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 Gemini 2.5 Flash and Llama 3.1-8B yielded accuracies of 87.6% and 80.8% 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 Gemini 2.5 Flash and Llama 3.1-8B yielding 87.6% and 80.8% 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 56 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 400 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).

We constructed a balanced evaluation dataset by selecting at least 15 military and 15 non-military usages for each of the 56 terms. MLCU contains 3,523 examples, with 1,984 labeled `Defense Flag = 1` and 1,539 labeled `Defense Flag = 0`. Per-term medians are 22 and 21.5 examples, and no single term accounts for more than 5% of the dataset.

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 Google's Gemini 2.5 Flash and Meta's Llama 3.1-8B-Instruct models.

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 each model 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 results across all 3,523 examples, normalized by term and defense flag. Gemini achieved 87.6% accuracy, and Llama 80.8%, corresponding to 438 and 645 total errors respectively. To avoid penalizing models for ambiguous or unstable outputs, we adopted a conservative scoring procedure. For Gemini, when an initially incorrect classification yielded a correct interpretation after explicitly prompting the model to explain its reasoning, the prediction was counted as correct. For Llama, responses were scored based solely on the final `yes/no` output, even when the accompanying reasoning was unclear or incoherent. As a result, the reported accuracies should be interpreted as upper-bound estimates rather than optimized performance.

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.

Model	Accuracy	Precision	Recall	F1 Score	TP	FP	FN	TN
Gemini	87.56%	60.71%	95.02%	96.65%	1923	377	61	1165
Llama	80.79%	58.93%	85.70%	90.20%	1802	463	182	1076

Table 1: Normalized per term per defense flag evaluation metrics for Gemini and Llama on the dataset.

Model	MMLU	MLCU
Gemini	88.4%	87.6%
Llama	69.4%	80.8%

Table 2: Benchmark performance on MMLU and our custom dataset (MLCU). Scores are reported from LLMDB.com (2025) and Meta AI (2024), except MLCU which reflects results from our evaluation.

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 scores. As shown in Table 2, MLCU 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 better understand incorrect responses, we further examined model outputs, noting that closed-source models such as Gemini can produce different classifications when prompted for explicit explanations, reflecting the dynamic nature of LLM behavior (Chen et al., 2024). Our MLCU benchmark was run via the Gemini API and on a local machine for Llama in December 2025.

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."*, Gemini answered: *"Yes, "Falcon" is used with a military meaning in this sentence. While it literally refers to the bird, its presence in the context of the "U.S. Air Force Academy's flying culture" strongly links it to the Academy's military identity and its mascot."* 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 Llama labeled it as having a military meaning due to context (Appendix F).

2. **Incorrect definition**: The model misinterprets the meaning of the term, leading to an incorrect classification. *Example*: `company` was misclassified by Llama as a group of soldiers or ships instead of the common definition of with a group (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 Gemini as part of the phrase `advance warning`, resulting in an incorrect military classification (Appendix H).

4. **Bad response**: The model produces incoherent or non-human-interpretable output, often consisting of repeated words or numerical sequences. This error type was observed only in the Llama outputs (Appendix I).

An analysis of error types reveals distinct patterns in how the two models fail. Among Gemini's 438 incorrect responses, the majority (89.5%) were due to incorrect reasoning, followed by misunderstood terms (9.1%) and incorrect definitions (1.4%). These correspond to 11.12%, 1.13%, and 0.17%

Error Type	Gemini			Llama		
	Count	% of Incorrect	% of Total	Count	% of Incorrect	% of Total
Incorrect reasoning	392	89.5%	11.12%	487	75.50%	13.82%
Incorrect definition	6	1.4%	0.17%	127	19.68%	3.60%
Misunderstood	40	9.1%	1.13%	17	2.63%	0.48%
Total Incorrect	438	100.00%	12.43%	645	100.00%	18.30%

Table 3: Comparison of Error Types Between Gemini 2.5 Flash and Llama

Error Type	Gemini		Llama	
	Yes Answer	No Answer	Yes Answer	No Answer
Incorrect reasoning	331	61	349	138
Incorrect definition	6	0	93	34
Misunderstood	40	0	15	2
Bad Response	0	0	6	8
Total	**377**	**61**	**463**	**182**

Table 4: Distribution of Error Types by Initial Response ("Yes"/"No") for Gemini and Llama

of the total dataset, respectively. Llama, on the other hand, demonstrated a weaker skew toward reasoning errors: 487 out of its 645 incorrect responses (75.50%) were due to incorrect reasoning, with incorrect definitions and misunderstood terms accounting for 19.68% and 2.63% of its errors, respectively. As a share of the total dataset, incorrect reasoning in Gemini alone affected 13.82% of examples (Table 3). 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 338 incorrect predictions, representing approximately 9.6% of the total dataset. This shared error set accounts for 58.68% of Gemini's incorrect responses and 52.40% of Llama's. Notably, 251 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 J). 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 Gemini is more likely to answer "Yes" when making a mistake, while Llama is more likely to answer "No" than the other model. (see Table 4). This could show that Gemini is more likely to be overconfident and Llama 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, Gemini 2.5 Flash and Llama 3.1-8B, and this may not reflect the full range of capabilities across other LLMs. 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.

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.

Meta AI. 2024. Meta Llama 3.1 8B Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

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. cell - A subordinate organization formed around a specific process, capability, or activity within a designated larger organization of a joint force commander's headquarters.

13. COIN - counter-insurgency

14. 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.

15. 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.

16. 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.

17. 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.

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

19. CONUS - Continental United States

20. COP - Combat Outpost

21. dispersal - Relocation of forces for the purpose of increasing survivability.

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

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

24. Eagle - F-15 Eagle

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

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

27. 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.

28. ERA - explosive reactive armour

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

30. Falcon - F-16 Fighting Faclon

31. Fifty - An M2 .50 caliber machine gun.

32. 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.

33. 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.

34. Galaxy - C-5 Galaxy

35. HE - High Explosive

36. Hercules - C-130 Hercules

37. KIA - Killed in Action

38. Klicks - Kilometers

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

40. Liberty - MC-12W Liberty

41. `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).

42. `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.

43. `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.

44. `Osprey` - Cv-22 Osprey

45. `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.

46. `Pegasus` - KC-46A Pegasus

47. `Predator` - MQ-1B Predator

48. `Raptor` - F-22 Raptor

49. `Scout` - A man specially trained in shooting, in using ground and cover, in observing, and in reporting the results of observation. A man who gathers information in the field. To reconnoiter a region or country to obtain information of the enemy or to accomplish any other military purpose.

50. `Sentry` - E-3 Sentry

51. `Spirit` - B-2 Spirit

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

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

54. `synchronization` - 1. The arrangement of military actions in time, space, and purpose to produce maximum relative combat power at a decisive place and time. 2. In the intelligence context, application of intelligence sources and methods in concert with the operation plan to answer intelligence requirements in time to influence the decisions they support

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

56. `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

Gemini was 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. Llama was prompted on a local machine and always generated a response with yes or no using temperature setting of 0. Due to output lengths, responses were limited to 256 tokens.

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 (Llama 3.1): *yes. the s-3 bird impact range is a military facility. the word* "`bird`" *in this context is a military term. it is not referring to a living creature. it is referring to a projectile, a chicken, that is used to test the strength of aircraft canopies. the word* "`bird`" *in this context is a military slang term. it is a colloquialism that is used to refer to a projectile, a chicken, that is used to test the strength of aircraft canopies. the word* "`bird`" *in this context is a military term that is used to refer to a chicken that is used as a projectile to test the strength of aircraft canopies. the word* "`bird`" *in this context is a military slang term that is used to refer to a chicken that is used as a projectile to test the strength of aircraft canopies. the word* "`bird`" *in this context is a military term that is used to refer to a chicken that is used as a projectile to test the strength of aircraft canopies. the word* "`bird`" *in this context is a military slang term that is used to refer to a chicken that is used as a projectile to test the strength of aircraft canopies. the word* "`bird`" *in this context is a military*

G Major Example

Q: *In the example: "Large, reliable, long-endurance, multimission destroyers are essential for any navy that aspires to conduct sustained operations an ocean away from its home waters. Modern destroyers are expected to provide air defense for themselves and other ships in company, defend against submarines, and use embarked helicopters to find and attack hostile submarines before they can close to firing range. Destroyers also are expected to prevail in encounters with other ships, using ASCMs and helicopters with antiship missiles."*
Does the word `column` *have a military meaning?*

A (Llama 3.1): *yes. in military contexts,* "`company`" *refers to a unit of soldiers, typically consisting of 60-200 soldiers. in this example, "ships in company" means ships that are together in a group, likely a task force or a fleet. the word* "`company`" *is used here in a military sense, referring to a group of ships rather than a unit of soldiers. the context of the sentence makes it clear that the word* "`company`" *is being used in a military sense, rather than its more common meaning of a business partnership or a group of people. therefore, the answer is yes. the word* "`company`" *does have a military meaning in this example. note: the word* "`company`" *in military contexts is often used to refer to a group of ships, aircraft, or other military units that are operating together as a single entity. this usage is distinct from the more common meaning of the word, which refers to a business partnership or a group of people. in the example "large, reliable, long-endurance, multimission destroyers are essential for any navy that aspires to conduct sustained operations an ocean away from its home waters. modern destroyers are expected to provide air defense for themselves and other ships in company, defend against submarines,*

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 (Gemini 2.5): *Yes,* "`advance`" *is used with a military meaning in this sentence. It refers to providing warning *before* an enemy attack, which is a critical tactical action in a combat scenario.*

I Bad Response

A (Llama 3.1): *the final answer is: yes. (source: fm 3-27) (source: fm 3*

J Overlap of Incorrect Predictions Between Gemini and Llama

Metric	Value
Total error overlap	338
% of Gemini errors (438)	58.68%
% of Llama errors (645)	52.40%
% of total dataset (3523)	9.59%
Same error type in overlap	251