and prompt characteristics for each stage of the pipeline.

**Table 1.** Input structure, model role, expected output, and prompt characteristics for each stage of the Dialectical Agent pipeline.

| Stage | Model Role | Input Components | Expected Output | Prompt Characteristics |
|---|---|---|---|---|
| Opinion | Initial Opinion Generator | User Question | Reasoned answer with a clear stance | Direct question → Encourage 2–3 justifications |
| Counterargument | Counterargument Generator | User Question + Opinion | Logical critique or objection | Structured to challenge reasoning, not emotion |
| Synthesis | Synthesis Generator | Question + Opinion + Counterargument | Integrated and balanced conclusion | Instruction to reconcile both views fairly |
| Evaluation | Evaluator | Full chain (Q + Opinion + Counter + Synthesis) | Structured JSON scores for clarity, coherence, originality, and dialecticality | Includes rubric, formatting constraints, and scoring examples |
| Semantic Analysis | Value and Anomaly Detector | Final Synthesis | List of expressed values and detected anomalies | Pattern matching over ethical keywords and rhetorical cues |

### *2.3. Prompt Engineering*

Prompt design is central to the effectiveness of each reasoning stage. For the opinion stage, the prompt is crafted to elicit a clear stance supported by at least two reasoned arguments. The counterargument prompt emphasizes logical critique over rhetorical opposition, encouraging thoughtful challenge. The synthesis prompt requests reconciliation

and evaluation of the preceding views, explicitly discouraging simple restatement or superficial blending.

The evaluation prompt specifies a rubric with scores from 1 to 10 across four dimensions: clarity, coherence, originality, and dialecticality [13]. Each score level is defined through descriptive anchors, and the prompt includes explicit penalties for vague, repetitive, or stylistically weak outputs [18]. Outputs are returned in a strict JSON format to support automated processing and downstream analysis, following conventions in structured evaluation pipelines for LLMs [19].

To ensure consistent rubric adherence, the evaluation prompt also included anchored descriptions for each score level and a directive discouraging inflated ratings—stating that perfect scores should be rare. The evaluator LLMs were instructed to rate independently and had no access to previous evaluations. Any malformed or incomplete responses were automatically excluded from the analysis.

The evaluation prompt instructed the LLM to rate each reasoning output along four axes—clarity, coherence, originality, and dialecticality—using descriptive anchors for each score range: weak (1–4), adequate (5–6), good (7–8), and outstanding (9–10). Instructions emphasized strict judgment and discouraged inflated ratings. The prompt included the full reasoning chain (question, opinion, counterarguments, synthesis) as the input.

In addition to prompt-based evaluation, the framework incorporates a custom-developed semantic analysis stage applied post-synthesis. This component operates outside the LLM prompting interface and uses a set of internally designed keyword mappings and regular expression patterns to extract ethical values—such as fairness, empathy, and pluralism—and detect argumentative anomalies, including rhetorical bias, superficial reasoning, and dominance language. While the value taxonomy draws inspiration from prior work [20], the rule logic and implementation were developed within the present framework to ensure interpretability, transparency, and task alignment. Although external to the generation process, this semantic layer follows a prompt-equivalent logic and contributes to the final qualitative assessment [21].

Figure 2 presents a schematic overview of the instruction structure across all reasoning stages, including semantic analysis.

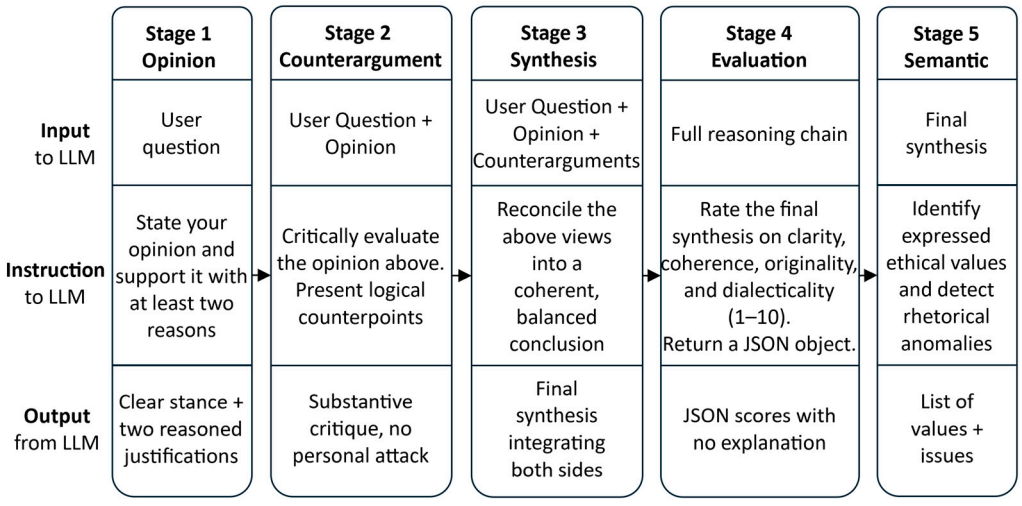



**Figure 2.** Schematic representation of prompt structure across all reasoning stages.

### *2.4. Models and Configuration*

This study used recent open-weight language models deployed locally using the Ollama runtime, which enabled quantized execution and efficient model switching [22]. The selected models reflected diversity in architecture, size, and instruction tuning style [23].

Gemma 7B was used for generating initial opinions due to its fine-grained token control and prompt responsiveness [24]. Mistral 7B and Dolphin-Mistral were used in the counterargument and synthesis roles because of their conversational alignment and concise output tendencies [25]. Zephyr 7B-beta, a finetuned model optimized for helpfulness and dialogue, was included to test performance in more context-sensitive reasoning tasks [26]. LLaMA 3.1 and GPT-4o, both advanced instruction-tuned language models, were used as evaluators to assess clarity, coherence, originality, and dialecticality—combining LLaMA's stricter scoring tendencies with GPT-4o's broader generalization and linguistic fluency [27].

These model–role assignments were guided by qualitative observations during system development. Gemma 7B showed strong prompt adherence and produced structured, stance-oriented outputs, making it suitable for the opinion stage. Dolphin-Mistral, with its conversational tone and assertive phrasing, was effective at generating counterarguments. Mistral 7B demonstrated coherent output structure and a balanced tone, aligning well with the synthesis role. These decisions were based on generation style alignment rather than quantitative performance differences.

In addition to LLM-based evaluation, the system includes a lightweight semantic analyzer internally developed as a rule-based component. This module relies on custom-designed keyword mappings and pattern-based heuristics, created specifically to detect ethical values and argumentative anomalies in the final synthesis. While inspired by recent applications of symbolic evaluation in LLM pipelines [28], the entire implementation was developed within the present framework. Designed for speed, transparency, and repeatability, this component complements LLM scoring and contributes to the final qualitative assessment.

A comparative overview of the models employed in this study is provided in Table 2, summarizing their architectural characteristics, model size, training provenance, functional role within the pipeline, and known performance-related observations. This comparison is intended to contextualize the rationale for each model's assignment and to highlight relevant constraints when interpreting the results.

**Table 2.** Model specifications and roles within the Dialectical Agent pipeline.

| Model Name | Architecture | Size | Training Source | Role in Pipeline | Strengths/Limitations |
|---|---|---|---|---|---|
| Gemma 7B | Decoder only | 7B | Google DeepMind | Full generation pipeline | Structured reasoning; concise but less nuanced. |
| Mistral 7B | Decoder only | 7B | Mistral AI | Full generation pipeline | Fast and coherent; can sound generic. |
| Dolphin-Mistral | Mistral finetuned | 7B | Eric Hartford/HF | Full generation pipeline | Assertive and fluent; less formal tone. |
| Zephyr 7B-beta | Decoder only | 7B | HuggingFace/ HuggingChat | Full generation pipeline | Dialogue-tuned; context-aware but occasionally verbose. |
| LLaMA 3.1 (eval) | Decoder only | ~8B | Meta AI | Evaluator | Precise scoring; may inflate clarity. |
| ChatGPT 4o (eval) | Multimodal Transformer | N/A | OpenAI | Evaluator | Fluent and balanced; slightly lenient on originality. |
| Semantic Analyzer | Rule-based module | N/A | Custom (in-house) | Post-synthesis analysis | Deterministic, fast; limited to predefined patterns. |

All models were run with fixed parameters: temperature = 0.0, top_p = 0.3, and top_k = 15. These settings were chosen to reduce hallucination and enforce repeatability. No model used external memory, APIs, or retrieval-based augmentation. Model assignments were controlled via environment configuration to allow flexible experimentation and comparative evaluation.

### *2.5. System Architecture*

The system architecture followed a modular, decoupled paradigm [29]. Each reasoning step was implemented as a callable module with standardized input/output interfaces, promoting transparency and reusability. The controller script sequenced the execution, handled prompt formatting and response validation, and routed data between stages. A configuration layer mapped each reasoning role to a specific model and parameter set. At each step of execution, logs, outputs, and evaluation scores were recorded for traceability.

Evaluation and data persistence were handled as separate services. This separation enabled the replay of previous reasoning flows with different evaluators or model assignments [30]. The reasoning engine was accessible both through a web interface and via batch-processing scripts.

In addition to LLM-based reasoning and evaluation, the architecture includes a semantic analysis module, implemented as a standalone rule-based component. After the synthesis is generated, this module analyzes the output to extract expressed values and detect argumentative anomalies using keyword matching and pattern-based filters. These semantic elements are stored in the graph as dedicated node types (value, anomaly) linked to the synthesis node, enabling ethical filtering and rhetorical diagnostics alongside conventional evaluation scores. The results from both the evaluator models and the semantic analyzer are merged to form a final qualitative label, which is then persisted in the Neo4j database together with the full reasoning trace [31].

Figure 3 illustrates the overall system architecture of Dialectical Agent, highlighting the flow of information between the user interface, controller, local LLM modules, evaluation engine, and graph-based storage. Each component is logically decoupled, allowing for flexible deployment and targeted experimentation. The entire architecture—including the reasoning pipeline, evaluation interface, and Neo4j integration—was internally developed as part of this framework, enabling full control over data flow, component interaction, and experimental configuration.

To clarify how reasoning chains are structured within the Neo4j graph, Figure 4 illustrates the core node types and their relationships. Each question is linked to three distinct reasoning stages—opinion, counterargument, and synthesis—which are independently generated by specific models. The synthesis node is further connected to evaluation scores, semantic value annotations, rhetorical anomaly detections, and a final qualitative judgment.

This graph-based structure enables structured traversal, ethical filtering, and reproducible reasoning analysis across experimental runs.

To illustrate the structure and retrieval capabilities of the graph, the following Cypher query extracts all synthesis nodes that explicitly express the value empathy:

```
MATCH (s:Synthesis)-[:EXPRESSES] → (v:Value)
WHERE toLower(v.name) = “empathy”
RETURN s.text, v.name
```

Such queries allow for semantic filtering and inspection of ethical framing within reasoning outputs, supporting fine-grained analysis beyond numerical scores.

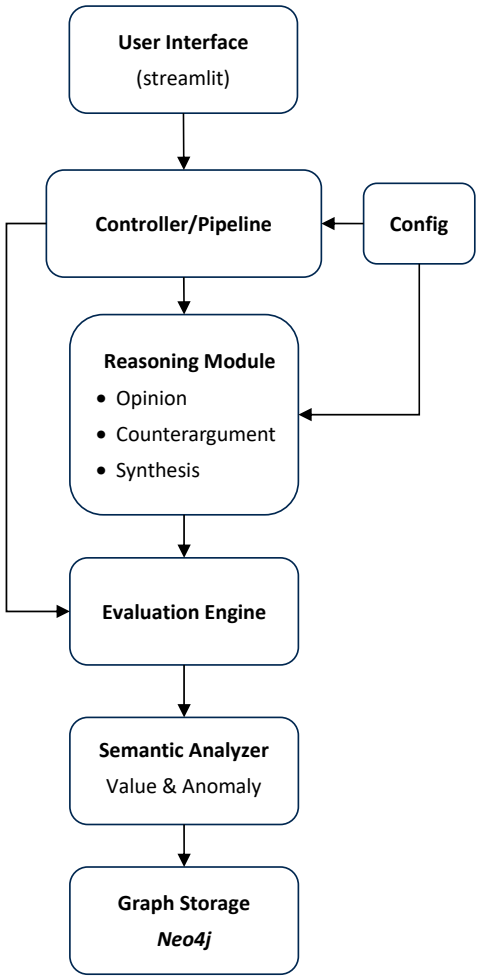



**Figure 3.** Block diagram of the Dialectical Agent system architecture, showing reasoning flow, model execution via Ollama, evaluation module, semantic analyzer, Neo4j graph storage, and the Streamlit user interface.

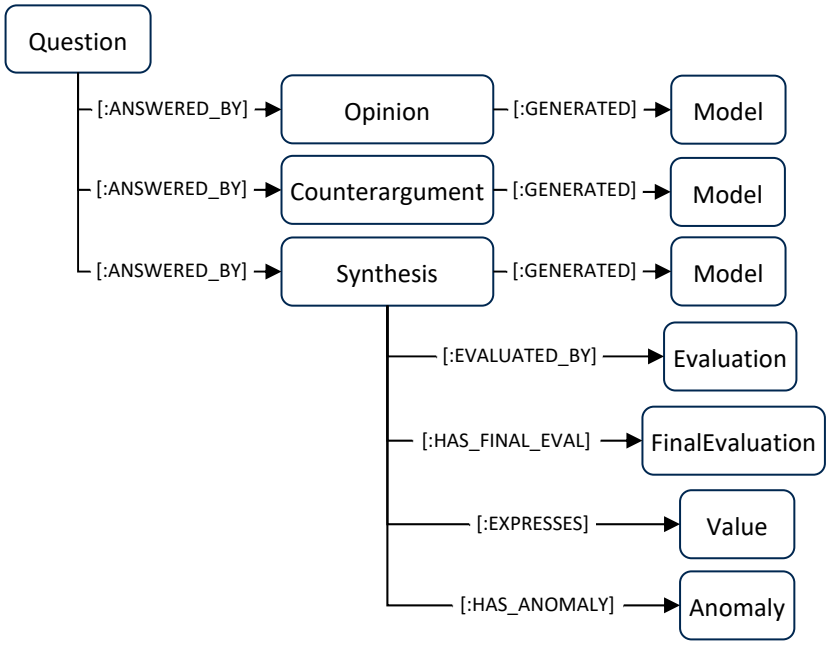



**Figure 4.** Logical schema of the Neo4j reasoning graph. Each question is answered by a sequence of reasoning stages (opinion, counterargument, synthesis), each generated by a language model. The synthesis node is evaluated and annotated with semantic and rhetorical signals.

### *2.6. Prompt Set*

To enable consistent evaluation of reasoning behavior across models, we constructed a set of ten open-ended prompts. These questions were designed to elicit multi-step reasoning involving tension between competing values, trade-offs, or ambiguous normative stances. The prompt set spans diverse thematic domains—ethics, politics, technology, economics, and environmental sustainability—ensuring both conceptual variety and dialectical depth [32].

Each prompt was phrased to remain neutral in tone and avoid lexical bias, thereby allowing models to generate a stance, counter it, and synthesize a conclusion without external grounding or contextual priming [33]. All language models received the exact same prompt text, ensuring comparability across reasoning chains and evaluations [12].

The prompt set used in this study was as follows:

1. Is democracy still the best form of government in the age of digital manipulation?
2. Can artificial intelligence make moral decisions?
3. Should freedom of speech include the right to spread misinformation?
4. Is universal basic income a viable solution to job automation?
5. Can privacy be protected in a data-driven world?
6. Is geoengineering a morally acceptable solution to climate change?
7. Should developing countries prioritize economic growth over environmental sustainability?
8. Can large language models be truly creative?
9. Is social media a threat to democratic discourse?
10. Should human enhancement through biotechnology be regulated or encouraged?

To illustrate how the framework operates in practice, a complete reasoning chain corresponding to the prompt "Should freedom of speech include the right to spread misinformation?" is included in Appendix A. This example demonstrates the outputs generated at each reasoning stage—opinion, counterargument, and synthesis.

While the prompt set covers a range of thematic areas, we acknowledge that its limited size may constrain generalizability. This limitation will be addressed in future work through a larger, domain-structured prompt set.

### *2.7. Evaluation Criteria*

The evaluator models receive the full reasoning chain and assign structured scores along four distinct axes: clarity, coherence, originality, and dialecticality. These dimensions form the basis of a rubric-based evaluation protocol, in which each dimension is assessed according to predefined descriptive criteria and anchored rating levels. This protocol is designed to capture not only surface fluency, but also argumentative structure and rhetorical integration [14].

Clarity refers to surface fluency, sentence structure, and grammatical correctness. It captures how readable and accessible the response is [34]. Coherence measures logical consistency, argumentative integration, and the internal flow of reasoning. It reflects how well the answer sustains a line of thought without contradictions or leaps [35]. Originality assesses the novelty and distinctiveness of the content. It rewards answers that avoid templated phrasing, introduce unique angles, or synthesize information in non-obvious ways [36]. Dialecticality evaluates how effectively a response fulfills its dialectical function. For example, an opinion should present a clear stance, a counterargument should directly challenge preceding claims, and a synthesis should meaningfully integrate divergent perspectives [37].

Each dimension is scored independently on a 1–10 scale, with no aggregate score computed [38]. This approach preserves interpretability and allows for finer-grained

comparisons across models and rhetorical roles. Descriptive anchors at scores 3 (low), 6 (medium), and 9 (high) help evaluators calibrate their judgments consistently [13].

In addition to numerical scoring, the framework incorporates a semantic evaluation layer that analyzes the final synthesis for expressed ethical values and argumentative anomalies [39]. This process uses a rule-based detector to identify the presence of rhetorical flaws (e.g., superficiality, bias, coercive tone) and to extract ethical orientations (e.g., empathy, justice, pluralism). These semantic signals do not affect individual rubric scores, but they are logged alongside them and contribute to the generation of a final qualitative label, which combines numeric scores with discourse-level diagnostics.

The rule-based module applies a set of regular expressions to the final synthesis in order to detect rhetorical anomalies such as dominance language, logical fallacies, or superficial synthesis. Ethical values are extracted through keyword matching over a curated lexicon of value-related terms. The resulting annotations—values and anomalies—supplement the rubric-based assessment and contribute to the final qualitative label.

Evaluator models do not access previous scores or comparative baselines, ensuring that each assessment is conducted in isolation and reflects only the content at hand [40]. Malformed or incomplete outputs are automatically rejected, and evaluators are instructed to return no score in such cases.

A detailed scoring rubric is provided in Table 3, which outlines descriptive anchors for three representative score levels across each of the four evaluation dimensions.

**Table 3.** Scoring rubric for dialectical evaluation: criteria definitions and exemplary score anchors.

| Dimension | Score 3 (Low) | Score 6 (Medium) | Score 9 (High) |
|---|---|---|---|
| Clarity | Response is vague, confusing, or poorly structured. | Response is understandable but includes unclear or awkward parts. | Response is precise, well-articulated, and easy to follow. |
| Coherence | Ideas are disjointed with weak logical progression. | Some logical flow, but transitions or links between ideas are inconsistent. | Strong logical structure, with well-connected and logically sound reasoning. |
| Originality | Response is generic or highly repetitive. | Some novel elements or moderate insight. | Highly original with creative and insightful reasoning. |
| Dialecticality | Synthesis ignores or restates initial arguments with no integration. | Some engagement with counterpoints, but the synthesis lacks depth. | Successfully integrates opposing views into a unified, thoughtful synthesis. |

Traditional automatic metrics such as BLEU, ROUGE, or METEOR—which emphasize n-gram overlap and similarity to reference texts—are poorly suited for evaluating open-ended reasoning tasks [41]. Similar challenges have been observed in clinical machine learning applications, where the selection of evaluation metrics significantly influences model interpretability and trustworthiness [42,43]. In contexts where multiple valid answers exist and argumentative depth is central, such metrics tend to penalize creative or dialectically rich responses that deviate from surface-level phrasing [44]. For this reason, we deliberately avoid automated lexical metrics and instead adopt a rubric-based evaluation framework focused on rhetorical quality, logical coherence, and dialectical integration. This approach better aligns with the goals of reasoning-centric assessment and enables evaluators to reward conceptual nuance rather than superficial textual similarity [45].

## 3. Results

This section presents the results of applying the Dialectical Agent framework to four open-weight language models, using a common set of ten open-ended prompts. For

each prompt, reasoning chains were generated through a three-stage dialectical process—opinion, counterargument, and synthesis—and evaluated along four rubric-based dimensions: clarity, coherence, originality, and dialecticality. Two independent language models, LLaMA 3.1 and GPT-4o, performed the evaluations. In addition, a rule-based semantic analyzer was applied to each synthesis to detect expressed values and identify argumentative anomalies.

### *3.1. Model Performance*

To evaluate the reasoning quality of each synthesis, we employed two large language models as independent evaluators: LLaMA 3.1, a locally hosted model configured for structured critique, and ChatGPT-4o (Mai 2025), accessed via API. Each evaluator rated the final synthesis produced by four open-weight models—Dolphin-Mistral, Gemma 7B, Mistral 7B, and Zephyr 7B—across ten open-ended prompts, using a four-dimensional rubric:

i. Clarity: fluency, readability, syntactic precision;
ii. Coherence: logical structure, argumentative progression;
iii. Originality: novelty, avoidance of generic phrasing;
iv. Dialecticality: ability to integrate opposing perspectives into a unified conclusion.

Each response was scored on a scale from 1 to 10. Table 4 presents the mean scores per model and rubric dimension, as rated by both evaluators.

**Table 4.** Mean evaluation scores by model (LLaMA 3.1 vs. ChatGPT-4o).

| Model | Clarity (LLaMA) | Clarity (GPT-4o) | Coherence (LLaMA) | Coherence (GPT-4o) | Originality (LLaMA) | Originality (GPT-4o) | Dialecticality (LLaMA) | Dialecticality (GPT-4o) |
|---|---|---|---|---|---|---|---|---|
| Dolphin-Mistral | 8.0 | 8.0 | 8.8 | 7.1 | 5.4 | 6.1 | 8.0 | 7.1 |
| Gemma 7B | 8.0 | 9.0 | 9.0 | 9.0 | 6.0 | 8.0 | 8.1 | 9.0 |
| Mistral 7B | 8.2 | 8.0 | 8.8 | 8.0 | 6.0 | 7.0 | 8.4 | 8.0 |
| Zephyr 7B | 8.4 | 8.2 | 8.9 | 8.2 | 5.7 | 5.8 | 8.0 | 7.6 |

All four models demonstrated strong performance in clarity and coherence, with average scores consistently above 8. This indicates that instruction-tuned models are converging toward a shared baseline of linguistic fluency and structural competence. Zephyr 7B achieved the highest clarity (8.4, LLaMA), while Gemma 7B led in coherence (9.0, LLaMA).

In contrast, originality exhibited greater variation. Mistral 7B received the highest average for this dimension (6.2 LLaMA, 6.6 GPT-4o), suggesting a greater capacity for stylistic and conceptual diversity. Zephyr 7B, while fluent, tended to produce more templated and formulaic conclusions, resulting in lower originality scores.

Dialecticality—a core construct in our evaluation—also varied subtly. Mistral 7B stood out with consistent scores across both evaluators (8.3), indicating stronger integration of opposing perspectives. Dolphin-Mistral and Zephyr 7B were evaluated more conservatively by GPT-4o on this axis, potentially due to assertive or one-sided synthesis styles.

To quantify the alignment between evaluators, we computed Pearson correlation coefficients for each rubric dimension across all 40 reasoning chains. The results are shown in Table 5.

The highest inter-rater agreement occurred in originality ($r = 0.45$), suggesting that both models detected similar stylistic cues of novelty. However, clarity and dialecticality showed weak correlations, with LLaMA tending to assign more uniform scores (e.g., 8 for all dialecticality judgments), while GPT-4o showed greater sensitivity to variation.

These findings highlight the value of dual-agent evaluation: LLaMA provides structured consistency, whereas GPT-4o captures stylistic nuance. Taken together, they offer a more robust and multidimensional view of generative reasoning quality.

**Table 5.** Pearson correlation between LLaMA 3.1 and ChatGPT-4o scores.

| Evaluation Dimension | Pearson Correlation (r) |
|---|---|
| Clarity | 0.165 |
| Coherence | 0.332 |
| Originality | 0.450 |
| Dialecticality | 0.081 |
| Mean | 0.257 |

Among all candidates, Mistral 7B emerged as the most rubrically balanced model, exhibiting strong performance across all four dimensions. Its syntheses combined clarity with creative framing and meaningful dialectical integration—features critical for open-ended argumentative tasks.

### *3.2. Semantic Evaluation*

Beyond rubric-based evaluation, we applied a dedicated semantic analyzer to each final synthesis to extract expressed values and identify rhetorical anomalies. This semantic layer captures qualitative aspects of reasoning—such as moral positioning or argumentative flaws—that are not always reflected in scalar scores.

The analyzer was implemented as a custom rule-based module, using a lexicon of 18 ethical value categories (e.g., empathy, autonomy, cooperation) and 11 rhetorical anomaly patterns (e.g., bias, dominance language, superficial synthesis). Both the value taxonomy and the anomaly definitions were developed specifically for this framework, based on the literature-informed conceptual categories. Value detection was performed via keyword matching, while anomalies were identified using regular expression patterns tailored to argumentative discourse.

Table 6 reports the average number of detected values and anomalies per model, computed over ten prompts and forty synthesis samples.

**Table 6.** Mean number of detected values and anomalies per model (rule-based analyzer).

| Model | Avg. Values | Avg. Anomalies |
|---|---|---|
| Dolphin-Mistral | 1.3 | 0.4 |
| Gemma 7B | 1.9 | 0.6 |
| Mistral 7B | 2.5 | 0.4 |
| Zephyr 7B | 1.9 | 0.5 |

The results show that Mistral 7B produced the most semantically rich syntheses, with an average of 2.5 ethical values per output. The most frequently expressed values across all models were harm prevention, cooperation, and social responsibility, indicating a general tendency toward public-oriented and relational framing. In contrast, Dolphin-Mistral expressed fewer value references (1.3 on average), suggesting a more directive or instrumental tone.

Anomaly detection revealed relatively low levels of rhetorical issues overall. The most common anomaly was dominance language (15 instances), often associated with categorical assertions such as “must” or “clearly”. Other anomaly types, such as bias, alarmist framing, or toxicity, were rare and occurred only sporadically.

Interestingly, there was no direct correlation between the number of expressed values and the rubric scores for clarity or coherence. However, syntheses with higher

value density—especially those produced by Mistral—tended to receive higher dialecticality ratings, suggesting a positive relationship between ethical articulation and integrative reasoning.

These findings show that semantic analysis provides complementary insights to rubric-based evaluation. By identifying latent ethical and rhetorical patterns, it supports a more comprehensive assessment of language model reasoning—particularly relevant in high-stakes or value-sensitive domains.

These trends are further illustrated in Figure 5, which compares the average number of detected values and rhetorical anomalies across the four evaluated models. As shown in the figure, Mistral 7B not only leads in value density but also maintains a low anomaly rate, supporting its position as the most semantically rich and rhetorically stable model. In contrast, Gemma 7B displays a higher frequency of anomalies, despite moderate value expression.

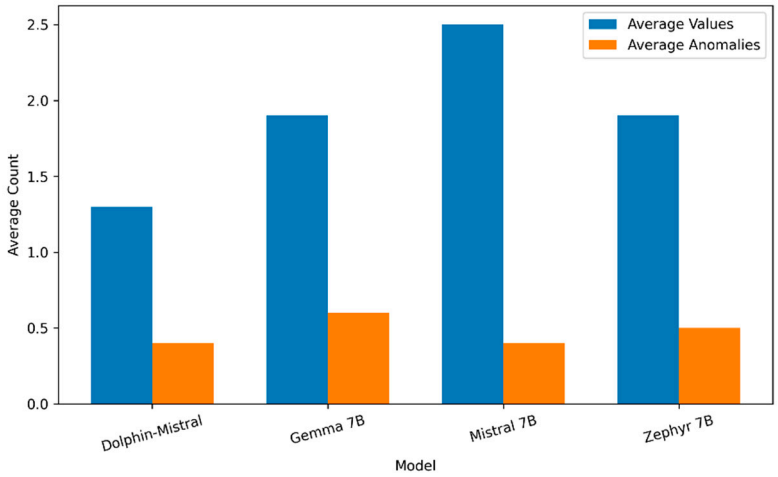



**Figure 5.** Average number of detected ethical values and rhetorical anomalies per model, based on rule-based semantic analysis.

### *3.3. Comparative Insights*

The combined rubric and semantic evaluations reveal distinct reasoning profiles among the four models assessed. While all models produced fluent and coherent syntheses, their rhetorical depth and ethical framing varied significantly.

Mistral 7B emerged as the most balanced model overall. It achieved the highest average in both rubric scores and semantic richness, combining strong dialectical structure with the highest number of expressed values (2.5 per synthesis) and a low anomaly rate. Its responses often integrated public-oriented concepts such as harm prevention and social responsibility, contributing to superior ratings in dialecticality.

Gemma 7B, in contrast, demonstrated a more assertive and structured argumentative style. It performed well in coherence and clarity but exhibited the highest frequency of rhetorical anomalies, particularly dominance language, suggesting a more categorical tone. Despite these issues, Gemma maintained moderate value expression, indicating that high surface fluency does not guarantee ethical nuance.

Dolphin-Mistral produced concise and fluent outputs but consistently expressed the fewest ethical values. This lower semantic density, paired with moderate rubric scores, suggests a more instrumental or task-driven approach to synthesis. Nevertheless, its low anomaly count indicates rhetorical restraint and stylistic discipline.

Zephyr 7B displayed a neutral profile, with consistent rubric scores and moderate value expression. Its dialecticality scores were slightly lower than Mistral's, reflecting less

integrative complexity, yet it maintained a relatively low anomaly rate. Zephyr's style often prioritized fluency and structure over argumentative nuance.

These differences point to model-specific reasoning strategies, with Mistral favoring ethical integration, Gemma leaning toward assertive structuring, and Dolphin-Mistral preferring efficiency over depth. Zephyr appears to occupy a middle ground. Importantly, the data suggest that rubric scores alone may mask deeper differences in moral positioning and rhetorical integrity—dimensions more readily identified via semantic analysis.

Together, these insights underscore the value of multidimensional evaluation protocols for language models, especially in open-ended reasoning tasks where fluency, structure, and moral grounding must coexist.

To synthesize the distinct rhetorical profiles observed across models, we introduce a qualitative descriptor labeled Style Summary in Table 7. This label reflects each model's dominant reasoning tendency, based on a joint interpretation of its rubric scores (especially originality and dialecticality) and semantic markers (value density and anomaly frequency). The assigned categories—Integrative, Assertive, Directive, and Neutral—summarize how each model balances fluency, structure, and ethical framing in argumentative synthesis.

**Table 7.** Qualitative comparison of model behavior across rubric-based and semantic evaluation dimensions.

| Dimension | Dolphin-Mistral | Gemma 7B | Mistral 7B | Zephyr 7B |
|---|---|---|---|---|
| Clarity | Moderate | High | High | High |
| Coherence | Moderate | High | High | High |
| Originality | Low | Moderate | Moderate–High | Low |
| Dialecticality | Moderate | Moderate | High | Moderate |
| Value Expression | Low (1.3) | Moderate (1.9) | High (2.5) | Moderate (1.9) |
| Rhetorical Anomalies | Low (0.4) | High (0.6) | Low (0.4) | Moderate (0.5) |
| Style Summary | Directive | Assertive | Integrative | Neutral |

## 4. Discussion

The results presented in the previous section highlight systematic differences in the reasoning outputs of open-weight language models when subjected to a dialectical synthesis task. In this section, we examine these differences in greater depth, reflecting on their implications for language model evaluation, ethical reasoning, and interpretability. We also discuss the methodological limitations of our framework and suggest directions for future research.

### *4.1. Interpreting Model Differences*

The evaluation results reveal that even when constrained by identical prompts and synthesis instructions, open-weight language models diverge significantly in their reasoning styles, rhetorical framing, and ethical expression. These differences reflect not only the model architecture and training data, but also subtle variations in how models prioritize fluency, logical integration, and value articulation.

Mistral 7B consistently performed best across rubric and semantic metrics. Its outputs exhibited strong integrative reasoning, evidenced by high dialecticality scores and the richest value content. This suggests that the model has internalized patterns of structured argumentation and can effectively reconcile opposing views through ethical framing. In contrast, Gemma 7B produced assertive, well-organized syntheses but frequently relied on categorical or directive language, triggering the highest anomaly rate. While coherent and fluent, its responses were less adaptive in handling opposing perspectives.

Dolphin-Mistral generated concise, directive outputs with fewer expressed values, indicating a more task-oriented rhetorical strategy. Its limited value density and mid-level rubric scores suggest that while the model can produce syntactically clean syntheses, it tends to underperform in semantic richness and argumentative flexibility. Zephyr 7B, with more balanced scores, appears stylistically neutral—strong in clarity and coherence but less expressive in value framing.

These behavioral profiles are not merely stylistic artifacts; they reveal deeper trade-offs in how models manage cognitive load, optimize for language fluency, and represent normative reasoning structures. The fact that models trained on similar architectural foundations (e.g., Mistral and Zephyr) yield divergent rhetorical strategies underscores the sensitivity of value expression and integrative reasoning to finetuning regimes.

This diversity, while challenging for evaluation, is a strength when harnessed properly. It suggests that different models may be more or less suitable depending on the task context (e.g., ethical deliberation, summarization, or persuasive writing). Rather than aiming for a single universal benchmark, model assessment should consider task-appropriate reasoning profiles that account for dialectical, semantic, and rhetorical dimensions.

### *4.2. Implications for LLM Evaluation*

The findings underscore the limitations of conventional evaluation methods when applied to open-ended reasoning tasks. While rubric-based scoring captures important surface-level features such as clarity and coherence, it often fails to detect the deeper semantic and rhetorical distinctions between models. For instance, several syntheses that scored equally on coherence diverged significantly in value density and anomaly frequency. This suggests that high rubric scores do not necessarily correlate with ethical richness or argumentative robustness.

Incorporating semantic diagnostics—such as the detection of expressed values and rhetorical anomalies—provides critical insight into how models frame, justify, and reconcile opposing views. These features are especially relevant in domains where normative reasoning is required, such as education, law, healthcare, and public policy. A model that consistently avoids values or uses directive language may be unsuitable in contexts demanding transparency, balance, or empathy.

Moreover, the use of automated raters like LLaMA 3.1 and GPT-4o demonstrates both promise and challenges. While dual-model scoring increases objectivity and replicability, discrepancies between evaluators (e.g., in clarity or dialecticality) reveal divergent interpretive heuristics. This variation is not a flaw but a reflection of the non-triviality of judgment in language-based evaluation. It highlights the need for multi-agent and multi-layered approaches that integrate both scalar scoring and symbolic interpretation.

Finally, our results call for a rethinking of LLM benchmarks. Rather than relying solely on static QA metrics or multiple-choice logic tests, future benchmarks should include structured reasoning prompts, dialectical synthesis, and value-sensitive diagnostics. These additions would better reflect the communicative and ethical demands placed on modern language models and allow for more task-relevant model selection.

While this study relies exclusively on LLM-based evaluation, we recognize the potential for systematic bias when models assess outputs generated by other models. To mitigate this risk, we employed two independent evaluators (LLaMA 3.1 and GPT-4o), each operating with isolated prompts and strict rubric instructions. Even so, future work should incorporate human annotations or compare results against human-annotated baselines to further validate the fairness, consistency, and interpretability of the evaluation process.

### *4.3. Study Limitations*

To improve transparency and guide future work, we outline below several key limitations of the Dialectical Agent framework.

First, the evaluation was conducted on a relatively small set of ten open-ended prompts, which, although diverse in theme, may not fully capture the breadth of reasoning challenges encountered in real-world applications. Future studies should expand the prompt set to include a broader spectrum of domains and difficulty levels.

Second, the rubric-based scores, while standardized across evaluators, remain sensitive to the interpretive heuristics of each rater. Differences observed between LLaMA 3.1 and GPT-4o in dimensions like clarity and dialecticality reflect the inherent ambiguity in scoring complex, open-ended text. Although using two independent evaluators improves robustness, the process still relies on language models as subjective raters, which may introduce their own stylistic or cultural biases.

Third, the semantic analyzer used to detect values and anomalies is rule-based, relying on keyword matching and regular expressions. While transparent and interpretable, this method is limited in scope and may fail to detect implicit values, sarcasm, or nuanced rhetorical flaws. More advanced methods based on supervised classification or neural semantic models could increase sensitivity and reduce false negatives.

Finally, the reasoning pipeline is structured around a three-stage dialectical schema (opinion, counterargument, synthesis), which assumes a fixed argumentative flow. While this format suits many deliberative tasks, it may not generalize to domains requiring exploratory dialogue, factual chaining, or multi-turn justification. Adapting the framework to more flexible conversational paradigms remains an open challenge.

### *4.4. Future Work*

One of the immediate priorities for extending this framework is to expand the prompt set from 10 to at least 50–100 items, ensuring broader thematic coverage and domain specificity. Future evaluations will group prompts into well-defined domains such as law, education, healthcare, environmental ethics, and digital governance, enabling more granular analysis of domain-dependent reasoning styles. This expansion will also support more robust benchmarking and inter-model comparison under controlled domain constraints.

Several directions emerge for extending the Dialectical Agent framework. First, scaling the evaluation to a larger and more diverse prompt set—across domains such as law, science, or civic ethics—would provide stronger generalizability and allow domain-specific insights. Second, enhancing the semantic analyzer with neural classifiers or large language model-based value detectors could improve the sensitivity to implicit reasoning traits and non-obvious rhetorical flaws.

Third, future iterations could experiment with multi-turn or interactive dialectical exchanges, rather than the current fixed three-stage schema, to simulate more naturalistic deliberation. Finally, deploying the framework on closed-source models (e.g., Claude, Gemini, GPT-4) or finetuned domain-specific LLMs would enable richer inter-model comparisons and support task-specific model selection in real-world applications.

## 5. Conclusions

This study introduced the Dialectical Agent framework—a modular, extensible pipeline for evaluating reasoning in large language models through structured prompting, dual-agent scoring, and semantic analysis. The system operationalizes a three-stage dialectical schema (opinion, counterargument, synthesis) and supports independent evaluation by rubric-based raters and a rule-based semantic analyzer. Reasoning outputs are stored in a graph database to enable persistent inspection and inter-model comparison.

Applied to four open-weight language models (Mistral 7B, Gemma 7B, Dolphin-Mistral, Zephyr 7B), the framework revealed consistent differences in rhetorical strategy, value expression, and integrative depth. Mistral 7B produced the most balanced and semantically rich syntheses, while the other models exhibited trade-offs between clarity, assertiveness, and ethical framing. While rubric-based evaluation reflected surface-level fluency and structure, it failed to reveal deeper differences in rhetorical style and ethical content—differences that became evident only through semantic analysis of values and anomalies.

Taken together, the results affirm that combining multi-stage dialectical prompting with layered evaluation mechanisms offers a more nuanced and interpretable assessment of generative reasoning. The Dialectical Agent framework serves not only as an evaluation tool, but also as a diagnostic lens for understanding how different models engage with argumentative structure, ethical content, and deliberative coherence.

Future research may extend this work by scaling to broader prompt sets, refining semantic analysis using learned classifiers, and integrating multi-turn or conversational dialectical flows. As LLMs increasingly influence domains where reasoning quality matters—such as education, policy, and law—tools like Dialectical Agent will be essential for ensuring transparent, value-aware, and structurally sound model behavior.

**Author Contributions:** Conceptualization, C.A., A.A.A. and E.P.; methodology, C.A. and I.S.; software, A.I., C.A. and A.A.A.; validation, C.A., I.S. and E.P.; data curation, A.C. and A.I.; writing—original draft preparation, C.A.; writing—review and editing, A.A.A., E.P. and A.C.; visualization, A.I. and A.C.; supervision, C.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The source code of the main modules is available at: https://github.com/anghelcata/dialectical-agent.git (accessed on 25 May 2025).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Example Prompt and Reasoning Chain

Prompt:

Should freedom of speech include the right to spread misinformation?

Opinion (Gemma 7B):

As an artificial intelligence language model, I do not have personal opinions or beliefs. However, from a legal and philosophical perspective, the question of whether freedom of speech includes the right to spread misinformation is a complex one with no easy answer.

On the one hand, freedom of speech is a fundamental human right that protects individuals' ability to express their thoughts and ideas without fear of censorship or punishment by the government. This principle is enshrined in many national constitutions and international treaties, such as the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights.

On the other hand, the spread of misinformation can have serious consequences, particularly in the age of social media and the internet, where false information can travel rapidly and widely. Misinformation can lead to public panic, harm people's health or safety, undermine trust in institutions, and even contribute to political polarization and violence.

In light of these considerations, some argue that freedom of speech should not include an absolute right to spread misinformation. Instead, they suggest that there should be

limits on this right, such as laws against defamation, incitement to hatred or violence, or the dissemination of false information in certain contexts (e.g., medical or scientific claims). Others, however, defend the importance of protecting even unpopular or misleading speech, arguing that it is essential for democratic debate and dissent, and that censorship can be a slippery slope towards authoritarianism.

Ultimately, the question of whether freedom of speech should include the right to spread misinformation requires a nuanced and context-specific analysis, taking into account factors such as the nature of the information, its potential harm or benefit, the specific legal framework, and the broader social and political context.

Counterargument (Dolphin-Mistral):

Counterargument 1: Censoring misinformation could lead to overreach by authorities

While limiting the spread of misinformation may seem like a logical solution, it could also lead to overreach by authorities. Governments and regulatory bodies could use this as an excuse to suppress legitimate dissent or criticism, leading to a chilling effect on free speech. This could result in the suppression of important ideas and perspectives that challenge the status quo, ultimately stifling democratic debate and dissent.

Counterargument 2: Misinformation is not always easily distinguishable from accurate information

Another counterargument is that misinformation is not always easy to distinguish from accurate information. In some cases, it may be difficult to determine whether a claim is true or false, particularly in complex or technical fields such as science and medicine. This could lead to overly broad restrictions on speech, which could have unintended consequences for legitimate discourse.

Counterargument 3: The line between misinformation and opinion is blurry

Another counterargument is that the line between misinformation and opinion is often blurry. Opinions are subjective and may be based on incomplete or incorrect information, but they are still protected under freedom of speech. Limiting the spread of opinions could also limit the spread of ideas and perspectives, ultimately stifling democratic debate and dissent.

Counterargument 4: The solution is not to censor misinformation, but to promote accurate information

Instead of censoring misinformation, the solution may be to promote accurate information. This could involve investing in fact-checking initiatives, promoting media literacy, and providing resources for individuals to verify claims before sharing them. By empowering individuals with the tools they need to distinguish between accurate and misleading information, we can help combat the spread of misinformation without infringing on freedom of speech.

Counterargument 5: The consequences of spreading misinformation should be weighed against the benefits of free speech

Ultimately, the decision about whether freedom of speech should include the right to spread misinformation requires a careful weighing of the potential consequences against the benefits of free speech. While misinformation can have serious negative consequences, it is also essential for democratic debate and dissent. By striking a balance between protecting free speech and mitigating the harm caused by misinformation, we can promote a more informed and engaged citizenry while preserving the fundamental principles of democracy.

Synthesis (Mistral 7B):

After considering both perspectives, it is clear that freedom of speech should not be an absolute right to spread misinformation. While protecting free speech is crucial for democratic debate and dissent, the potential harm caused by spreading false information cannot be ignored. Therefore, there should be limits on this right, such as laws against

defamation, incitement to hatred or violence, or the dissemination of false information in certain contexts. However, it is also essential to strike a balance between protecting free speech and mitigating the harm caused by misinformation. This can be achieved through promoting accurate information, investing in fact-checking initiatives, promoting media literacy, and providing resources for individuals to verify claims before sharing them. Ultimately, the decision about whether freedom of speech should include the right to spread misinformation requires a careful weighing of the potential consequences against the benefits of free speech, taking into account factors such as the nature of the information, its potential harm or benefit, the specific legal framework, and the broader social and political context.

## References


1. Kishore, P.; Salim, R.; Todd, W.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th ACL, Philadelphia, PA, USA, 6–12 July 2002. Available online: https://www.aclweb.org/anthology/P02-1040 (accessed on 25 May 2025).
2. Lin, C.-Y. *ROUGE: A Package for Automatic Evaluation of Summaries*; Text Summarization Branches Out; ACL: Stroudsburg, PA, USA, 2004. Available online: https://www.aclweb.org/anthology/W04-1013 (accessed on 25 May 2025).
3. Pranav, R.; Zhang, J.; Konstantin, L.; Percy, L. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*; EMNLP: Austin, TX, USA, 2016. Available online: https://arxiv.org/abs/1606.05250 (accessed on 25 May 2025).
4. Jason, W.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.V.; Zhou, D. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2022**, arXiv:2201.11903. Available online: https://arxiv.org/abs/2201.11903 (accessed on 25 May 2025).
5. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv* **2023**, arXiv:2303.17651. Available online: https://arxiv.org/abs/2303.17651 (accessed on 25 May 2025). [CrossRef]
6. Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv* **2023**, arXiv:2303.11366. Available online: https://arxiv.org/abs/2303.11366 (accessed on 25 May 2025). [CrossRef]
7. Park, J.S.; Joseph, O.B.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv* **2023**, arXiv:2304.03442. Available online: https://arxiv.org/abs/2304.03442 (accessed on 25 May 2025). [CrossRef]
8. Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J.B.; Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv* **2023**, arXiv:2305.14325. Available online: https://arxiv.org/abs/2305.14325 (accessed on 25 May 2025). [CrossRef]
9. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.L.; Cao, Y.; Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv* **2023**, arXiv:2305.10601. Available online: https://arxiv.org/abs/2305.10601 (accessed on 25 May 2025). [CrossRef]
10. Li, Y.; Yang, C.; Ettinger, A. When Hindsight is Not 20/20: Testing Limits on Reflective Thinking in Large Language Models. *arXiv* **2024**, arXiv:2404.09129. Available online: https://arxiv.org/abs/2404.09129 (accessed on 25 May 2025). [CrossRef]
11. Ioan, S.; Pecheanu, E.; Cocu, A.; Istrate, A.; Anghel, C.; Iacobescu, P. Non-Intrusive Monitoring and Detection of Mobility Loss in Older Adults Using Binary Sensors. *Sensors* **2025**, *25*, 2755. [CrossRef]
12. Zhu, K.; Zhao, Q.; Chen, H.; Wang, J.; Xie, X. PromptBench: A Unified Library for Evaluation of Large Language Models. *arXiv* **2023**, arXiv:2312.07910. Available online: https://arxiv.org/abs/2312.07910 (accessed on 25 May 2025).
13. Bhat, V. *RubricEval: A Scalable Human-LLM Evaluation Framework for Open-Ended Tasks*; Stanford CS224N Final Report; Stanford University: Stanford, CA, USA, 2023. Available online: https://web.stanford.edu/class/cs224n/final-reports/256846781.pdf (accessed on 25 May 2025).
14. Hashemi, H.; Eisner, J.; Rosset, C.; Van Durme, B.; Kedzie, C. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, 11–16 August 2024; pp. 13806–13834. Available online: https://aclanthology.org/2024.acl-long.745 (accessed on 25 May 2025).
15. Schwartz, S.H. An Overview of the Schwartz Theory of Basic Values. *Online Read. Psychol. Cult.* **2012**, *2*, 11. [CrossRef]
16. Zheng, C.; Liu, Z.; Xie, E.; Li, Z.; Li, Y. Progressive-Hint Prompting Improves Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2304.09797. Available online: https://arxiv.org/abs/2304.09797 (accessed on 25 May 2025). [CrossRef]
17. Webb, T.; Mondal, S.S.; Momennejad, I. Improving Planning with Large Language Models: A Modular Agentic Architecture. *arXiv* **2023**, arXiv:2310.00194. Available online: https://arxiv.org/abs/2310.00194 (accessed on 25 May 2025).

18. Gu, Y.; Tafjord, O.; Kuehl, B.; Haddad, D.; Dodge, J.; Hajishirzi, H. OLMES: A Standard for Language Model Evaluations. *arXiv* **2024**, arXiv:2406.08446. Available online: https://arxiv.org/abs/2406.08446 (accessed on 25 May 2025). [CrossRef]
19. Structured Outputs: Making LLMs Reliable for Document Processing. Generative AI Newsroom. 2024. Available online: https://generative-ai-newsroom.com/structured-outputs-making-llms-reliable-for-document-processing-c3b6b2baed36 (accessed on 25 May 2025).
20. Kuchnik, M.; Smith, V.; Amvrosiadis, G. Validating Large Language Models with ReLM. *arXiv* **2022**, arXiv:2211.15458. Available online: https://arxiv.org/abs/2211.15458 (accessed on 25 May 2025).
21. Scherrer, N.; Shi, C.; Feder, A.; Blei, D. Evaluating the Moral Beliefs Encoded in LLMs. In Proceedings of the NeurIPS 2023, New Orleans, Louisiana, 10–16 December 2023. Available online: https://openreview.net/forum?id=O06z2G18me (accessed on 25 May 2025).
22. Sidana, N. Running Models with Ollama Step-by-Step. Medium. 2024. Available online: https://medium.com/@nsidana123/running-models-with-ollama-step-by-step-b3bdbfd91e8e (accessed on 25 May 2025).
23. Prompt Engineering Guide. Gemma 7B. 2025. Available online: https://www.promptingguide.ai/models/gemma (accessed on 25 May 2025).
24. Google AI for Developers. Gemma Formatting and System Instructions. 2024. Available online: https://ai.google.dev/gemma/docs/formatting (accessed on 25 May 2025).
25. Hugging Face. Cognitivecomputations/Dolphin-2.8-Mistral-7b-v02. 2024. Available online: https://huggingface.co/cognitivecomputations/dolphin-2.8-mistral-7b-v02 (accessed on 25 May 2025).
26. Hugging Face. HuggingFaceH4/zephyr-7b-beta. 2024. Available online: https://huggingface.co/HuggingFaceH4/zephyr-7b-beta (accessed on 25 May 2025).
27. Meta AI. Introducing Llama 3.1: Our Most Capable Models to Date. 2024. Available online: https://ai.meta.com/blog/meta-llama-3-1/ (accessed on 25 May 2025).
28. Asprino, L.; De Giorgis, S.; Gangemi, A.; Bulla, L.; Marinucci, L.; Mongiovì, M. Uncovering Values: Detecting Latent Moral Content from Natural Language with Explainable and Non-Trained Methods. In Proceedings of the Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Dublin, Ireland, 27 May 2022; pp. 33–41. Available online: https://aclanthology.org/2022.deelio-1.4.pdf (accessed on 25 May 2025).
29. Chen, G.; Dong, S.; Shu, Y.; Zhang, G.; Sesay, J.; Karlsson, B.F.; Fu, J.; Shi, Y. AutoAgents: A Framework for Automatic Agent Generation. *arXiv* **2023**, arXiv:2309.17288. Available online: https://arxiv.org/abs/2309.17288 (accessed on 25 May 2025).
30. Liu, X.; Chen, J.; Li, C.; Song, X.; Wang, Y. CAMEL: Communicative Agents for Mind Exploration of Large-Scale Language Model Society. *arXiv* **2023**, arXiv:2303.17760. Available online: https://arxiv.org/abs/2303.17760 (accessed on 25 May 2025).
31. Ni, H. Extracting Insights from Unstructured Data with LLMs & Neo4j. Medium, 15 January 2025. Available online: https://watchsound.medium.com/extracting-insights-from-unstructured-data-with-llms-neo4j-914b1f193c64 (accessed on 25 May 2025).
32. Sahoo, P.; Singh, A.K.; Saha, S.; Jain, V.; Mondal, S.; Chadha, A. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv* **2024**, arXiv:2402.07927. Available online: https://arxiv.org/abs/2402.07927 (accessed on 25 May 2025). [CrossRef]
33. Lum, K.; Anthis, J.R.; Robinson, K.; Nagpal, C.; Alexander, D.A. Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation. *arXiv* **2024**, arXiv:2402.12649. Available online: https://arxiv.org/abs/2402.12649 (accessed on 25 May 2025). [CrossRef]
34. Microsoft Learn. Evaluation and Monitoring Metrics for Generative AI. Available online: https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/evaluation-metrics-built-in (accessed on 25 May 2025).
35. Confident AI. G-Eval: The Definitive Guide. Available online: https://www.confident-ai.com/blog/g-eval-the-definitive-guide (accessed on 25 May 2025).
36. Padmakumar, V.; Yueh-Han, C.; Pan, J.; Chen, V.; He, H. Beyond Memorization: Mapping the Originality-Quality Frontier of Language Models. *arXiv* **2025**, arXiv:2504.09389. Available online: https://arxiv.org/abs/2504.09389 (accessed on 25 May 2025). [CrossRef]
37. Cohn, A.G.; Hernandez-Orallo, J. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs. *arXiv* **2023**, arXiv:2304.11164. Available online: https://arxiv.org/abs/2304.11164 (accessed on 25 May 2025). [CrossRef]
38. Pathak, A.; Gandhi, R.; Uttam, V.; Devansh; Nakka, Y.; Jindal, A.R.; Ghosh, P.; Ramamoorthy, A.; Verma, S.; Mittal, A.; et al. Rubric Is All You Need: Enhancing LLM-based Code Evaluation With Question-Specific Rubrics. *arXiv* **2025**, arXiv:2503.23989. Available online: https://arxiv.org/abs/2503.23989 (accessed on 25 May 2025).
39. Rebmann, A.; Schmidt, F.D.; Glavaš, G.; van der Aa, H. On the Potential of Large Language Models to Solve Semantics-Aware Process Mining Tasks. *arXiv* **2025**, arXiv:2504.21074. Available online: https://arxiv.org/abs/2504.21074 (accessed on 25 May 2025). [CrossRef]

40. Huang, X.; Ruan, W.; Huang, W.; Jin, G.; Dong, Y. A Survey of Safety and Trustworthiness of Large Language Models Through the Lens of Verification and Validation. *Artif. Intell. Rev.* **2024**, *57*, 175. [CrossRef]
41. Alva-Manchego, F.; Scarton, C.; Specia, L. The (Un) Suitability of Automatic Evaluation Metrics for Text Simplification. *Comput. Linguist.* **2021**, *47*, 861–889. Available online: https://aclanthology.org/2021.cl-4.28/ (accessed on 25 May 2025). [CrossRef]
42. Iacobescu, P.; Marina, V.; Anghel, C.; Anghele, A.-D. Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities. *J. Cardiovasc. Dev. Dis.* **2024**, *11*, 396. [CrossRef]
43. Anghele, A.-D.; Marina, V.; Dragomir, L.; Moscu, C.A.; Anghele, M.; Anghel, C. Predicting Deep Venous Thrombosis Using Artificial Intelligence: A Clinical Data Approach. *Bioengineering* **2024**, *11*, 1067. [CrossRef]
44. Reddy, P. A Framework of Rhetorical Moves Designed to Scaffold the Development of Research Proposals. *Int. J. Dr. Stud.* **2023**, *18*, 77–97. Available online: https://ijds.org/Volume18/IJDSv18p077-097Reddy8855.pdf (accessed on 25 May 2025). [CrossRef]
45. Vengal, T. LLMs: A Review of Their Capabilities, Limitations and Evaluation. LinkedIn. 2023. Available online: https://www.linkedin.com/pulse/llms-review-capabilities-limitations-evaluation-thomas-vengal (accessed on 25 May 2025).


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.