



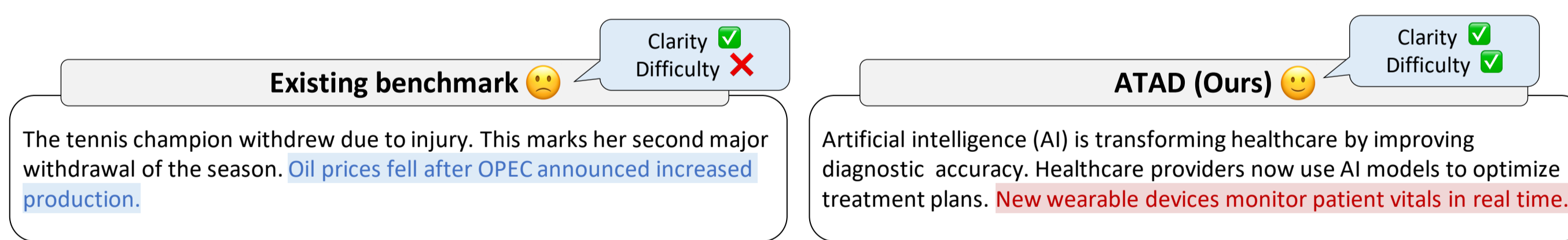
INTRODUCTION

Limitations of Static Benchmarks

- Performance Saturation:** Frontier LLMs surpass human baselines on static benchmarks, **losing their discriminative power.**
- Data Contamination:** Large-scale pretraining inadvertently **memorize test data, inflating leaderboard scores.**
- Format Overfitting:** Tuning to dataset details creates **feedback loops**, not genuine reasoning improvements.
- Short Lifespan:** Once a benchmark is "solved," creating a new one requires **significant time and resources**, limiting **sustainable evaluation** (rapid benchmark saturation).

Motivation

- Need for Dynamic Evaluation:** Continuous evolution is essential to mitigate contamination and overfitting.
- Clarity vs. Difficulty Trade-off:** Increasing difficulty often **sacrifices clarity**, while clear problems tend to be **overly simple or trivial.**
- Why Text Anomaly? (Stress Test):** Unlike math or code, language lacks formal rules. It serves as a stringent stress test for scaling difficulty without ambiguity, revealing subtle cross-sentence flaws.



Task Design Examples (7 Types Total)

T1. Sentence Context Anomaly

Identify the sentence that semantically or topically deviates from the rest within a paragraph.

The dynamics of social change are deeply influenced by technological advancements. Globalization has led to increased interconnectedness. Climate change threatens crop yields, rising concerns about food security. Digital currencies are changing how people interact economically. The theory of relativity explains time dilation, which helps us understand how society views time.

Correct Answer: 5

Context misalignment between society and physics.

T5. Referential Ambiguity

Detect sentences with unclear or conflicting pronoun references.

Jane and Mr. Bennet frequently walked the countryside, talking about their family matters. She found solace in her library, especially when tension rose due to marriage discussions. Lady Catherine, Bennet's aunt, expressed concerns about her family's social standing. After Jane met Catherine, that changed relations between the families. Since then, their interactions became noticeably more cordial and respectful.

Correct Answer: 3

'Her' can be Catherine or Jane.

T7. Tone / Style Violation

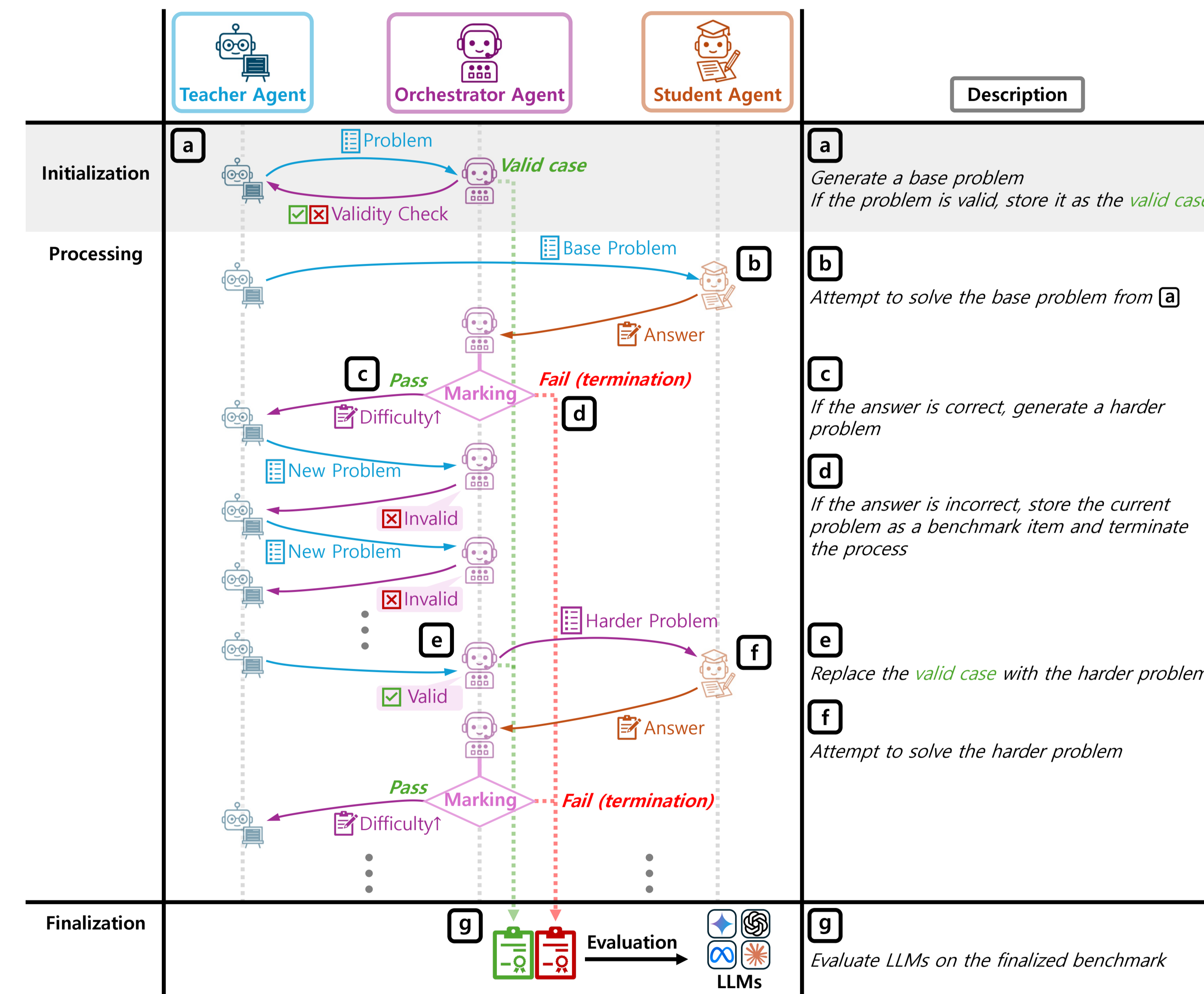
Identify stylistic or tonal shifts that disrupt the overall narrative tone.

In Moby-Dick, Melville explores obsession, revenge, and nature's power through a deep psychological lens. His symbols and complex characters raise big questions about existence and knowledge. The story blends adventure with philosophy, showing the range of 19th-century American lit. Detailed whale science adds a cool layer of realism and depth. The characters' emotional chaos? Pure literary beast mode.

Correct Answer: 5

Informal and conversational expression.

ATAD: Benchmark Protocol Design and Operation



- Three-Agent System:** Teacher generates problems, Orchestrator strictly validates quality/fairness, and Student attempts to solve them, driving the difficulty calibration.
- Difficulty Scaling via T-S Competition:** Teacher generates harder variants directly targeting the Student's reasoning weaknesses.
- Orchestrator-Regulated Difficulty Control:** Prevents uncontrolled or adversarial difficulty escalation by ensuring logical coherence and clarity.
- Failure-Driven Sample Finalization:** Problems are finalized only when the Student fails, anchoring benchmark difficulty to actual model limitations.
- Dynamic Difficulty Localization:** Adjusts difficulty at the instance level, enabling precise probing of model-specific blind spots.
- Cross-Agent Instantiability:** Modular design supports different model pairings) to track evolution over time.

Results & Analysis

Main Results & Difficulty Scaling

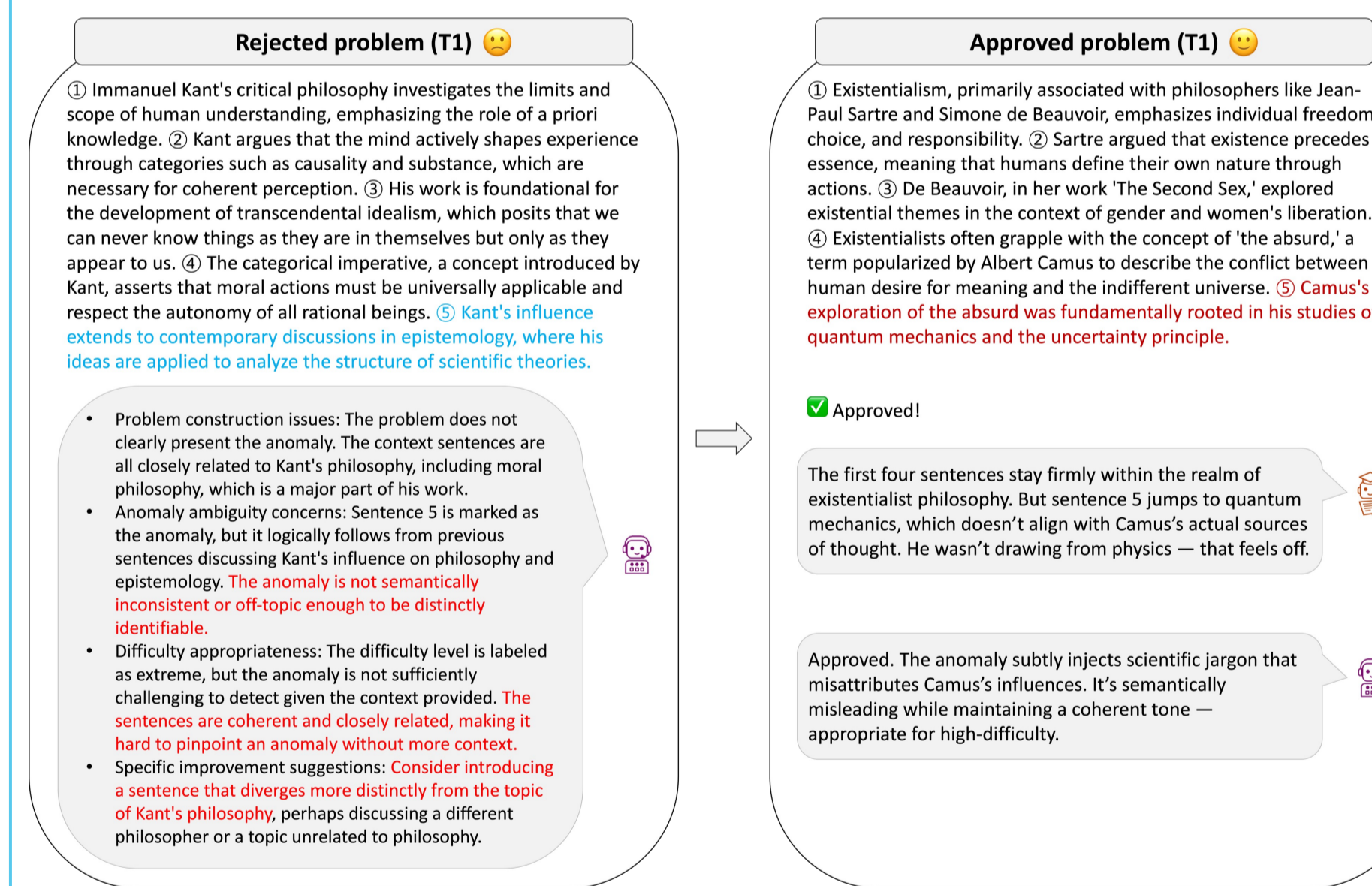
Evaluation Model	T1	T2	T3	T4	T5	T6	T7	Avg.
GPT-3.5-Turbo	59.00	16.00	66.75	48.50	55.75	51.75	81.50	54.18
GPT-4o-mini	57.25	17.00	62.50	54.00	52.50	58.75	83.00	55.00
GPT-4o	62.00	21.25	68.25	53.25	49.25	56.75	81.00	55.96
GPT-o4-mini	63.25	30.25	68.50	53.00	47.25	57.25	80.00	57.07
Gemini-1.5-Flash	6.00	11.25	62.00	48.75	17.50	10.75	21.00	25.32
Gemini-2.0-Flash-Lite	64.00	10.75	63.50	52.25	62.75	62.00	86.25	57.36
Gemini-2.0-Flash	65.25	25.00	63.00	58.25	51.00	62.00	88.00	58.93
Claude-3-Haiku	63.75	12.00	61.00	51.75	53.50	60.00	72.75	53.54
Claude-3.5-Haiku	19.75	55.00	7.25	5.00	5.00	8.50	35.50	19.50
Claude-3.5-Sonnet	65.75	31.75	65.00	59.50	53.50	57.50	86.75	59.96
LLaMA-3.1-8B	39.50	12.75	35.50	24.50	53.00	38.75	68.75	38.96
LLaMA-3.3-70B	60.75	27.75	63.25	60.00	52.25	57.75	84.25	58.00

Avg. Acc. (%)	Benchmark Generator Family			
	GPT-4o	Gemini-2.0	Claude-3.5	LLaMA-3.3
Base Problems	82.19	73.07	76.36	78.26
Final Problems	60.93	39.20	52.29	45.51
Acc. Drop	↓ 21.26%p	↓ 33.87%p	↓ 24.07%p	↓ 32.75%p

Effective Scaling: Accuracy drops significantly from base to final, exposing true reasoning weaknesses.

Diverse Reasoning Tasks: Across 7 distinct anomaly types (T1–T7), no single model dominates. SOTA models exhibit varying strengths and unique blind spots depending on the task.

Robustness through Orchestrator-Led Quality Control



Rejected problem (T1) (Clarity \times , Difficulty \times): Immanuel Kant's critical philosophy investigates the limits and scope of human understanding, emphasizing the role of a priori knowledge. Kant argues that the mind actively shapes experience through categories such as causality and substance, which are necessary for coherent perception. His work is foundational for the development of transcendental idealism, which posits that we can never know things as they are in themselves but only as they appear to us. The categorical imperative, a concept introduced by Kant, asserts that moral actions must be universally applicable and respect the autonomy of all rational beings. Kant's influence extends to contemporary discussions in epistemology, where his ideas are applied to analyze the structure of scientific theories.

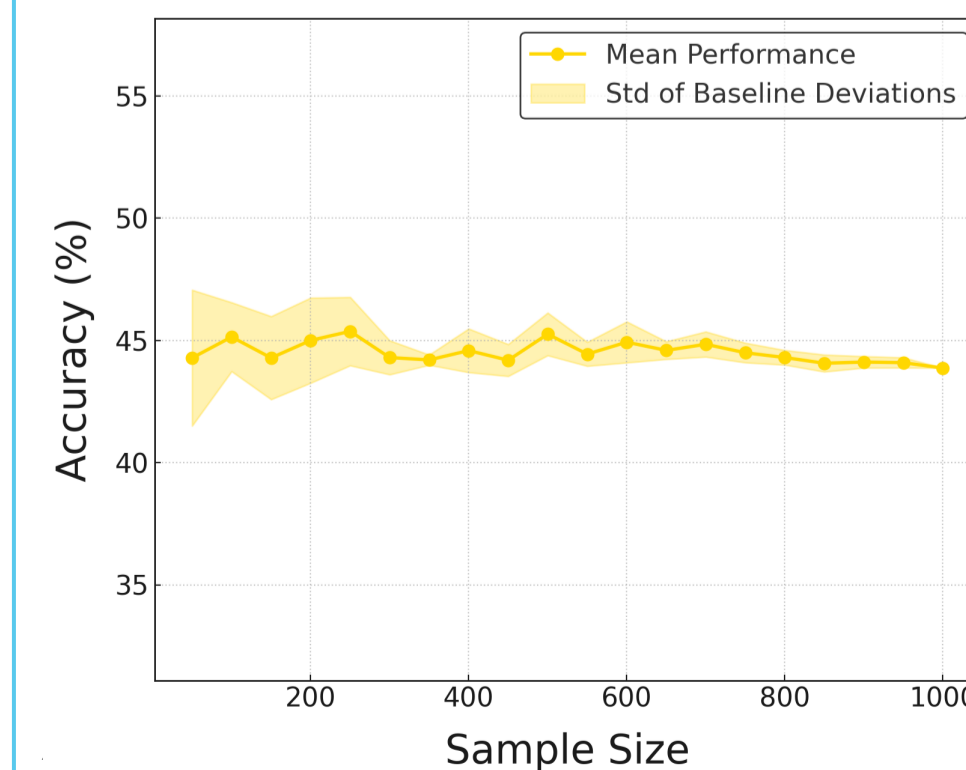
Approved problem (T1) (Clarity \checkmark , Difficulty \checkmark): Existentialism, primarily associated with philosophers like Jean-Paul Sartre and Simone de Beauvoir, emphasizes individual freedom, choice, and responsibility. Sartre argued that existence precedes essence, meaning that humans define their own nature through actions. De Beauvoir, in her work "The Second Sex," explored existential themes in the context of gender and women's liberation. Existentialists often grapple with the concept of the absurd, a term popularized by Albert Camus to describe the conflict between human desire for meaning and the indifferent universe. Camus's exploration of the absurd was fundamentally rooted in his studies of quantum mechanics and the uncertainty principle.

	w/o Orch.	w/ Orch.
[Evaluator: GPT-4o]		
Performance (%)	68.29	72.43
Validity (1–5)	4.30	4.85
Coherence (1–5)	3.71	4.74
Fairness (1–5)	3.20	4.65
Approval Rate (%)	38.14	87.14

	w/o Orch.	w/ Orch.
[Evaluator: Claude-3.5-Sonnet]		
Performance (%)	65.00	72.86
Validity (1–5)	4.61	4.92
Coherence (1–5)	4.11	4.69
Fairness (1–5)	3.41	4.42
Approval Rate (%)	55.57	90.43

Analysis & Future-Proofing

Average Performance across Tasks by Sample Size



Stability: Consistent evaluation results with low variance across sample sizes.

Evaluation Model	GPT-4o	
	Base	Final
GPT-o3-mini	93.71	72.14
GPT-o4-mini	91.86	72.43
GPT-4o	94.29	72.43
GPT-4o-mini	93.00	68.29

Future-Proof: Sustains discriminative power even for simulated future models.

Future Work

- Game-Theoretic Validation:** Quantifying agent contributions using Shapley values.
- Meta-Agent Ecosystems:** Automated discovery of superior teachers and validators.
- External Knowledge:** Integrating RAG for domain-specific anomalies.