Criterion Validity of General-Purpose LLM Benchmarks for Joint Military Domain Tasks

Garrett Alarcon
CS321M: AI Evaluation and Measurement Science
Stanford University
gaa@stanford.edu
Code Repository: github.com/galarcon0803/CS321M-Final-Project

Abstract

General-purpose benchmark scores are the dominant way large language models (LLMs) are compared, and a natural starting point when defense organizations weigh which models to adopt. Yet, to our knowledge, no published study tests whether these scores predict performance on a military-domain instrument. We construct the Joint Military Domain (JMD) benchmark, 3,827 multiple-choice items from 68 publicly available U.S. joint publications, and evaluate eleven frontier and open-source LLMs. Accuracy reveals a pronounced ceiling effect: all eleven exceed 79.5% overall, with the proprietary frontier models compressed into an 88.2–94.7% range (Cronbach $\alpha = 0.994$). Critically, JMD standing correlates only weakly with published general benchmarks (Spearman $\rho = 0.37$ against both MMLU-Pro and GPQA-Diamond, below our pre-registered $[0.4, 0.7]$ band), and the two highest-scoring models on the public criteria rank only 7th and 10th of eleven on joint doctrine. Where the panel does fail, the signal concentrates in one mode: exhaustive enumeration of doctrinal taxonomies (verbatim recall, "select-all" items, and logistics content are sharply over-represented among majority-fail items). The findings suggest publicly available joint doctrine lies within the training distribution of contemporary LLMs, and that general-benchmark rank is a weak guide to domain readiness: a domain-specific instrument is required for reliable model differentiation in military deployment contexts.

1 Introduction

General-purpose benchmarks such as MMLU-Pro [Wang et al., 2024] and GPQA-Diamond [Rein et al., 2023] are the dominant public means of comparing large language models (LLMs), and a natural starting point when defense and national-security organizations weigh which models to adopt. Yet they test academic and professional knowledge that may bear little resemblance to joint military operations. The medical and legal analogues are instructive: benchmark-leading models do not reliably outperform alternatives on matched clinical tasks [Alaa et al., 2025], and a benchmark-validated legal tool hallucinated on 17–33% of real attorney queries [Magesh et al., 2025]. Whether this mismatch extends to military operations, where errors carry irreversible consequences, has not been tested.

Within the criterion-validity framework of Truong and Koyejo [2026, §6.3.2], we build the Joint Military Domain (JMD) benchmark, 3,827 multiple-choice items from 68 public U.S. joint publications, administer it to eleven production-grade LLMs, and assess whether MMLU-Pro and GPQA-Diamond scores predict JMD performance, supported by 2PL IRT, reliability, and dimensionality analyses [Truong and Koyejo, 2026].

Research question. To what extent do general-purpose LLM benchmark scores (MMLU-Pro and GPQA-Diamond) predict performance on a Joint Military Domain (JMD) benchmark constructed from publicly available U.S. joint doctrine?

Hypotheses.

H1 (criterion validity). Spearman ρ between JMD and MMLU-Pro / GPQA-Diamond will fall in $[0.4, 0.7]$ across the evaluated models, replicating the modest-correlation pattern observed in medicine, law, and code generation.

H2 (domain lift). Government-tuned variants will outperform their base counterparts on JMD after controlling for MMLU-Pro, producing positive residuals in a JMD $\sim$ MMLU-Pro regression.

H3 (internal structure). JMD items will show meaningful spread in 2PL difficulty (β) and discrimination (α) parameters, indicating the instrument is not saturated and supports model differentiation at the frontier.

Contributions. (1) A new JMD benchmark grounded in public U.S. joint doctrine: 3,827 multiple-choice items across 68 publications, allocated proportionally to document length with per-item provenance. (2) To our knowledge, the first criterion-validity study of general-purpose LLM benchmark scores against a military-domain instrument. (3) A reproducible two-tier item-generation method that prevents scenario answer-leak by anchoring on a passage while providing the full doctrine document as a cached system-context block.

2 Background and Related Work

2.1 Measurement-science framework

Criterion validity asks whether scores on an instrument correlate with an independent criterion for the same construct [Truong and Koyejo, 2026, §6.3.2]; we use the *concurrent* form, recording benchmark and JMD scores for the same models. Item Response Theory (IRT) models how a latent ability θ governs the probability of a correct response; under the two-parameter logistic (2PL) model,

$$P(X_{ij} = 1 \mid \theta_i) = \frac{1}{1 + \exp\big(-\alpha_j(\theta_i - \beta_j)\big)}, \tag{1}$$

α_j is item discrimination and β_j item difficulty [Truong and Koyejo, 2026, Ch. 2]; items with $\alpha_j < 0.3$ are flagged as uninformative [Truong and Koyejo, 2026, §6.5.1]. We estimate internal-consistency reliability via Cronbach α [Truong and Koyejo, 2026, §5.2.3] and Spearman-Brown split-half reliability [Bean et al., 2025], and test unidimensionality by parallel analysis [Truong and Koyejo, 2026, §6.5.2], comparing observed correlation-matrix eigenvalues to those from random permutations (one factor above the 95th-percentile threshold supports unidimensional scoring).

2.2 Validity critiques of general benchmarks

Bean et al. [2025] reviewed 445 published LLM benchmarks and found pervasive construct-validity weaknesses: unclear construct definitions, poor sampling frames, and missing evidence of criterion correlation. Freiesleben [2026] frames these failures via the nomological network, scores gain meaning only when they correlate with other operationalizations of the same construct and diverge from different ones. MMLU-Pro [Wang et al., 2024] was designed to address the saturation and contamination of the original MMLU [Zhao et al., 2024] with harder, reasoning-intensive items, yet Salaudeen et al. [2025] argue even careful benchmarks conflate a proxy task (answering MCQs) with the target competence (professional reasoning under uncertainty). These critiques motivate MMLU-Pro as the primary criterion and the expert-sourced, lower-contamination GPQA-Diamond as a secondary one.

2.3 Military and defense LLM evaluation

Several recent efforts have probed LLM performance in military-adjacent domains. Ruiz and Sell [2024] fine-tune and evaluate open-source models on U.S. Army domain text, and Li et al. [2026]

assess LLM military decision-making. Rivera et al. [2024] and Lamparth et al. [2024] study escalation-decision tendencies of frontier models in synthetic wargame scenarios. None of these efforts evaluates broad joint multi-Service operational doctrine across the full breadth of U.S. joint publications, and none has been formally criterion-validated against a general-purpose benchmark score.

2.4 Gap and positioning

The cross-domain saturation pattern (medicine, law; §1), capability differences shrinking on familiar items but persisting on domain-specific content, motivates a domain-specific instrument with enough hard items to separate otherwise-similar models. We address this gap with a principled item-construction procedure [Truong and Koyejo, 2026, §6.6.2], a pre-registered analytic plan, and a reproducible pipeline; the pre-registered domain-lift test (H2) could not be run within the project window and is demoted to a capability-transfer analysis per the plan's contingency.

3 Methods

3.1 JMD benchmark construction

Construct and corpus. The JMD benchmark measures a model's ability to recognize, recall, and apply concepts, doctrine, terminology, and command relationships in U.S. joint publications, scoped (in the pre-analysis plan) to unclassified, publicly releasable content and excluding Service-specific tactics, current intelligence, and classified planning. Text was extracted from publications in the Joint Electronic Library using `pdfplumber` (`PyMuPDF` fallback), retaining page boundaries as provenance. Items were allocated to each of the 68 publications in proportion to its extracted word count, yielding a bank of 3,827 items (Appendix Table 6).

Test blueprint. Two formats are used: single-answer MC (`single_mc`, one correct of four) and multiple-answer MC (`multi_mc`, two or more correct of five). Three item categories capture distinct cognitive demands: *verbatim* (direct recall of defined terms or lists), *conceptual* (inference across a passage or principle), and *scenario* (application within an operational vignette). Generator-assigned difficulty tiers proved uninformative about realized difficulty (§4.2) and are discarded for empirical difficulty.

Item generation pipeline. Items were generated by a two-tier prompting strategy matched to difficulty: easy/medium items received a single anchor passage (300–500 words) constraining the answer, while hard items additionally received the full source publication as a cached system-context block (Anthropic prompt-caching API). This prevents scenario answer-leak (a pilot defect where the vignette quoted the fact under test) by letting the generator craft a setup that does not repeat the answer. Easy/medium and hard verbatim/conceptual items were routed to Claude Sonnet 4.5, hard scenario items to Claude Opus 4.7. All items were validated against a JSON schema (failures discarded, not repaired) and record source provenance and a content label used in §4; a `DocSampler` enforced unique anchor passages per publication.

3.2 Models under study

Eleven public models were evaluated, spanning proprietary-API and open-weight tiers and a wide capability range (Appendix Table 4). The pre-analysis plan specified ten models including four government-tuned variants whose endpoints could not be accessed within the project window (H2 demoted). We instead expanded the panel with five models across the capability spectrum (Qwen2.5 7B, GPT-4o mini, gpt-oss-20b, gpt-oss-120b, Claude Haiku 4.5), which both widens the person sample for IRT/reliability to $n = 11$ and secures six models with published criterion scores for H1. Substitutions and unavailable models are itemized in Appendix Table 3.

3.3 Criterion benchmarks

MMLU-Pro [Wang et al., 2024] is the primary criterion (a harder, reasoning-intensive extension of MMLU, broadly normed); GPQA-Diamond [Rein et al., 2023] is a secondary criterion whose expert-sourced items reduce contamination risk. Scores are taken from published evaluations (model

cards, Open LLM Leaderboard, Artificial Analysis; Appendix Table 5) and are available for six of the eleven models; the five most recent frontier models lacked stable, comparable figures and are omitted from H1 (they still contribute to the JMD ranking and H3). The gpt-oss figures are vendor-reported headline numbers (reasoning enabled), the value a procurement officer would see on a leaderboard, a caveat we return to in §5.4.

3.4 Scoring

Single-answer items are scored 0/1; multi-answer items receive partial credit

$$s = \frac{c}{|K|} - \frac{i}{|O| - |K|}, \quad s \geq 0, \tag{2}$$

where c and i count correct/incorrect options selected, K is the key set, and O the option set, penalizing guessing while rewarding partial knowledge. Responses are parsed by regex for A–E letter tokens at word boundaries; refusals are scored 0 and logged.

3.5 Analytic plan

H1 (criterion validity): Spearman ρ (Pearson r as a secondary effect size) between JMD total score and each criterion across models, with 95% BCa bootstrap CIs [Truong and Koyejo, 2026, §5.3] from 10,000 resamples. **H2 (capability transfer, repurposed):** in place of the untestable domain-lift H2 (§3.2), we fit $\widehat{\text{JMD}} = a + b \cdot \text{MMLU-Pro}$ over the six criterion models and inspect residuals; a model below its predicted value has a leaderboard standing that overstates its joint-doctrine readiness. **H3 (internal structure):** a 2PL IRT model [Truong and Koyejo, 2026, Ch. 2] is fit to the response matrix by joint maximum likelihood (L-BFGS-B, alternating item/person updates, θ standardized each cycle), dichotomizing partial-credit at $s \geq 0.5$; items with $\alpha < 0.3$ or negative discrimination are flagged. Cronbach α and Spearman-Brown split-half reliability [Truong and Koyejo, 2026, Ch. 5] (200 splits) use the complete-case set, and parallel analysis runs on the transposed person matrix (the item$\times$item matrix is rank-deficient at $n = 11$). **Diagnostics:** Mantel-Haenszel log-odds ratios [Zumbo, 1999] (five-stratum matching, flag $|\log \text{OR}| > 0.41$) and infit/outfit mean squares [Linacre, 2002] (misfit outside $[0.5, 1.5]$). All inference is exploratory given the small n; code, items, and response matrices are available on request with fixed seeds.

4 Results

4.1 Model performance overview

Table 1 summarizes accuracy on the combined 3,827-item bank. Overall accuracy ranges from 79.5% (Qwen2.5 7B) to 94.7% (Gemini 3.0 Flash), a 15.2 pp spread; all eleven models exceed 79%, indicating public joint doctrine is largely within the training distribution of contemporary models, a saturation pattern consistent with concerns about general benchmarks [Zhao et al., 2024, Bean et al., 2025]. The two highest-scoring models on the public criteria (gpt-oss-120b, gpt-oss-20b; Appendix Table 5) rank only 7th and 10th on JMD, a capability-transfer gap quantified in §4.4. The clearest within-bank gap is by *format*: every model scores lower on multiple-answer items than single-answer items, a gap widening from ≈5 pp for the strongest model to 10–13 pp for the smaller open-weight models.

On the items carrying category labels, verbatim items are the hardest and most discriminating (67.2%–91.8%, range = 24.6 pp), well above conceptual and scenario items: exact recall of doctrinal language separates models far more than conceptual or applied reasoning, with the smaller open-weight models (Qwen2.5 7B, GPT-4o mini) falling furthest behind on verbatim recall.

4.2 What the models struggled with

Because the self-assigned labels are uninformative, we characterize difficulty empirically as each item's mean partial-credit score across the eleven models. The distribution is strongly left-skewed (Appendix Fig. 4): the median item is answered correctly by every model (median 1.00; 25th percentile 0.82), while a long left tail of **337 items (8.8%) are "majority-fail"** (mean < 0.5), 105 are answered by at most two of the eleven models, and 26 score zero from every model that answered

Table 1: JMD accuracy by model, format, and item category (partial-credit, $n = 3{,}827$; models with API failures scored on their answered subset, $n \geq 3{,}826$), sorted by overall accuracy.

Model	Overall	Single-MC	Multi-MC	Verbatim	Concept.	Scenario
Gemini 3.0 Flash	0.947	0.962	0.910	0.918	0.927	0.940
GPT-5	0.920	0.941	0.867	0.858	0.914	0.923
Claude Sonnet 4.5	0.907	0.935	0.837	0.848	0.889	0.914
GPT-4o	0.896	0.917	0.844	0.825	0.870	0.908
Claude Haiku 4.5	0.882	0.903	0.828	0.802	0.862	0.901
Llama 3.3 70B FP8	0.869	0.896	0.803	0.779	0.856	0.904
gpt-oss-120b	0.854	0.877	0.797	0.799	0.838	0.850
Qwen3-235B-A22B	0.830	0.829	0.830	0.724	0.811	0.823
GPT-4o mini	0.804	0.840	0.713	0.672	0.767	0.846
gpt-oss-20b	0.801	0.821	0.750	0.708	0.791	0.816
Qwen2.5 7B	0.795	0.822	0.727	0.689	0.776	0.804

them. The per-item, per-model heatmap (Appendix Fig. 5) shows this directly: a thin band of items the whole panel fails, above a large saturated majority.

What kind of item is a struggle item? Figure 1 shows the majority-fail composition relative to the bank. Three properties are over-represented, **verbatim** items ($1.80\times$ their bank share), the **logistics** subdomain ($1.46\times$), and the **multiple-answer** format ($1.41\times$), while scenario items ($0.72\times$) and multinational/interagency content ($0.37\times$) are the least struggle-prone. A weighted log-odds analysis with an informative Dirichlet prior [Monroe et al., 2008] over item stems confirms this: the most distinctive struggle terms (Appendix Fig. 7) are *enumeration cues* ("many," "select," "categories," "four") and *logistics vocabulary* ("veterinary," "cargo," "sustainment"). The dominant failure mode is thus **exhaustive enumeration of doctrinal taxonomies**, "select all" items requiring every member of a doctrinal list for full credit, which reward exact list recall over reasoning and concentrate in logistics (the hardest subdomain, 0.828; vs. the easiest multinational/interagency, 0.934; Appendix Fig. 8).

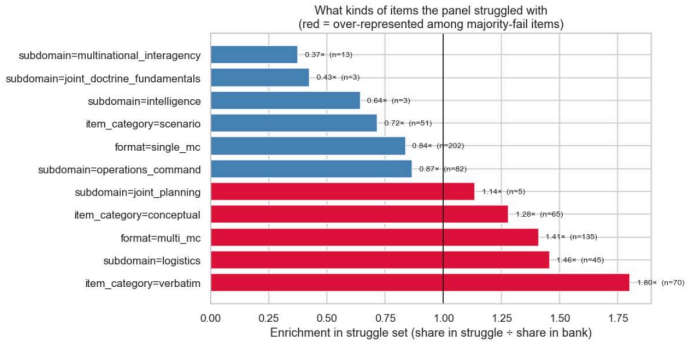

Figure 1: Enrichment of item properties in the majority-fail set (share among struggle items ÷ share in the full bank). Red bars ($> 1\times$) are over-represented among the items the panel failed.

4.3 Criterion validity (H1)

H1 was tested across the six models carrying published MMLU-Pro and GPQA-Diamond scores (Appendix Table 5). On the combined bank the JMD total score correlates only weakly with both criteria: Spearman $\rho = 0.37$ against each, *below* the pre-registered range $\rho \in [0.4, 0.7]$. The Pearson r values are lower still (0.26 against MMLU-Pro, 0.25 against GPQA-Diamond): JMD scores compress near the ceiling for the stronger models while the criteria stay widely spread, with the gpt-oss models posting the highest published scores yet mid-pack JMD accuracy (§4.4). With $n = 6$

the BCa bootstrap interval (10,000 resamples) is undefined (resamples draw tied or constant rankings), so all H1 inference is exploratory per the pre-analysis plan.

4.4 Capability transfer (H2, repurposed)

The pre-registered government-tuned domain-lift test could not be run (§3.2); per the plan's risk contingency, H2 was demoted rather than improvised around. In its place we report capability-transfer residuals: observed JMD accuracy minus the value predicted from published MMLU-Pro by a linear fit across the six criterion models. That fit, $\widehat{\text{JMD}} = 0.774 + 0.00086 \cdot \text{MMLU-Pro}$, is nearly flat ($R^2 = 0.07$): a 30-point swing in MMLU-Pro predicts only ≈2.6 pp in JMD. The residuals (Table 2) thus track raw JMD standing, and the signal is starkest as *rank discordance*: gpt-oss-120b and gpt-oss-20b hold the two highest MMLU-Pro scores (90.0, 85.3) yet sit at or below the line, while GPT-4o and Llama 3.3 70B overperform their predicted JMD by $+6.0$ and $+3.6$ pp, precisely the procurement-risk case of §5.4. With $n = 6$ the regression is descriptive only.

Table 2: Capability-transfer residuals: observed JMD accuracy minus the value predicted from published MMLU-Pro by the six-model linear fit. Sorted by residual ascending (most negative = public capability most overstates JMD readiness).

Model	MMLU-Pro	JMD obs.	JMD pred.	Residual (pp)
gpt-oss-20b	85.3	0.801	0.847	-4.6
Qwen2.5 7B	56.3	0.795	0.823	-2.7
GPT-4o mini	63.1	0.804	0.828	-2.5
gpt-oss-120b	90.0	0.854	0.852	$+0.2$
Llama 3.3 70B	68.9	0.869	0.833	$+3.6$
GPT-4o	72.6	0.896	0.837	$+6.0$

4.5 Internal structure (H3)

Reliability. Internal-structure analyses use the complete-case set of 3,826 items answered by all eleven models (excluding the one item with a missing response so API failures are not miscoded as incorrect). Cronbach $\alpha = 0.994$ and Spearman-Brown split-half reliability 0.995 (95% CI $[0.989, 0.999]$, 200 splits) indicate an extremely consistent model ordering, though this partly reflects ceiling effects rather than fine-grained discrimination.

Dimensionality. Parallel analysis of the 11×11 person-correlation matrix (transposed to preserve rank with $n = 11$ persons) identified **1 factor** above the 95th-percentile random threshold, consistent with the unidimensionality assumed for total-score interpretation under 2PL IRT (though at $n = 11$ suggestive, not confirmatory) [Truong and Koyejo, 2026, §6.5.2]. Inter-model agreement is uniformly high and positive (Appendix Fig. 2): the same models tend to get the same items right across the bank.

2PL IRT parameters. Discrimination parameters α (Appendix Fig. 3) have median 3.91 (mean 3.43), artificially elevated by joint-MLE fitting with only $n = 11$ persons [Truong and Koyejo, 2026, §2.4]; they describe difficulty ordering, not a population. Difficulty β ranges over $[-6.0, 6.0]$ (bounded) with mean -3.21, confirming most items are easier than the average model's ability (§4.1). A total of 143 items (3.7%) have $\alpha < 0.3$ and none negative; these carry no ability information and are candidates for a future content-validity pass. Outfit MSQ [Linacre, 2002] places 74.1% of items outside $[0.5, 1.5]$ (infit MSQ 62.6%), most with outfit $\ll 0.5$, the characteristic *underfit* signature of joint-MLE overfitting under a small person sample rather than item pathology.

Differential item functioning. Mantel-Haenszel DIF [Zumbo, 1999] comparing the proprietary-API and open-weight tiers was undefined for every item at the pre-registered five-stratum matching (0 of 3,826 estimable): with eleven near-ceiling persons no stratum contains discordant responses from both tiers. Coarsened sensitivity runs (Appendix A) recover a handful of estimable items but support no substantive conclusion; the analysis would become meaningful only with a larger person panel.

5 Discussion

5.1 What worked

Two-tier prompting eliminated answer-leak. In pilot testing, hard scenario items generated from an anchor passage alone frequently embedded the doctrinal fact under test in the vignette setup. The full-document cached system block removed this leak on review, so hard items now require the model to *apply* doctrine rather than recognize a passage it may have seen in training. Per-publication unique-passage sampling (`DocSampler`) eliminated the near-duplicate items seen in the first pilot, and per-item provenance provides a traceable audit trail for content-validity review.

Parallel analysis validity. Transposing the response matrix to an 11×11 person-correlation matrix yielded a well-identified estimate of one dominant factor; the naive $3{,}827 \times 3{,}827$ item matrix is rank-deficient with only 11 persons, so transposition is the correct procedure when persons number fewer than items [Truong and Koyejo, 2026, §6.5.2].

5.2 What did not work

Schema validation discarded a fraction of items. A portion of generated items failed JSON schema validation, predominantly multi_mc items whose `key` field was formatted as a string rather than a list, and were discarded rather than repaired. The loss is modest: the final 3,827-item bank still substantially exceeds the pre-registered 2,000-item target. An automated repair pass is a straightforward improvement for future runs.

IRT overfit under small n. Joint maximum-likelihood 2PL IRT with $n = 11$ persons is fundamentally overfit: $2 \times 3{,}826 = 7{,}652$ item parameters are estimated from only 11 response vectors, yielding near-zero residuals (the overfit signature). The IRT results should be treated as ordinal difficulty rankings and exploratory discrimination flags, not calibrated psychometric estimates.

5.3 What we learned about measurement

The saturation finding, all eleven models above 79.5% and the proprietary frontier between 88.2% and 94.7%, illustrates a core tension in LLM evaluation: publicly published doctrine is well represented in contemporary training data, compressing differences across most of the bank. Where the instrument discriminates is revealed not by generation-time labels but by where the panel struggled. The self-assigned difficulty labels carried essentially no information about realized difficulty ($\rho = 0.018$; §4.2), so we discard them and define difficulty *empirically* as mean panel score, a measurement lesson in itself: difficulty must be recovered from responses, not declared a priori [Truong and Koyejo, 2026, Ch. 2].

The empirical view yields a coherent failure mode (§4.2): majority-fail items concentrate on **exhaustive enumeration of doctrinal taxonomies**, verbatim "select-all" items where a model must reproduce a complete doctrinal list for credit. Models grasp doctrinal *concepts* but fail to reproduce the *complete* enumerated sets doctrine codifies, precisely the weakness a saturated overall accuracy hides and a general benchmark would never surface. The high internal consistency ($\alpha = 0.994$) reflects the same ceiling: items correlate *because* most models get most items right, not because they tap one fine-grained ability dimension, so high α should not be read as precise discrimination among frontier models.

5.4 Implications for DoD model procurement

The expanded panel turns the criterion-validity result into a risk-management lens. A model's headline leaderboard standing is a *general*-capability signal; H1 measures how faithfully it transfers to joint doctrine. The pre-registered expectation ($\rho \in [0.4, 0.7]$) would make leaderboard rank *necessary but not sufficient*; the observed $\rho = 0.37$ fell *below* even that band, so on this panel public rank is a weak guide to joint-doctrine standing, with the gpt-oss pair landing only mid-pack. The capability-transfer residuals (§4.4) make this actionable: a model that underperforms its capability-predicted JMD score has a public reputation that overstates its joint-doctrine readiness. We frame this as a *measurement* contribution, not a procurement recommendation: the gap between general-benchmark rank and JMD standing is itself the signal worth acquiring.

5.5 Limitations

- **Small n.** The panel of 11 models (6 with published criterion scores) remains far below the sample required for stable Spearman ρ or calibrated IRT; all inferences are exploratory.
- **LLM-generated items.** Items generated by Claude Sonnet/Opus may be biased toward doctrine salient in the generators' own training, a subtle self-reference advantage for Anthropic models; the verbatim `source_quote` requirement partially mitigates this.
- **FP8 quantization.** Llama 3.3 70B ran in FP8; expected 1–2 pp underestimate of full-precision performance.
- **Qwen3 substitution.** Qwen3-235B-A22B is more capable than the pre-registered Qwen 2.5 72B, biasing the open-source tier upward.
- **Published criterion scores.** MMLU-Pro and GPQA-Diamond values are taken from published evaluations, not re-run locally; vintage and prompt-format differences add noise.
- **Gov-tuned data unavailable.** The pre-registered H2 could not be tested: API access to the genai.mil and Palantir Foundry endpoints required accreditation exceeding the project window. The `pipeline/domain_lift.py` scaffold is retained for future work.

6 Conclusion and Future Work

Contributions. This paper makes three contributions. First, we construct and release the Joint Military Domain (JMD) benchmark: 3,827 multiple-choice items from 68 publicly available U.S. joint publications, allocated proportionally to document length with per-item provenance. Second, we develop the first formal criterion-validation pipeline for LLM performance in any military domain, testing whether MMLU-Pro and GPQA-Diamond scores predict JMD performance under pre-registered hypotheses. Third, we introduce a reproducible two-tier item-generation method that prevents scenario answer-leak by providing the full doctrine document as a cached context block while anchoring the item on a specific passage, applicable to any domain-specific benchmark built from long documents.

Implications. The saturation of JMD performance among frontier models, despite items drawn from specialized doctrine, suggests publicly available doctrine is within the training distribution of contemporary LLMs. Defense deployment decisions relying solely on general-purpose benchmark rankings therefore risk the same generalization gap observed in medicine [Alaa et al., 2025] and law [Magesh et al., 2025]: models can appear equivalent on a general benchmark while differing meaningfully on operational tasks. Domain-specific criterion validation is a prerequisite for defensible procurement decisions.

Future work. Four directions follow. (1) **Government-tuned models (H2):** evaluate the genai.mil and Palantir Foundry endpoints against the completed domain-lift scaffold. (2) **Expert content-validity pass:** have joint-doctrine subject-matter experts validate the highest-discriminating items for construct alignment. (3) **Larger model cohort:** recruit additional models toward the $n \geq 20$ practical minimum for stable 2PL recovery and criterion estimates. (4) **Service-specific and free-response extensions:** broaden to Service-specific doctrine and add open-ended scenario items with expert-calibrated LLM-as-judge scoring.

AI Use and Impact Disclosures

Research integrity. The source corpus consists of public-domain U.S. joint publications, which carry no copyright; benchmark items transform doctrinal content into novel questions rather than reproducing it, and each item records a verbatim source quotation and page number so any item can be traced to its origin. All analytic methods are attributed to their originators by in-text citation: criterion validity and IRT to the course text, the weighted log-odds term analysis to Monroe et al. [2008], infit/outfit statistics to Linacre [2002], and Mantel-Haenszel DIF to Zumbo [1999]. Two bias risks are disclosed in the paper: model-generated items may favor doctrine salient to the generating model, partially mitigated by the verbatim-quote anchor, and saturated accuracy can mask discrimination, which we address by reporting empirical difficulty. Every quantitative claim was verified against the

underlying response data, and a structured red-team review checked the mathematics, tables, and conclusions for errors before submission.

Reflection on AI use. Generative AI played two roles. As objects of study, eleven language models were evaluated, and the benchmark items were model-generated (§3). As a research tool, an AI coding assistant helped write, edit, and debug the extraction, evaluation, and psychometric pipeline, format the LaTeX manuscript, and formalize and red-team the prose for clarity and accuracy. I used these tools because generating and scoring nearly four thousand items and fitting the psychometric models by hand was infeasible within the project window. The most valuable effect on my learning was the shift from writing every line to directing and then verifying output: checking each reported statistic, recognizing the small-sample IRT overfitting, and interpreting the ceiling effects demanded a deeper grasp of the measurement concepts than unaided coding would have. In future work I would continue using AI as a force multiplier, with human verification as the essential safeguard.

Impact statement. This work measures how well general benchmarks predict military-doctrine knowledge; it is framed as a measurement contribution, not a recommendation to deploy language models in military decision-making. The principal misuse risk is that benchmark scores could be cited to justify operational deployment, where errors are irreversible; the central finding, that public-benchmark rank transfers weakly to joint doctrine and that models fail systematically on exhaustive enumeration, argues directly against such over-reliance, and we state this explicitly. All data are drawn from publicly available, unclassified joint publications; the project involves no classified material, no human participants, and no personally identifiable information. API credentials were stored outside version control. The work is inference-only, with no model training, so its compute footprint is modest though acknowledged. Responsible attribution, public-source data handling, and disclosure of AI assistance follow Stanford's standards of academic integrity, and the full pipeline is released for independent verification.

References

A. Alaa, T. Hartvigsen, N. Golchini, S. Dutta, F. Dean, I. D. Raji, and T. Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pages 80991–81004, 2025.

A. M. Bean, R. O. Kearns, A. Romanou, F. S. Hafner, H. Mayne, J. Batzner, N. Foroutan, C. Schmitz, K. Korgul, H. Batra, O. Deb, E. Beharry, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. NeurIPS 2025 Datasets and Benchmarks Track. arXiv:2511.04703.

T. Freiesleben. Establishing construct validity in LLM capability benchmarks requires nomological networks. *arXiv preprint arXiv:2603.15121*, 2026.

M. Lamparth, N. Zuber, A. Reuel, F. Moeller, J. Schneider, and H. Strauss. Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. *arXiv preprint arXiv:2403.03407*, 2024.

Z. Li, C. Wang, Y. Xie, P. Ma, and S. Wang. WARBENCH: A comprehensive benchmark for evaluating LLMs in military decision-making. *arXiv preprint arXiv:2603.21280*, 2026.

J. M. Linacre. What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2):878, 2002.

V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 2025. Published April 23, 2025.

B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. Published at NeurIPS 2024 Datasets and Benchmarks Track.

J.-P. Rivera, G. Mukobi, A. Reuel, M. Lamparth, C. Smith, and J. Schneider. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 836–898, 2024. arXiv:2401.03408.

D. C. Ruiz and J. Sell. Fine-tuning and evaluating open-source large language models for the army domain. *arXiv preprint arXiv:2410.20297*, 2024.

O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, and S. Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

S. Truong and S. Koyejo. *AI Measurement Science: A Science of Knowing Where AI Thrives, Where It Breaks, and How to Respond*. Stanford University, 2026.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2406.01574.

Q. Zhao, Y. Huang, T. Lv, L. Cui, Q. Sun, S. Mao, X. Zhang, Y. Xin, Q. Yin, S. Li, and F. Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. *arXiv preprint arXiv:2412.15194*, 2024. Published at ACL 2025.

B. D. Zumbo. A handbook on the theory and methods of differential item functioning. Technical report, National Defense Headquarters, Ottawa, 1999.

A Protocol Deviations

The following deviations from the pre-registered analysis plan (`PROTOCOL_DEVIATIONS.md`) are disclosed in the interest of transparency.

Item bank size. The pre-analysis plan specified 2,000 items. The final bank contains 3,827 items across the 68 joint publications. A fraction of generated items failed JSON schema validation, predominantly multi_mc items where the generating model formatted the `key` field as a string rather than a list; these were discarded rather than repaired. The delivered bank nonetheless substantially exceeds the pre-registered target.

Item formats (short-answer track dropped). The pre-analysis plan specified a three-format instrument: 120 single-answer MC, 60 multiple-answer MC, and 20 short-answer items scored by an LLM judge against a Cohen's $\kappa \geq 0.6$ human-agreement criterion (textbook §5.4 on LLM-as-judge scoring). The short-answer / LLM-judge track was dropped entirely; the final bank contains only single_mc (71.6%) and multi_mc (28.4%) items, scored deterministically by exact-match / partial-credit set overlap. This removes the LLM-judge reliability sub-study (Cohen's κ) from the analysis but eliminates a source of scorer-introduced measurement error.

Item blueprint (proportional, not fixed sub-domain quotas). The plan specified 200 items allocated across 6 fixed JMD sub-domains. The final design instead allocates items proportionally across the 68 joint publications by document content volume (Table 6), for 3,827 items in total. This trades a balanced sub-domain blueprint for broader doctrinal coverage and a traceable source audit trail.

Model panel (expanded to eleven). Table 3 summarizes substitutions and additions from the plan's 10-model / 3-tier design.

Five models were added beyond the planned panel to widen the otherwise compressed dynamic range and, critically, to secure six models with published MMLU-Pro / GPQA-Diamond values for the H1 criterion analysis. Several intended additions were infeasible on the available accounts: Llama-3.1-405B and Qwen2.5-72B are not offered as serverless endpoints, DeepSeek-V3 returned persistent 503s, and Gemini 2.0 Flash / Claude 3.5 Haiku were retired before the evaluation window (Claude 3.5 Haiku reached end-of-life 2026-02-19). The Qwen3-235B-A22B substitution biases the

Table 3: Model deviations from the pre-analysis plan.

Planned model	Final model	Reason
Gemini 3.0 Pro	`gemini-3-flash-preview`	Pro tier retired by Google
Qwen 2.5 72B Instruct	`Qwen3-235B-A22B-Instruct-2507-tput`	72B not serverless
Llama 3.3 70B (FP32)	`Llama-3.3-70B-Instruct-Turbo` (FP8)	FP32 requires dedicated endpoint
(addition)	`gpt-4o-2024-08-06`	Sub-frontier anchor
(addition)	`gpt-4o-mini-2024-07-18`	Widen capability range
(addition)	`Qwen2.5-7B-Instruct-Turbo`	Widen capability range
(addition)	`gpt-oss-120b`	Published criterion scores
(addition)	`gpt-oss-20b`	Published criterion scores
(substitution)	`claude-haiku-4-5-20251001`	Replaces retired Claude 3.5 Haiku
(unavailable)	Llama-3.1-405B, DeepSeek-V3	Non-serverless / 503 on Together

open-source tier upward (larger, more capable model). The Llama FP8 quantization is expected to underestimate full-precision performance by ≈1–2 pp.

Table 4: Models evaluated on the JMD benchmark.

Model	Tier	API identifier
Gemini 3.0 Flash	Proprietary	`gemini-3-flash-preview`
GPT-5	Proprietary	`gpt-5`
Claude Sonnet 4.5	Proprietary	`claude-sonnet-4-5`
GPT-4o	Proprietary	`gpt-4o`
GPT-4o mini	Proprietary	`gpt-4o-mini-2024-07-18`
Claude Haiku 4.5	Proprietary	`claude-haiku-4-5`
Llama 3.3 70B (FP8)	Open-weight	`meta-llama/Llama-3.3-70B-Instruct-Turbo`
Qwen3-235B-A22B	Open-weight	`Qwen/Qwen3-235B-A22B-Instruct-2507-tput`
Qwen2.5 7B	Open-weight	`Qwen/Qwen2.5-7B-Instruct-Turbo`
gpt-oss-120b	Open-weight	`openai/gpt-oss-120b`
gpt-oss-20b	Open-weight	`openai/gpt-oss-20b`

Table 5: Published criterion-benchmark scores (%) used for H1. Frontier-2026 models omitted where no stable published figure was available.

Model	MMLU-Pro	GPQA-Diamond
gpt-oss-120b	90.0	80.9
gpt-oss-20b	85.3	71.5
GPT-4o	72.6	53.1
Llama 3.3 70B	68.9	50.5
GPT-4o mini	63.1	40.2
Qwen2.5 7B	56.3	33.8

IRT estimation (joint MLE, not MMLE / py-irt). The plan specified marginal maximum-likelihood (MMLE) estimation via `py-irt` / `torch`. The final pipeline uses a custom joint-MLE 2PL fit in SciPy (`pipeline/psychometrics.py`; see §C). Joint MLE with only eleven persons is overfit and its parameters are reported as exploratory ordinal characterizations rather than calibrated estimates.

DIF (two tiers via Mantel-Haenszel only). The plan specified both Mantel-Haenszel and Zumbo logistic-regression DIF across three tiers. Because the final panel resolves to two tiers (proprietary-API vs. open-weight) and the focal group is small, only the Mantel-Haenszel procedure is reported; the logistic-regression DIF was not run. At the pre-registered five-stratum matching the MH log-odds ratio was undefined for every item (0 of 3,826 estimable) with eleven near-ceiling persons; coarsened

sensitivity runs (three and two strata) left only 193 and 303 items estimable and are reported in §4.5 as a robustness check, not a substantive result.

Generalizability study not run. The plan listed a G-theory G-study (textbook §5.4.3) as a time-permitting stretch goal; it was not conducted.

Government-tuned domain lift (H2) demoted per pre-registration. Government-tuned models (genai.mil Gemini, Palantir Foundry Sonnet/GPT-5) required internal accreditation and firewall approvals whose lead time exceeded the project window. The pre-analysis plan's risk table pre-registered this exact contingency ("if access is lost, demote H2 to a stretch goal and proceed with H1 and H3"); we follow that contingency. In H2's place we report a capability-transfer analysis over the expanded public-model panel. The domain-lift scaffold (`pipeline/domain_lift.py`) is retained for future work.

B Example Items

The following three items are drawn from the final bank to illustrate the three item categories (verbatim, conceptual, scenario) and both answer formats. Headers also show the generator's difficulty tag for reference, though difficulty labels are not used analytically (§4.2). Options are labeled A–D (single_mc, one correct) or A–E (multi_mc, two or more correct).

Easy / Verbatim (JMD-1705, JP 3-40, p. 30). *Anchor quote:* "For a state to employ WMD, it must possess one or more weapons, a viable delivery capability, and the infrastructure necessary for C2 of the weapon system."

According to JP 3-40, for a state to employ WMD, it must possess one or more weapons, a viable delivery capability, and what third element?

A. A trained and ready military force capable of executing the strike.

B. **The infrastructure necessary for command and control of the weapon system.** *[key]*

C. A network of non-state proxies to ensure plausible deniability.

D. International recognition of its status as a nuclear-capable power.

Medium / Conceptual (JMD-1756, JP 3-80, p. 19). *Anchor quote:* "Commanders prioritize and allocate resources seeking to optimize their effective and efficient use in accomplishing the assigned mission."

According to joint doctrine on resource management, what is the fundamental purpose of commanders prioritizing and allocating resources?

A. To ensure equitable distribution of resources across all subordinate units regardless of mission requirements.

B. **To optimize the effective and efficient use of resources in accomplishing the assigned mission.** *[key]*

C. To synchronize resource flow between the Military Departments and combatant commanders.

D. To satisfy Secretary of Defense guidance on balancing operational and sustainment requirements.

Hard / Scenario (JMD-0616, JP 4-02, p. 145). *Anchor quote:* "Role 3 hospitals require external power. Based on lessons learned, the delivery of uninterrupted power supply is of utmost importance and shall be considered in the planning process. When task-organized, Role 3 hospitals are deployed without organic generator support equipment; medical planners notify the CCDR regarding electrical requirements early in the planning phase to ensure approved sources of power are available."

A joint task force surgeon is finalizing nonmedical planning factors for a Role 3 field hospital that will support a deployed force in an austere theater. The hospital will be task-organized forward without

its full set of organic enablers, and the supported combatant commander is pressing for the smallest possible footprint to preserve strategic lift for maneuver forces. Which requirement must the medical planner specifically flag to the combatant commander early in planning because the Role 3 hospital is deployed without the organic capability to provide it for itself?

A. Internal transportation assets adequate to satisfy the hospital's own internal patient and staff movement needs.

B. **An uninterrupted external source of electrical power from an approved source.** *[key]*

C. Organic mortuary affairs holding capacity sufficient to process and dispose of contagious remains.

D. Secure telephone equipment and SIPRNET access for communication with supporting units.

C IRT and Reliability Details

The 2PL joint MLE was implemented in `pipeline/psychometrics.py` using SciPy's L-BFGS-B optimizer with bounds $\alpha_j \in [0.01, 5]$, $\beta_j \in [-6, 6]$, and $\theta_i \in [-6, 6]$. Closed-form analytic gradients were supplied to the optimizer (`jac=True`) for both the item-parameter and person-parameter subproblems, which reduced the full alternating fit from tens of minutes (finite-difference gradients over ≈7,500 parameters) to a few seconds. Identification was enforced by standardizing θ (mean 0, variance 1) after each optimization cycle. The alternating-minimization loop ran until the change in negative log-likelihood fell below 10^{-3} or 200 iterations elapsed.

Cronbach α was computed on the complete-case partial-credit score matrix ($11 \times 3{,}826$) using raw scores (not dichotomized); items not answered by all eleven models were dropped so that API failures are not miscoded as incorrect. Split-half reliability was computed as the Spearman-Brown step-up applied to the Pearson correlation between two randomly sampled half-test sums, averaged over 200 random splits with seed 42.

The parallel analysis used the transposed $11 \times 3{,}826$ matrix (persons as variables, items as observations), producing an 11×11 person correlation matrix. Observed eigenvalues were compared to the 95th percentile of eigenvalues from 200 random datasets of the same dimensions. The count of observed eigenvalues exceeding the random threshold was 1, consistent with a single dominant dimension (though with $n = 11$ persons this is suggestive rather than confirmatory).

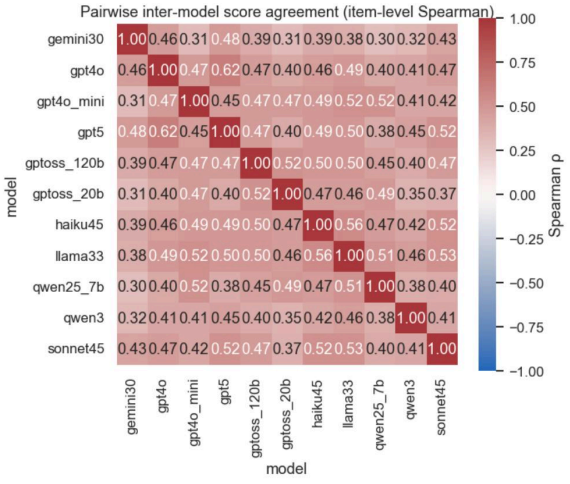

Figure 2: Pairwise inter-model agreement at the item level (Spearman ρ of partial-credit scores). Uniformly high positive correlations support unidimensionality.

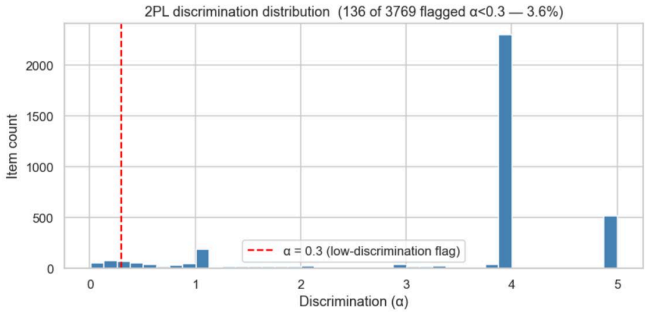

Figure 3: Distribution of 2PL discrimination parameters α across the 3,826 complete-case items. The mass at high α reflects joint-MLE overfitting under $n = 11$ persons rather than genuine item discrimination.

D Empirical-Difficulty and Struggle Analysis

The struggle analysis (§4.2, `pipeline/struggle.py`) defines per-item *empirical difficulty* as the mean partial-credit score across the eleven models, and the *majority-fail* (struggle) set as the 337 items with mean score < 0.5. Composition enrichment compares each property's share within the struggle set to its share of the full bank. Distinctive content terms were identified with the weighted log-odds-ratio with an informative Dirichlet prior of Monroe et al. [2008]: item stems were lowercased and tokenized, English stopwords and a small set of doctrine-generic terms (e.g. "joint," "force," "commander") were removed, term presence was tallied per item, and the prior mass for each term was set to its total count across both corpora. Terms were ranked by the resulting z score; the top terms appear in Figure 7.

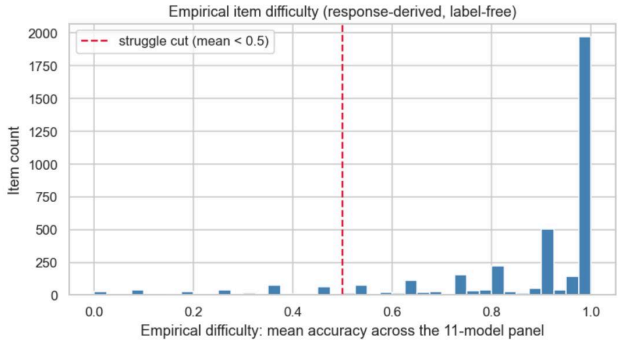

Figure 4: Empirical item difficulty: mean partial-credit accuracy across the eleven-model panel. Most items are saturated (mean near 1.0); the left tail past the dashed line marks the 337 majority-fail items analyzed in §4.2.

Item-fit infit and outfit mean squares follow the Wright/Linacre convention [Linacre, 2002]:

$$\text{Infit}_j = \frac{\sum_i (X_{ij} - P_{ij})^2}{\sum_i P_{ij}(1 - P_{ij})}, \tag{3}$$

$$\text{Outfit}_j = \frac{1}{n} \sum_i \frac{(X_{ij} - P_{ij})^2}{P_{ij}(1 - P_{ij})}. \tag{4}$$

Values outside $[0.5, 1.5]$ indicate misfit; values $\ll 0.5$ indicate overfit (model fits better than expected, diagnostic of the small-n regime, not item pathology).

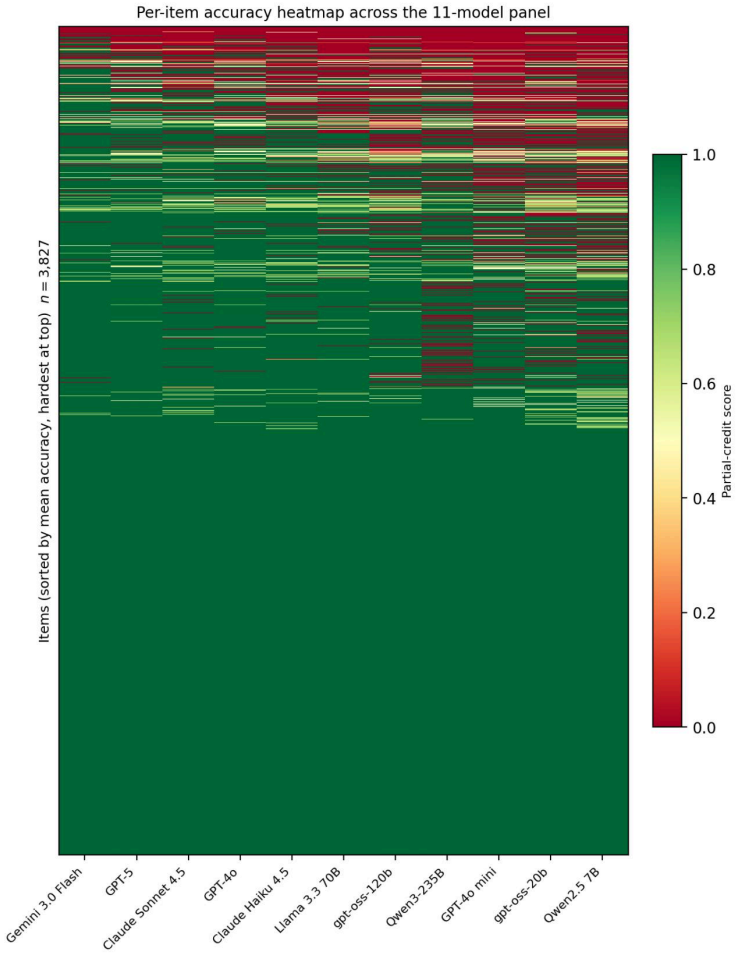

Figure 5: Per-item, per-model partial-credit accuracy across the eleven-model panel (n = 3,827 items). Rows are items sorted ascending by mean accuracy (hardest at top); columns are models sorted by overall accuracy. The thin red band at the top is the set of items the whole panel fails; the large green region below is the saturated majority. Color runs red (0) to green (1).

E Per-Publication Item Allocation

Items were allocated to each joint publication in proportion to its extracted word count, as described in §3. The table below lists all 68 publications, their word counts, share of the corpus, and item allocation, which sums to the full 3,827-item bank. Publication identifiers follow the extracted source-file names; chapter/volume suffixes (e.g. `ch1`, `vol2`) denote the specific change or volume used.

Table 6: Proportional item allocation across the 68 joint publications. Items were allocated to each publication in proportion to its extracted word count (total 3,827 items).

Pub.	Title	Words	Share %	Items
JP 3-0	Joint Campaigns and Operations	122,480	3.286	126
JP 3-09-3	Joint Close Air Support	116,801	3.133	120
JP 333ch2	Joint Force Headquarters	112,636	3.022	116

continued on next page

Table 6 continued from previous page

Pub.	Title	Words	Share %	Items
JP 3-50	Personnel Recovery	108,814	2.919	112
JP 2-0ch1	Joint Intelligence	102,624	2.753	105
JP 3-07	Joint Stabilization Activities	100,447	2.695	103
JP 3-08	Interorganizational Cooperation	105,181	2.822	108
JP 3-24	Counterinsurgency	96,679	2.593	99
JP 3-25	Joint Countering Threat Networks	97,954	2.628	101
JP 5-0	Joint Planning	91,488	2.454	94
JP 4-02	Joint Health Services	88,259	2.368	91
JP 3-02	Amphibious Operations	83,732	2.246	86
JP 4-09	Joint Distribution Operations	82,341	2.209	85
JP 3-29	Foreign Humanitarian Assistance	75,999	2.039	78
JP 4-10	Operational Contract Support	71,199	1.91	73
JP 3-22	Foreign Internal Defense	66,437	1.782	68
JP 3-31ch2	Joint Land Operations	68,786	1.845	71
JP 3-34	Joint Engineer Operations	66,402	1.781	68
JP 3-36	Joint Air Mobility and Sealift Operations	65,324	1.752	67
JP 3-04	Information in Joint Operations	64,870	1.74	67
JP 3-41	Chemical, Biological, Radiological, and…	61,742	1.656	63
JP 4-0ch1	Joint Logistics	64,440	1.729	66
JP 3-01ch1	Countering Air and Missile Threats	61,295	1.644	63
JP 3-06ch1	Joint Urban Operations	60,527	1.624	62
JP 3-16ch1	Multinational Operations	58,229	1.562	60
JP 3-42	Joint Explosive Ordnance Disposal	60,980	1.636	63
JP 3-68	Joint Noncombatant Evacuation Operations	60,469	1.622	62
JP 3-32	Joint Maritime Operations	57,327	1.538	59
JP 3-05	Special Operations	53,104	1.425	54
JP 3-11	Operations in Chemical, Biological, Radi…	48,269	1.295	50
JP 3-20ch1	Security Cooperation	50,867	1.365	52
JP 3-26	Joint Combating Terrorism	53,727	1.441	55
JP 3-28	Defense Support of Civil Authorities	51,528	1.382	53
JP 3-53	Joint Military Information Support Opera…	51,741	1.388	53
JP 335sig-v2	Joint Deployment and Redeployment Operat…	47,923	1.286	49
JP 3-09	Joint Fires	45,843	1.23	47
JP 3-10	Joint Security Operations in Theater	45,325	1.216	47
JP 3-23	Joint Peace Operations	44,943	1.206	46
JP 3-30	Joint Air Operations	43,280	1.161	44
JP 3-60ch1	Joint Targeting	43,921	1.178	45
JP 1vol2	The Joint Force	39,856	1.069	41
JP 3-07-4	Counterdrug Operations	40,065	1.075	41
JP 3-12	Joint Cyberspace Operations	39,612	1.063	41
JP 3-15	Barriers, Obstacles, and Mine in Joint O…	40,760	1.093	42
JP 4-05	Joint Mobilization Planning	42,498	1.14	44
JP 1-0	Joint Personnel Support	36,353	0.975	37
JP 1vol1	Joint Warfighting	38,679	1.038	40
JP 3-27	Joint Homeland Defense	36,821	0.988	38
JP 3-40	Joint Countering Weapons of Mass Destruc…	38,707	1.038	40
JP 3-54	Joint Doctrine for Military Deception	38,757	1.04	40
JP 3-80	Resource Management	37,384	1.003	38
JP 3-14	Joint Space Operations	34,047	0.913	35
JP 3-52	Joint Airspace Control	33,304	0.893	34
JP 3-03	Joint Interdiction	29,548	0.793	30
JP 3-18	Joint Forcible Entry Operations	29,480	0.791	30
JP 3-61	Joint Public Affairs	28,707	0.77	29
JP 3-85	Joint Electromagnetic Spectrum Operations	30,289	0.813	31
JP 4-03	Joint Bulk Petroleum and Water Doctrine	28,541	0.766	29
JP 6-0	Joint Communications	31,678	0.85	32
JP 3-57	Joint Civil-Military Operations	26,715	0.717	27
JP 3-59	Meteorological and Oceanographic Operati…	26,394	0.708	27
JP 3-84	Legal Support to Military Operations	25,076	0.673	26
JP 4-01	The Defense Transportation System	27,164	0.729	28
JP 4-04	Contingency Basing	27,605	0.741	28

continued on next page

Table 6 continued from previous page

Pub.	Title	Words	Share %	Items
JP 3-55	Joint Operations Security	22,043	0.591	23
JP 3-72	Joint Nuclear Operations	19,611	0.526	20
JP 3-0appd	Fundamentals of Joint All-Domain Operati…	12,434	0.334	13
JP 3-83	Religious Affairs in Joint Operations	11,738	0.315	12

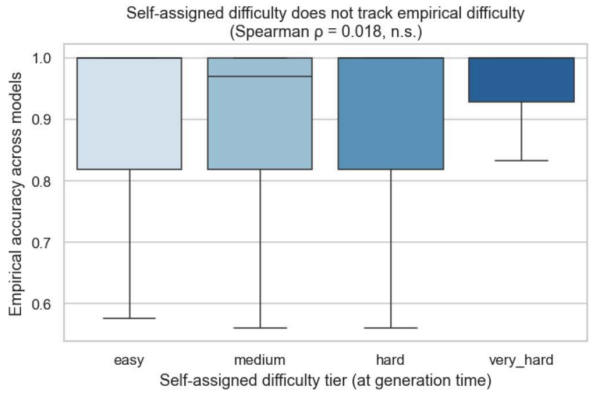

Figure 6: Self-assigned difficulty tier (x) versus empirical accuracy across the panel (y). The easy/medium/hard distributions are nearly identical (Spearman $\rho = 0.018$, n.s.); the generation-time labels are uninformative, motivating the response-derived difficulty used in §4.2.

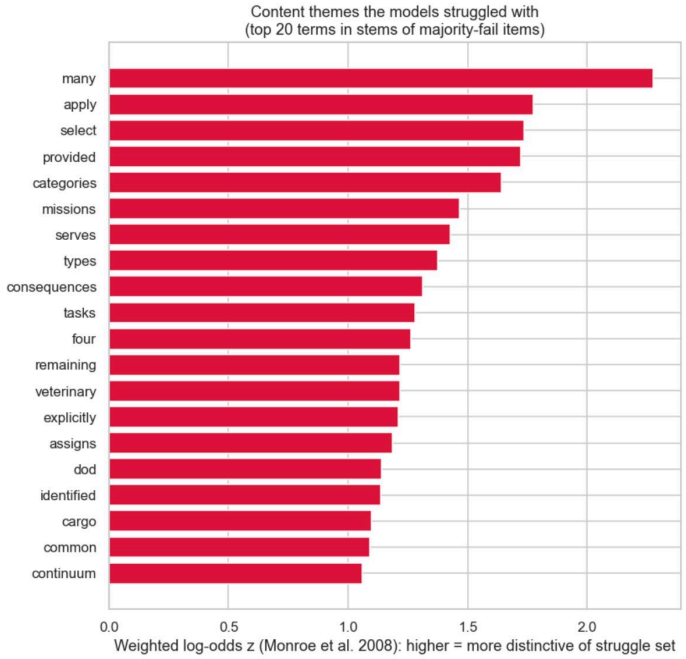

Figure 7: Content terms most distinctive of the majority-fail item stems (weighted log-odds z with informative Dirichlet prior, Monroe et al. 2008). Enumeration cues and logistics vocabulary dominate.

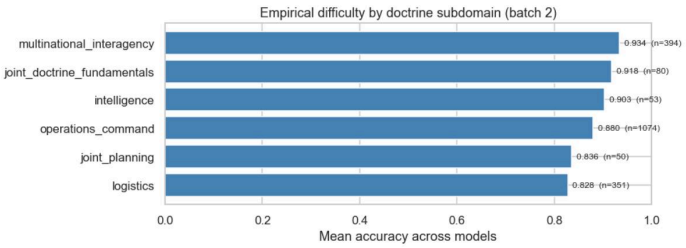

Figure 8: Empirical difficulty (mean accuracy across models) by doctrine subdomain. Logistics is the hardest subdomain and multinational/interagency the easiest.

Criterion Validity of General-Purpose LLM Benchmarks for Joint Military Domain Tasks Pre-Analysis Plan

Garrett Alarcon
CS321M: AI Evaluation and Measurement Science
Stanford University
gaa@stanford.edu

1 Overview and Motivation

General-purpose LLM benchmarks drive deployment decisions in defense and national-security contexts, yet no published study tests whether their scores predict performance on joint military operational tasks. Following the criterion validity framework in Truong and Koyejo [2026] (Section 6.3.2), benchmark scores must correlate with an external criterion that independently captures the construct—a requirement that becomes especially acute in AI evaluation, where no agreed-upon gold standard exists for many target capabilities. In medicine, where criterion validation has been attempted, Alaa et al. [2025] show that benchmark performance does not strongly predict real-world clinical performance: models that top leaderboards on MedQA do not reliably outperform others on matched real-world patient cases. The same question has never been asked of any military domain.

Research question. To what extent do general-purpose LLM benchmark scores (MMLU-Pro and GPQA-Diamond) predict performance on a Joint Military Domain (JMD) benchmark constructed from publicly available U.S. joint doctrine?

Three pre-registered hypotheses operationalize the question:

- **H1 (criterion validity):** Spearman ρ between JMD and MMLU-Pro / GPQA-Diamond will fall in $[0.4,\ 0.7]$ across the ten models, replicating the modest-correlation pattern observed in medicine, law, and code generation.
- **H2 (domain lift):** Government-tuned variants (Gemini 3.0 and 3.1 via genai.mil; Sonnet 4.5 and GPT-5 via Palantir Foundry) will outperform their base counterparts on JMD after controlling for MMLU-Pro, producing positive residuals in a JMD $\sim$ MMLU-Pro regression.
- **H3 (internal structure):** JMD items will show meaningful spread in 2PL difficulty (β) and discrimination (α) parameters, indicating the instrument is not saturated and supports model differentiation at the frontier.

Course connection. The benchmark construction draws on Chapter 2 (2PL IRT and factor models), Chapter 3 (parameter estimation), and the item-construction framework in Chapter 6 (Borsboom warrants at 6.1, validity taxonomy at 6.3, threats at 6.4, diagnostics at 6.5, and principled item construction at 6.6.2). Reliability and generalizability analysis draws on Chapter 5 (classical reliability at 5.2.3, G-theory at 5.4, and LLM-as-judge methodology at 5.5).

Scope. This is a solo project with a 16-day execution window. Items are drawn entirely from publicly available doctrine, removing any SME-scheduling bottleneck. Criterion benchmark scores are taken from published leaderboards, so no additional model evaluations are needed for the general benchmarks. The novel empirical work is limited to running 10 models on 200 JMD items, which is tractable given confirmed API access to all targets including the government-tuned variants.

2 Literature Review

Validity critiques of general benchmarks. Bean et al. [2025] reviewed 445 LLM benchmarks and found that almost all have construct-validity weaknesses: 21.7% fail to define the target phenomenon and only 16% conduct any statistical testing. Freiesleben [2026] argues that capability claims require nomological-network evidence, aligning with the framework in Truong and Koyejo [2026, Section 6.6.3]. These general validity concerns are acute for the criterion benchmarks used in this study. Wang et al. [2024] show that MMLU-Pro produces 16 to 33 point accuracy drops relative to the original MMLU and lower prompt sensitivity, which motivated its selection over the original. Zhao et al. [2024] document 14 to 16 points of contamination-driven score inflation in MMLU-family benchmarks, which motivates the inclusion of GPQA-Diamond as a less-contaminated secondary criterion.

Military and defense LLM evaluation. Existing work spans several sub-domains: MilBench [Ruiz and Sell, 2024] (Army doctrine), MilSCORE [Palnitkar et al., 2026] (geospatial planning), WAR-BENCH [Li et al., 2026] (tactical decision-making), COA-GPT [Goecks and Waytowich, 2024] (course-of-action generation), and the Rivera et al. and Lamparth et al. wargame escalation studies [Rivera et al., 2024, Lamparth et al., 2024]. None tests joint, multi-Service operational generalization, and none has been subjected to formal criterion validation. The military domain therefore represents the largest unaddressed gap in benchmark validation among high-stakes professional fields.

Cross-domain criterion validity. Medicine provides the closest template. Alaa et al. [2025] demonstrate that LLMs which top medical benchmark leaderboards do not reliably outperform others on matched real-world patient cases, showing that benchmark rankings can fail to generalize to operational performance. In law, Magesh et al. [2025] find that benchmark-validated AI legal research tools (Lexis+ AI and Westlaw AI-Assisted Research) still hallucinate on 17 to 33% of queries which is evidence that strong benchmark performance does not guarantee reliable real-world output. The pattern is consistent across domains: general-benchmark scores correlate weakly with operational outcomes, and the gap widens at the frontier where benchmarks saturate.

Gap and positioning. No published study correlates general-benchmark scores with operational performance in any military domain. Validity tools applied in medicine (MedIRT, Fluid Benchmarking) and the construct-validity framework of Bean et al. [2025] and Salaudeen et al. [2025] (textbook Section 6.6.3) have not been applied together in any military domain validation pipeline. Whether government domain fine-tuning produces measurable lift beyond general capability has not been tested. This project addresses all three gaps.

3 Proposed Methods and Analysis

3.1 JMD Benchmark Construction

Following the principled item-construction procedure in Truong and Koyejo [2026, Section 6.6.2]:

- **Construct definition.** JMD competence is defined as the ability to recognize, recall, and apply concepts, doctrine, terminology, and command relationships specified in U.S. joint publications, scoped to unclassified, publicly releasable content. The construct excludes Service-specific tactical knowledge, current intelligence, classified planning, and competence outside the doctrinal corpus.
- **Test blueprint.** 200 items allocated across six sub-domains: joint doctrine fundamentals (JP 1, JP 3-0) 50; joint planning (JP 5-0) 40; operations and command relationships (JP 3-0, JP 1-02) 40; intelligence (JP 2-0) 25; logistics (JP 4-0) 25; multinational and interagency (JP 3-08, JP 3-16) 20.
- **Item formats.** 120 single-answer multiple choice, 60 multiple-answer (partial credit $=$ proportion correct $-$ proportion incorrect, floored at zero), and 20 short-answer items scored by LLM-as-judge with a fixed rubric validated against author scoring on a 20-item subset (per Section 5.5 LLM-as-judge reliability guidance in Truong and Koyejo 2026).
- **Sourcing and quality.** All items are derived from joint publications, Joint Staff training materials, and unclassified CJCSI/CJCSM releases. Items are paraphrased from source

language to reduce contamination risk; per-item source citation is logged for later content-validity audit. Items with point-biserial correlation below 0.10 or 2PL discrimination $\alpha < 0.3$ are flagged; items with negative discrimination are removed.

3.2 Models Under Study

Ten models in three tiers are evaluated (Table 1).

Table 1: Models under study

Tier	Models	Access
Frontier (base)	GPT-5, Claude Sonnet 4.5, Gemini 3.0, Gemini 3.1	Vendor APIs
Government-tuned	Gemini 3.0 and 3.1 (genai.mil); Sonnet 4.5 and GPT-5 (Palantir Foundry)	genai.mil, Foundry
Open-source	Llama 3.3 70B Instruct, Qwen 2.5 72B Instruct	Inference API

3.3 Criterion Benchmarks

MMLU-Pro serves as the primary criterion (broad professional knowledge, harder than MMLU, lower prompt sensitivity per Wang et al. 2024). GPQA-Diamond serves as the secondary criterion (graduate-level reasoning, lower contamination risk than MMLU-family benchmarks). Using two criterion benchmarks tests convergent validity of the JMD instrument across both knowledge-recall and reasoning constructs. Scores are taken from published evaluations (model cards, technical reports, Open LLM Leaderboard, Artificial Analysis), with all reported figures versioned and cited in the final manuscript.

3.4 Analytical Approach

Criterion validity (primary analysis). The primary analysis computes Spearman ρ between JMD total scores and each criterion benchmark (MMLU-Pro and GPQA-Diamond separately) across the ten models. Pearson r is reported as a secondary effect size. Because $n = 10$ limits statistical power, the analysis is framed as exploratory; results are interpreted descriptively against the pre-registered H1 range rather than against a significance threshold. Confidence intervals are 95% bias-corrected and accelerated (BCa) bootstrap with 10,000 resamples.

Domain lift (H2). To test whether government fine-tuning adds domain-specific capability beyond what general benchmarks predict, residuals are computed from the regression JMD $\sim$ MMLU-Pro for each of the four base-vs-tuned model pairs. Positive residuals for the tuned variants would constitute evidence of domain lift. A paired Wilcoxon signed-rank test is applied across the four pairs, though with $n = 4$ the result is interpreted descriptively rather than inferentially.

Internal structure (H3). A 2PL IRT model is fit to the full 10×200 response matrix via marginal maximum likelihood. Item difficulty (β) and discrimination (α) distributions are reported to characterize the instrument. Parallel analysis per Section 6.5.2 of Truong and Koyejo [2026] is used to test whether a single dominant dimension underlies JMD performance, which is a precondition for valid total-score interpretation.

Reliability. Cronbach α and Spearman-Brown corrected split-half reliability are computed for the full JMD item set. If time permits within the analysis window, a fully crossed items $\times$ models G-study will be run to decompose score variance into model, item, and interaction components and produce a generalizability coefficient per Section 5.4.3 of Truong and Koyejo [2026]. For the short-answer subset, Cohen's κ is computed between the LLM judge and author scores on the 20-item validation set, following the LLM-as-judge guidance in Section 5.5.

Validity diagnostics. Differential Item Functioning is screened across the three model tiers using Mantel-Haenszel [Truong and Koyejo, 2026, Section 6.5.1] and Zumbo logistic regression [Zumbo, 1999] to identify items that behave differently for frontier, government-tuned, or open-source models.

Potential contamination is assessed via item-fit statistics (infit and outfit mean squares) on the response matrix per Section 6.5.3 of Truong and Koyejo [2026].

3.5 Metrics, Data, and Code

The primary validity metric is Spearman ρ between JMD and each criterion benchmark, reported with 95% BCa bootstrap confidence intervals (10,000 resamples); Pearson r is reported as a secondary effect size. At the item level, 2PL discrimination (α) and difficulty (β) distributions are reported alongside item information curves for the highest-discriminating items. Reliability is assessed via Cronbach α (target ≥ 0.80) and Spearman-Brown corrected split-half, with a generalizability coefficient added if the G-study completes within the timeline. DIF is flagged using the Mantel-Haenszel log-odds-ratio at a $|\log \text{OR}| > 0.41$ threshold, and dimensionality is assessed via a parallel-analysis eigenvalue plot.

All JMD items, model responses, judge prompts, scoring rubrics, and analysis code will be deposited in a public repository alongside the final manuscript. Per-item provenance is logged in a CSV file. IRT estimation uses `py-irt`; broader analysis uses Python with `torch_measure` where applicable. The full pipeline is reproducible from raw response matrices to all reported figures and tables.

4 Timeline, Risks, and Responsibilities

4.1 Week-by-Week Milestones (Solo Project)

Table 2: Project timeline

Window	Milestone	Output
May 11	Pre-Analysis Plan submission	This document
May 12–18	Doctrine sourcing; blueprint freeze; drafting to 200 items; judge rubric v1	Full 200-item pool
May 19–20	Run 10 models on JMD; LLM-judge scoring of short-answer subset	10×200 response matrix
May 21–23	Reliability, IRT, DIF, dimensionality; criterion-validity and residual analysis	Result tables and figures
May 24–27	Manuscript drafting (NeurIPS format, 8 pages); final submission	Final paper
May 28–Jun 3	Code cleanup, reproducibility verification, repository finalization	Code deliverable

4.2 Risks and Mitigations

Table 3: Identified risks and mitigations

Risk	Mitigation
Item quality without SME review	Authoritative doctrine only; dual-source key facts; pilot-and-revise per Section 6.6.2; drop items with $\alpha < 0.3$ or negative discrimination.
Small model sample ($n = 10$)	Exploratory framing; BCa bootstrap CIs; rank-based statistics; pre-registered effect-size range rather than significance test.
MMLU-Pro contamination in newer models	GPQA-Diamond as less-contaminated secondary criterion; report both correlations; contamination-adjusted interpretation discussed in manuscript.
Access loss on government-tuned variants	Front-load gov-tuned runs May 19–20; if access is lost, demote H2 to a stretch goal and proceed with H1 and H3 on remaining models.
Judge or schedule slippage	If Cohen's $\kappa < 0.6$, drop short-answer items from the primary analysis. Blueprint frozen May 14; fallback to 150 items with the same sub-domain proportions if drafting slips; document any deviation in the manuscript.

References

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical Large Language Model Benchmarks Should Prioritize Construct Validity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80991–81004. PMLR, 2025.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. NeurIPS 2025 Datasets and Benchmarks Track. arXiv:2511.04703.

Timo Freiesleben. Establishing construct validity in LLM capability benchmarks requires nomological networks. *arXiv preprint arXiv:2603.15121*, 2026.

Vinicius G. Goecks and Nicholas Waytowich. COA-GPT: Generative pre-trained transformers for accelerated course of action development in military operations. *arXiv preprint arXiv:2402.01786*, 2024.

Max Lamparth, Niklas Zuber, Anka Reuel, Friederike Moeller, Jacquelyn Schneider, and Hendrick Strauss. Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. *arXiv preprint arXiv:2403.03407*, 2024.

Zongjie Li, Chaozheng Wang, Yuchong Xie, Pingchuan Ma, and Shuai Wang. WARBENCH: A comprehensive benchmark for evaluating LLMs in military decision-making. *arXiv preprint arXiv:2603.21280*, 2026.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 2025. Published April 23, 2025.

Aadi Palnitkar, Mingyang Mao, Nicholas Waytowich, Vinicius G. Goecks, and Xiaomin Lin. MilSCORE: Benchmarking long-context geospatial reasoning and planning in large language models. *arXiv preprint arXiv:2601.21826*, 2026.

Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 836–898, 2024. arXiv:2401.03408.

Daniel C. Ruiz and John Sell. Fine-tuning and evaluating open-source large language models for the army domain. *arXiv preprint arXiv:2410.20297*, 2024.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

S. Truong and S. Koyejo. *AI Measurement Science: A Science of Knowing Where AI Thrives, Where It Breaks, and How to Respond*. Stanford University, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. NeurIPS 2024. arXiv:2406.01574.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. MMLU-CF: A contamination-free multi-task language understanding benchmark. *arXiv preprint arXiv:2412.15194*, 2024. Published at ACL 2025.

B. D. Zumbo. *A Handbook on the Theory and Methods of Differential Item Functioning*. National Defense Headquarters, Ottawa, 1999.