MLCU: A Dataset for Evaluating Domain-Specific Language Understanding in LLMs

Anonymous ACL submission

Abstract

Modern Natural Language Processing systems are typically trained on large general purpose corpora where domain specific language is often underrepresented and relegated secondary to fine-tuning. This omission limits the effectiveness of language models in critical applications, such as for military specific functions. This paper presents the development and analysis of MLCU, the Military Language Comprehension and Understanding dataset, which is a novel comprehensive domain specific language dataset tailored for military speech detection. We introduce a comprehensive curated lexicon and example phrases in English drawn from authentic and representative military communications for the basis of benchmarking and annotation of contextual military term disambiguation. Evaluation of our dataset on GPT-4o and and Gemini 2.0 Flash yielded accuracies of 83.6% and 86.2% respectively, indicating that state of the art LLMs have difficulties with disambiguating language when used in a nonstandard manner such as military language. This work highlights the limitations of existing models and the room for improvement of domain specific language in NLP.

1 Introduction

Persistent challenges remain with LLMs' ability to disambiguate terms whose meaning is highly dependent on context especially in domains where polysemous or homonymous expressions are common (Haber and Poesio, 2024). It is especially critical in military language, where many terms have situation specific meanings that diverge significantly from their general-purpose interpretations.

In such domains, successful language understanding depends on the model's ability to resolve the meaning of the term based on contextual clues - similar in some respects to the use of figurative or coded language, such as metaphor or dog whistles (Mendelsohn et al., 2023). In the case of specialized terms used in the armed forces, there exists a large glossary of terms used in everyday language that hold distinct meanings in military-related situations. For example, the word "bird" may describe a helicopter in a military context or refer to a flying living creature in regular language. These challenges extend to other specialized domains requiring high operational accuracy.

Despite progress with tools such as military chatbots that use intent detection and response generation for task-oriented training and dialogue in operational scenarios (Chuang and Cheng, 2022), NLP systems still face significant gaps in interpreting military terminology. To address this, we introduce MLCU (Military Language Comprehension and Understanding), a novel dataset designed to evaluate LLMs on contextual term disambiguation in authentic military communication settings. Our main contributions are as follows:

- A curated dataset of military-specific terms and contextual examples for specialized language understanding.
- An empirical analysis of GPT-4o and Gemini 2.0 Flash yielding accuracies of 83.6% and 86.2% respectively demonstrating challenges in disambiguating military terminology.
- A qualitative analysis identifying distinct reasoning and interpretation errors.

2 Data Collection

2.1 Term Selection

After conducting extensive research using various online sources, we identified a wide range of sources of reputable military terminology[1]. From

[1] Sources include US Department of Defense Dictionary of Military and Associated Terms, NATO Glossary of Abbreviations used in NATO, US Army Military Intelligence Service British Military Terminology, US Air Force Fact Sheet, The Economist The A to Z of military terms.

this collection, we carefully selected 50 relevant terms in English (Appendix A) that are potentially ambiguous: those that could be mistaken for civilian language in everyday contexts. Each of the terms has an alternative military definition that is different from the standard definition; for instance, "COIN" refers to counter-insurgency. The MLCU Terms includes the term itself, the military definition, and the source index (see Appendix B for an example). This deliberate curation of terms ensures that our analysis focuses on terminology that may pose challenges in interpretation, making it more useful for natural language processing applications with contextual disambiguation.

2.2 Contextual Example Selection

To compile contextual examples for the selected terms, we collected a diverse range of real-world textual examples from over 350 sources. These include film transcripts (e.g., Black Hawk Down, Avatar), literary works (e.g., Snakes in the Eagle's Nest), articles from major news outlets (e.g., AP, CBS), and official publications from military organizations (e.g., the U.S. Air Force, NATO). Our sources span both civilian and military domains, covering topics such as strategy, history, culture, science, and policy. By relying exclusively on naturally occurring language, we ensure that all examples are contextually grounded and verifiable.

For each target term, we extracted the sentence in which the term appears, along with the two surrounding sentences from the same paragraph, in order to preserve local discourse context. We intentionally retained longer excerpts, as shorter military communications are often ambiguous or compressed in meaning, making them more difficult to interpret in isolation (Lebanoff et al., 2021).

Each example is annotated with a binary `Defense Flag` indicating whether the term is being used in a military sense (`1`) or not (`0`). Additional metadata includes the original term, its military definition, a numeric source index (linking to the dataset's `Sources` sheet), and optional clarifying notes (see Appendix C for format example).

After collecting and annotating a comprehensive set of examples, we constructed a balanced dataset for evaluation. For each of the 50 selected terms, we included 15 examples where the term is used with its military meaning and 15 where it is not. Therefore, MLCU has a total of 1,500 examples: 750 labeled with `Defense Flag = 1`, and 750 labeled with `Defense Flag = 0`.

3 Dataset Processing

Our core task defined by our dataset is *contextual military term disambiguation* - determining whether a given term in a sentence is used with its military specific meaning. To determine MLCU's effectiveness, we utilized OpenAI's ChatGPT-4o and Google's Gemini 2.0 Flash.

We employed a zero-shot prompting setup, using the following template for each example:

> *In the example "{example}", does the word "{term}" have a military meaning? Answer yes or no.*

In this prompt, {*example*} refers to the sample example found in a real piece of writing (a target sentence plus two adjacent sentences), and {*term*} refers to the ambiguous term being tested. We recorded responses from the model API in natural language and normalized to fit our binary `Defense Flag` label: `1` for "yes" (military meaning), and `0` for "no" (non-military meaning). The models always answered yes or no (see Appendix D for details) using a temperature of 0. Predicted labels were then compared to the gold-standard `Defense Flag` annotations in the dataset. These outcomes form the basis for downstream evaluation metrics such as accuracy, precision, recall, and F1 score, which we report in the following section.

4 Evaluation

Table 1 summarizes the results across all 1,500 examples. GPT-4o achieved an accuracy of 83.6%, while Gemini reached 86.2%. These correspond to 246 and 207 total errors, respectively.

In both cases, the models exhibited a consistent pattern: more false negatives than false positives. This means they were more likely to predict a term as *non-military* when it was actually used in a military sense, risking real world misunderstanding. This tendency may suggest a conservative bias in zero-shot disambiguation or reflect a broader gap in model exposure to domain-specific usage, particularly in military contexts.

To contextualize MLCU's difficulty relative to existing evaluation benchmarks, we compare results on our dataset to MMLU and MMLU-Pro scores from the LLMDB-2025 benchmark suite (LLMDB.com, 2025). As shown in Table 2, MLCU

[2] Gemini 2.0 Flash does not have a published standard MMLU score; we use its Global-MMLU-Lite score from LLMDB as a proxy.

Model	Accuracy	Precision	Recall	F1 Score	TP	FP	FN	TN
GPT 4o	83.6%	85.0%	81.6%	83.3%	612	108	138	642
Gemini 2.0 Flash	86.2%	88.5%	83.2%	85.8%	624	81	126	669

Table 1: Evaluation metrics comparing GPT4o and Gemini models on the balanced Defense Flag dataset.

Model	MMLU (Standard)	MMLU-Pro	MLCU
GPT-4o	88.7%	74.7%	83.6%
Gemini 2.0 Flash	∼83.4%[2]	76.4%	86.2%

Table 2: Benchmark performance on MMLU, MMLU-Pro, and our custom dataset (MLCU). All scores are reported from LLMDB.com (2025), except MLCU which reflects results from our evaluation.

elicits model performance in a comparable range to these established benchmarks. This shows that MLCU is well-calibrated in difficulty and offers a valuable diagnostic for evaluating LLMs on contextual understanding in specialized domains.

5 Analysis

To understand the incorrect responses, we analyzed these results. When we manually re-prompted the LLMs with the same incorrect entries, we observed that some responses had changed. This variability may be attributable to the dynamic nature of LLMs, which can evolve over time (Chen et al., 2024). For example, our original MLCU benchmark was run via the API in May 2025, while the manual re-prompting was conducted in July 2025 with accuracy results of 91.87% and 89.40% for GPT-4o and Gemini respectively (see Table 3) - highlighting how changes in model behavior or training over short time frames can affect reproducibility.

Metric	GPT-4o	Gemini 2.0
Incorrect from API	246	207
Incorrect from Manual	122	159
Manual Accuracy[3]	91.87%	89.40%
Manual Error Rate (%)	8.13%	10.60%

Table 3: Performance of Manual Results

In many instances, we noticed our errors were related to the fact that the LLM would classify mostly based on context rather than the actual meaning of the word. For example, for the term "falcon" in the sentence *"This hybrid falcon performs at halftime of our home football games. Ace is a nickname short for his full name, Achilles, but fits with the U.S. Air Force Academy's flying culture."*, GPT-4o answered: *"Yes, the term "falcon" in the sentence has a military meaning. While it refers to an actual bird, it is also the mascot and symbol of the U.S. Air Force Academy, representing the Academy's heritage, values, and flying culture. In this context, "falcon" carries symbolic military significance tied to Air Force identity and tradition."*. Despite correctly identifying the falcon as a mascot and a bird, the model's association of the term with military symbolism led to an erroneous classification. This highlights how broader context can override interpretation of a word.

To better understand the types of errors, we identified three general categories of classification mistakes made by the two LLMs tested:

1. **Incorrect reasoning**: The model provides a correct definition of the term but incorrectly classifies its usage. *Example*: The term `bird` was correctly interpreted as a literal chicken, yet GPT-4o labeled it as military due to context (Appendix F).

2. **Incorrect definition**: The model misinterprets the meaning of the term, leading to an incorrect classification. *Example*: `Major Matsutoshi` was misclassified by Gemini as a geographic term instead of a military rank (Appendix G).

3. **Misunderstood**: The model combines the correct definition with local context in a flawed way, leading to misclassifications. *Example*: The term `advance` (military meaning: movement toward an enemy) was misinterpreted by GPT-4o as part of the phrase `advance warning`, resulting in an incorrect military classification (Appendix H).

An analysis of error types reveals distinct patterns in how the two models fail. Among GPT-

[3]This accuracy is after manually rerunning only the incorrectly marked results from the API. We did not rerun correct examples manually.

Error Type	GPT-4o			Gemini 2.0 Flash		
	Count	% of Incorrect	% of Total	Count	% of Incorrect	% of Total
Incorrect reasoning	92	75.41%	6.13%	148	93.08%	9.87%
Incorrect definition	8	6.56%	0.53%	7	4.40%	0.47%
Misunderstood	21	17.21%	1.40%	4	2.52%	0.27%
Total Incorrect	122	100%	8.13%	159	100%	10.60%

Table 4: Comparison of Error Types Between GPT 4o and Gemini 2.0 Flash

Error Type	GPT-4o		Gemini 2.0 Flash	
	Yes Answer	No Answer	Yes Answer	No Answer
Incorrect reasoning	61	31	49	99
Incorrect definition	5	4	2	5
Misunderstood	21	0	4	0
Total	**87**	**35**	**55**	**104**

Table 5: Distribution of Error Types by Initial Response ("Yes"/"No") for GPT-4o and Gemini 2.0 Flash

4o's 122 manual incorrect responses, the majority (75.41%) were due to incorrect reasoning, followed by misunderstood terms (17.21%) and incorrect definitions (6.56%). These correspond to 6.13%, 1.40%, and 0.53% of the total dataset, respectively. Gemini, on the other hand, demonstrated an even stronger skew toward reasoning errors: 148 out of its 159 manual incorrect responses (93.08%) were due to incorrect reasoning, with incorrect definitions and misunderstood terms accounting for only 4.40% and 2.52% of its errors, respectively. As a share of the total dataset, incorrect reasoning in Gemini alone affected 9.87% of examples (Table 4). These findings suggest that both models are predominantly challenged by reasoning-based distinctions. Often, the context of a sentence overpowers the definition of the term.

The two models exhibited an overlap of 74 incorrect predictions, representing approximately 4.93% of the total dataset. This shared error set accounts for 60.66% of GPT-4o's incorrect responses and 46.54% of Gemini's. Notably, 63 of these overlapping errors were not only misclassified by both models but were also assigned the same error type, indicating a substantial convergence in the nature of their failures (see Appendix I). This alignment suggests that certain examples systematically challenge both models, likely due to inherent ambiguities or conceptual complexities in the data . Interestingly, the remaining overlapping cases involved different error types across models, underscoring not only the difficulty of the task but also the distinct reasoning pathways and decision boundaries employed by each model. These discrepancies highlight the value of comparative error analysis in uncovering nuanced differences in model behavior.

Additionally, we observe that when the models make a mistake, there is a slight bias to answer one way or another. In our testing, we found that GPT-4o is more likely to answer "Yes" when making a mistake, while Gemini is more likely to answer "No" (see Table 5). This could show that GPT-4o is more likely to be overconfident and Gemini conservative.

6 Conclusion

In this work, we introduced MLCU, a novel dataset designed to evaluate large language models on context-sensitive understanding of ambiguous terms. Our results show that current LLMs often fail to disambiguate terms whose meanings vary by domain or usage, indicating a gap in their contextual reasoning abilities. Unlike traditional benchmarks that focus on general linguistic capability, MLCU highlights the need for fine-grained semantic understanding without relying on task-specific fine-tuning. We believe that ambiguous terms—especially those shared across domains—represent a critical challenge for general-purpose LLMs. Our dataset provides a foundation for systematic evaluation of this phenomenon and can support future work in improving model understanding, training data selection, and robustness to semantic ambiguity.

7 Limitations

We note that this work was conducted by a student team with limited access to research funding and high-performance computational resources. These constraints may have influenced the scale of our experiments and the choice of models to two, GPT 4o and Gemini 2.0 Flash. Additionally, we notice that sometimes LLMs will be able to correctly identify the source of the example either from in its knowledge base or internet search, which loses some of the ambiguity of that example.

This work was conducted by a student team with limited access to research funding and high-performance computational resources. These constraints may have influenced the scale of our experiments and the choice of models to two, GPT-4o and Gemini 2.0 Flash, and this may not reflect the full range of capabilities across other LLMs.

Additionally, as we retrieved our examples from sources found on the internet, some examples in our dataset may be partially known to the models through their pretraining data or retrieval mechanisms, which can reduce the intended ambiguity and artificially inflate performance.

Our dataset focuses exclusively on English and covers a specific subset of ambiguous military-related terms, limiting its generalization to other languages or domains.

The examples were hand-selected and manually annotated by our team, including classification of error types. While this allowed for careful curation, it introduces the potential for human bias and inconsistency in the dataset and error analysis.

We used a fixed prompt structure for all model evaluations; thus, performance may vary with alternative prompting strategies. Finally, our primary reliance on accuracy and related metrics may not fully capture the nuanced understanding or partial correctness in model predictions.

We understand there are inherent risks with sensitive domain types such as military language and that misinterpretations of the results of this project may create dysfunctional or dangerous real world consequences. Additionally, over-reliance on this or any benchmark may produce adverse results.

References

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2).

Hsiu-Min Chuang and Ding-Wei Cheng. 2022. Conversational ai over military scenarios using intent detection and response generation. *Applied Sciences*, 12(5).

Janosch Haber and Massimo Poesio. 2024. Polysemy—Evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Logan Lebanoff, Charles Newton, Victor Hung, Beth Atkinson, John Killilea, and Fei Liu. 2021. Semantic parsing of brief and multi-intent natural language utterances. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 255–262, Kyiv, Ukraine. Association for Computational Linguistics.

LLMDB.com. 2025. Llm leaderboard benchmarks.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

A List of Terms

1. `Action` — An engagement or battle, usually one on a small scale.

2. `Advance` - The progress of a command toward the enemy. To move forward. To make progress in the direction of the enemy.

3. `Approach` - A route by which a place or position can be approached by an attacking force. The route leading to anything, as a bridge.

4. `apron` - A defined area on an airfield intended to accommodate aircraft for purposes of loading or unloading passengers or cargo, refueling, parking, or maintenance. (JP 3-34)

5. `backfill` - Reserve Component units and individuals recalled to replace deploying active units and/or individuals in the continental United States and outside the continental United States. See also Reserve Component.

6. `barrier` - A coordinated series of natural or man-made obstacles designed or employed to channel, direct, restrict, delay, or stop the movement of an opposing force and to impose additional losses in personnel, time, and equipment on the opposing force.

7. `base` - 1. A locality from which operations are projected or supported. 2. An area or locality

containing installations which provide logistic or other support. 3. Home airfield or home carrier.

8. `beach party` - The Navy component of the landing force support party under the tactical control of the landing force support party commander. See also beachmaster unit; shore party.

9. `Bird` - Helicopter

10. `Black Hawk` - UH-60 Black Hawk

11. `campaign` - A series of related major operations aimed at achieving strategic and operational objectives within a given time and space.

12. `COIN` - counter-insurgency

13. `Column` - A formation in which the elements are Same or line ahead. placed one behind another. A march column comprises all elements of a command marching on one route under the control of one commander, ineluding such forward, flank, and rear security forces as may be employed.

14. `Company` - A military unit made up of at least three platoons and generally led by a captain or a major. The size and structure of a company can differ between armies, but they tend to contain no more than 250 soldiers.

15. `component` - 1. One of the subordinate organizations that constitute a joint force. 2. In logistics, a part or combination of parts having a specific function, which can be installed or replaced only as an entity.

16. `Concentration` - a. An assembly of troops in a particular locality, on mobilization, for training, attack, or defense. b. The amount of toxic chemical vapor in a given volume of air at any particular time and place. c. A volume of fire placed on an area within a limited time.

17. `constellation` - A system consisting of a number of like satellites acting in concert to perform a specific mission.

18. `dispersal` - Relocation of forces for the purpose of increasing survivability.

19. `display` - In military deception, a static portrayal of an activity, force, or equipment intended to deceive the adversary's visual observation.

20. `download` - An operation that removes airborne weapons or stores from an aircraft.

21. `Eagle` - F-15 Eagle

22. `element` - An organization formed around a specific function within a designated directorate of a joint force commander's headquarters.

23. `employment` - The strategic, operational, or tactical use of forces.

24. `engagement` - 1. In air defense, an attack with guns or air-to-air missiles by an interceptor aircraft, or the launch of an air defense missile by air defense artillery and the missile's subsequent travel to intercept. 2. A tactical conflict, usually between opposing lower echelons maneuver forces.

25. `ERA` - explosive reactive armour

26. `exercise` - A military maneuver or simulated wartime operation involving planning, preparation, and execution that is carried out for the purpose of training and evaluation.

27. `Falcon` - F-16 Fighting Faclon

28. `Fifty` - An M2 .50 caliber machine gun.

29. `footprint` - 1. The area on the surface of the earth within a satellite's transmitter or sensor field of view. 2. The amount of personnel, spares, resources, and capabilities physically present and occupying space at a deployed location.

30. `Front` - The direction of the enemy. The line of contact of two opposing forces. The space occupied by an element, measured from one flank to the opposite flank.

31. `Galaxy` - C-5 Galaxy

32. `HE` - High Explosive

33. `Hercules` - C-130 Hercules

34. `KIA` - Killed in Action

35. `lead` - In intelligence usage, a person with potential for exploitation, warranting additional assessment, contact, and/or development.

36. `Liberty` - MC-12W Liberty

37. `Major` - A rank of commissioned officer in the army, above a captain but below a lieutenant colonel. Major is also an officer rank in America's air force (equivalent to its level in the army).

38. `master` - The commander of a United States Naval Ship, a commercial ship, or a government-owned general agency agreement ship operated for the Military Sealift Command by a civilian company to transport Department of Defense cargo.

39. `organic` - Assigned to and forming an essential part of a military organization as listed in its table of organization for the Army, Air Force, and Marine Corps, and are assigned to the operating forces for the Navy.

40. `Osprey` - Cv-22 Osprey

41. `Patriot` - A point and limited area defense surface-to-air missile system capable of intercepting aircraft and theater missiles, including short-, medium-, and intermediate range ballistic missiles in the terminal phase.

42. `Pegasus` - KC-46A Pegasus

43. `Predator` - MQ-1B Predator

44. `Raptor` - F-22 Raptor

45. `Sentry` - E-3 Sentry

46. `Spirit` - B-2 Spirit

47. `staging` - Assembling, holding, and organizing arriving personnel, equipment, and sustaining materiel in preparation for onward movement.

48. `strike` - An attack to damage or destroy an objective or a capability

49. `theater` - The geographical area for which a commander of a geographic combatant command has been assigned responsibility

50. `TOW` - tube-launched optically-tracked wire-guided (missle)

B Example Term

Term	Definition	Source
`Bird`	Helicopter	4

C Example Entry

Term	Example	Defense Flag	Source	Notes
`bird`	Super Six one is down! We got a bird down in the city!	1	50	

D Term Processing

The models were prompted using the API and always generated a yes or no response using temperature setting of 0. All examples were processed in full without truncation and no token-length issues were encountered. Due to API rate limits, the benchmark dataset was run in small batches over several days. Any results that timed out or did not produce full results were rerun again to get a full output.

E Flag Labels

For the defense flag predictions, we defined the following outcome categories:

- **True Positive (TP)**: Model predicts 1 and ground truth is 1.
- **False Positive (FP)**: Model predicts 1 and ground truth is 0.
- **False Negative (FN)**: Model predicts 0 and ground truth is 1.
- **True Negative (TN)**: Model predicts 0 and ground truth is 0.

F Bird Example

Q: *In the example: "Thanksgiving is just around the corner, but before you sit down to enjoy a freshly-cooked turkey, picture this: a chicken carcass flying out of an eight-inch naval gun at more than 700 miles per hour and into a glass canopy like the ones found on fighter jets. As strange as the image seems, it played out more than 1,000 times at Arnold Air*

Force Base, Tennessee, home of the S-3 Bird Impact Range, better known as the 'chicken gun.' First fired 50 years ago this fall, the chicken gun at Arnold has been used to test glass canopies, windshields and other materials for some of the most famous U.S. military aircraft, including the A-10 attack jet, the C-130 transport plane, and the F-15, F-16 and F/A-18 fighter jets."
Does the word `bird` *have a military meaning?*

A (GPT-4o): *Yes, the term* `bird` *in the sentence has a military meaning. In the phrase "S-3 Bird Impact Range,"* `bird` *refers to avian strike testing—specifically, simulating bird impacts on aircraft components like canopies and windshields. This is a standard part of military aviation safety testing, and while* `bird` *literally refers to a chicken in this case, it's used within a military technical context, giving it a military meaning.*

G Major Example

Q: *In the example: "The Japanese then began bringing in coastal defense and anti-aircraft guns, aircraft and midget submarines, turning Kiska into their major bastion in the North Pacific. Concurrently, Major Matsutoshi Hozumi landed the 1,143-man North Sea (Hokkaido) Detachment, consisting of the 301st Independent Infantry Battalion and 301st Engineer Company, at Holtz Bay, Attu Island, 190 miles west of Kiska, during the evening of June 6. Climbing over the difficult snow-covered Holtz Bay–Chichagof Harbor Pass, they reached Attu Village the next morning, Sunday, as the villagers prepared to go to church."*
Does the word `major` *have a military meaning?*

A (Gemini 2.0): *No, in that context, the term "major" does not have a military meaning. It is used as an adjective to describe Kiska as their principal or most important stronghold in the North Pacific.*

A (GPT-4o): *Yes, the term "Major" in this sentence has a military meaning — it refers to Major Matsutoshi Hozumi, indicating his military rank within the armed forces.*

H Advance Example

Q: *In the example: "The 3/26 elements on Hill 881S were a constant thorn in the enemy's side because the men on that most isolated of the Marine outposts could observe all three of the main NVA firing positions—881N, 305, and Co Roc. When rockets lifted off of 881N or the guns at Co Roc lashed out, the men of Company I could see the flashes and provided advance warning to the base. Whenever possible they directed retaliatory air strikes on the offenders."*
Does the word `advance` *have a military meaning?*

A (GPT-4o): *Yes, the term "advance" in the sentence has a military meaning. In the phrase "advance warning," it refers to a military alert or early notification given ahead of an enemy attack, allowing defensive or counteractive measures to be taken. Given the combat context involving enemy fire and retaliatory air strikes, "advance warning" clearly denotes a tactical military function.*

I Overlap of Incorrect Predictions Between GPT 4o and Gemini 2.0 Flash

Metric	Value
Total error overlap	74
% of 4o errors (122)	60.66%
% of Gemini errors (159)	46.54%
% of total dataset (1500)	4.93%
Same error type in overlap	63