Overview of the ClinIQLink 2025 Shared Task on Medical Question-Answering

Brandon C. Colelough, Davis Bartels, and Dina Demner-Fushman
National Library of Medicine, National Institutes of Health
Bethesda, MD, USA
{firstname.lastname}@nih.gov

Abstract

In this paper, we present an overview of CLINIQLINK a shared task, collocated with the 24th BioNLP workshop at ACL 2025, designed to stress-test large language models (LLMs) on medically-oriented question answering aimed at the level of a General Practitioner. The challenge supplies 4 978 expert-verified, medical source-grounded question–answer pairs that cover seven formats - *true/false*, *multiple choice*, *unordered list*, *short answer*, *short-inverse*, *multi-hop*, and *multi-hop-inverse*. Participating systems, bundled in Docker or Apptainer images, are executed on the CodaBench platform or the University of Maryland's *Zaratan* cluster. An automated harness (Task 1) scores closed-ended items by exact match and open-ended items with a three-tier embedding metric. A subsequent physician panel (Task 2) audits the top model responses.

1 Introduction

LLMs have increasingly demonstrated their ability to memorize information and answer questions (Carlini et al., 2023). This has led to their increased use by consumers to ask medically relevant questions (Yun and Bickmore, 2025). However, LLMs have been shown to "hallucinate", that is, to generate factually incorrect, or even harmful answers (Singhal et al., 2023). In high-stakes domains, such as medicine, it is incredibly important to be able to evaluate the veracity of any question answering system. While there exist datasets, such as MultiMedQA (Singhal et al., 2023), designed to do just this, recent LLMs have been trained over their data. This limits their usefulness in evaluating the ability of these models to generalize to out-of-distribution data.

New datasets are necessary for the evaluation of medical question-answering systems and new systems are needed to increase accuracy and mitigate hallucinations.

To this end, we introduce the ClinIQLink shared task, inviting participants to submit question-answering systems to be evaluated on a novel dataset of medical questions. Participants are encouraged to submit systems that are capable of demonstrating medical knowledge, while mitigating hallucinations. Our dataset consists of seven question types, both closed and open ended, and a wide range of medical topics. Our task had a total of three runs from one team. Our contributions are as follows:

- A dataset of 4,978 vetted medical question-answer pairs
- Automated evaluation metrics
- A task design for participant-submitted systems
- A physician audit of system responses

2 Task Description

ClinIQLink [1] is a shared task that evaluates the ability of generative models to produce factually accurate medical information aimed at the knowledge level of a general practitioner. The submitted systems are executed in a containerized environment on CodaBench [2] or via the University of Maryland (UMD) HPC Zaratan [3] (depending on the size and model/system complexity), where the submitted systems answered a corpus of expert-curated atomic medical questions. Answers provided from the systems submitted were judged only on factual accuracy, so leaderboard ranking reflects a model's ability to retrieve correct information from its own parametric memory or any retrieval mechanism the team elected to integrate.

[1] https://cliniqlink.org/
[2] https://www.codabench.org/
[3] https://hpcc.umd.edu/hpcc/zaratan.html

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 378–387

The question sets were divided into two types (closed and open-ended QA pairs) and spanned seven modalities, including true/false, multiple choice, unordered list, short answer, short-inverse, multi-hop and multi-hop-inverse. Across all of the seven QA pair modalities, the ground truth was anchored in standard open-source medical texts, and each item targets a single, clearly defined concept such as a procedure, drug, diagnostic finding, or anatomical fact.

The challenge comprised two sequential components. Task 1 executed all baseline systems and participant submissions within our automated benchmarking harness. The script marked closed-ended items strictly for precision and evaluated open-ended answers with a semantic-similarity module that awards full or partial credit according to their closeness to the hidden ground-truth. Leaderboard rankings are derived solely from these automatic scores. Task 2 began after the leaderboard was frozen: a panel of human-expert annotators reviewed the highest-scoring outputs, ranking them from best to worst and annotating each answer on a spectrum from "good" to "bad". Participants were allowed to employ any architecture, external knowledge base, or retrieval-augmented pipeline to generate answers to questions posed, provided the final system can run end-to-end inside the supplied containerised harness. Teams were limited to three leaderboard submissions and were required to accompany their final entry with a short paper that details model design, data usage, and inference strategy for inclusion in the BioNLP 2025 proceedings. The full evaluation dataset remains private to preserve its viability for later use.

3 Dataset Description

3.1 Generation and Vetting

A neuro-symbolic pipeline was employed to produce roughly $\sim$20K atomic question–answer pairs from open-source medical texts. Each pair was linked to its supporting passage so that later reviewers could verify every biomedical fact. The QA Pairs were then ported to our online annotation portal[4], (which is now open to accredited medical schools and hospitals who wish to contribute further judgments), where human-experts (paid medical students) confirmed correctness, rated *general-practitioner (GP) relevance* on a five-point scale, and could file structured feedback or formal disputes.

3.2 Human-verification Workflow

1. **Primary review**: an expert validated factual accuracy against the source excerpt, assigned a GP-relevance score, and could flag issues or supply comments.

2. **Secondary review**: $\sim 45\%$ of items received an independent second pass; disagreements triggered adjudication. By 1 May 2025 reviewers had lodged 601 feedback notes and 461 disputes. The 1062 QA Pairs that had been flagged as feedback or disputes were not used for testing and are presently still being held for later review.

3.3 Benchmark Snapshot (1 May 2025)

At the dataset freeze the repository contained $5{,}118$ verified QA pairs (Table 1): $5{,}118$ had a single expert judgement and $2{,}505$ were double-annotated. For leaderboard scoring, we retained only the $4{,}978$ items rated maximally relevant ($score = 5$); 140 lower-relevance items were set aside for future analysis. The sample dataset plus the full evaluation architecture are available at [5].

3.4 Question Modalities

Seven formats cover both machine-gradable *closed-ended* items and semantically scored *open-ended* prompts:

- **Closed-ended**
  - True/False (TF)
  - Multiple Choice (MC) — single-best answer
  - Unordered List (LIST) — enumerate all correct elements

- **Open-ended**
  - Short Answer (SHORT) — concise factoid
  - Short-Inverse (SHORT_INV) — explain why the supplied wrong answer is incorrect
  - Multi-hop (MULTI_HOP) — required several leaps in knowledge to arrive at a final answer; models must return answer *and* knowledge leaps
  - Multi-hop Inverse (MULTI_HOP_INV) — locate the faulty step in a provided, erroneous multi-hop rationale

[4] https://bionlp.nlm.nih.gov/ClinIQLink/NIHLogin

[5] https://github.com/Brandonio-c/ClinIQLink_Sample-dataset

Table 1: ClinIQLink benchmark composition at the baseline freeze (1 May 2025). "High" denotes GP-relevance score 5 items used for leaderboard evaluation; "Low" items (< 5) were withheld.

QA Format	Counts			Subset with Two Independent Reviews			
	High	Low	Total	High	Low	Total	Percent Double
True/False (TF)	813	38	851	369	–	369	43.4%
Multiple Choice (MC)	765	29	794	346	–	346	43.6%
Unordered List (LIST)	714	28	742	341	–	341	46.0%
Short Answer (SHORT)	427	9	436	339	–	339	77.8%
Short-Inverse (SHORT_INV)	742	16	758	353	–	353	46.6%
Multi-hop (MULTI_HOP)	771	8	779	331	–	331	42.5%
Multi-hop Inverse (MULTI_HOP_INV)	746	12	758	318	–	318	42.0%
Totals	**4 978**	**140**	**5 118**	**2 497**	–	**2 505**	**48.8%**

4 Evaluation Protocol

Our assessment of CLINIQLINK was conducted in two sequential phases. First, we relied on a fully automated evaluation script (Task-1) that ingested model/participant system responses; second, we complemented the automated evaluation with an expert preference study (Task-2) in which paid medical students compared top-performing model responses.

4.1 Task-1: automatic scoring

Each submission returned answers for **seven distinct question classes**. *True/False* and single-best *multiple-choice* items were judged by straightforward **accuracy**

$$\text{Accuracy} \;=\; \frac{\#\text{correct}}{N},$$

whereas *multiple select list* questions were graded with both macro- and micro F_1 (Manning et al., 2008):

$$F_1^{\text{macro}} = \frac{1}{N}\sum_{i=1}^{N} F_1^{(i)},$$

$$F_1^{\text{micro}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}.$$

All free-text tasks (`short`, `multi hop`, and their inverse variants) were assessed twice; once with the ClinIQLink **semantic-similarity** score and again with the conventional n-gram metrics BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

ClinIQLink semantic-similarity score. The score blended *three* complementary cosine layers:

(1) **Token layer:** an IDF-weighted, greedy token-alignment F_1, rewarding exact overlap on infrequent clinical terms.

(2) **Sentence layer:** cosine similarity of SBERT–MINILM [6] CLS embeddings, capturing broader paraphrase.

(3) **Paragraph layer:** cosine similarity of the raw answer strings, offering global context.

Let $C_{\text{tok}}, C_{\text{sent}}, C_{\text{para}} \in [0,1]$ denote these three cosines. With weights $w_{\text{tok}} = w_{\text{sent}} = 0.4$ and $w_{\text{para}} = 0.2$ the raw score is

$$S_{\text{raw}} \;=\; 0.4\,C_{\text{tok}} + 0.4\,C_{\text{sent}} + 0.2\,C_{\text{para}}.$$

Because SBERT assigns unrelated sentence pairs a baseline similarity of about $\beta = 0.25$, we subtract that offset, floor negatives, and snap near-perfect matches:

$$S = \min\Big(1,\; \max\big(0,\; S_{\text{raw}} - \beta\big)\Big),$$

$$S \geq 0.95 \implies S := 1.$$

Penalty for `multi-hop inverse`. If a model highlighted the wrong reasoning step, the semantic score was down-weighted. Let $d = |\,\text{predicted step} - \text{gold step}\,|$ be the absolute distance; then

$$\alpha(d) = \begin{cases} 1 & d = 0, \\ 0.7 & d = 1, \\ 0.3 & d = 2, \\ 0.3\,2^{-(d-2)} & d \geq 3, \end{cases} \qquad S^* = \alpha(d)\,S.$$

Hence, the final similarity S^* combined graded lexical alignment, distributional semantics, and explicit reasoning correctness, while conventional

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

BLEU/ROUGE/METEOR offered secondary diagnostics. A complete implementation of the evaluation script that implements the above can be found with the testing harness[7].

4.2 Task-2: expert preference study

We found that the automated metrics employed for analysis of the open-ended QA pairs were not effective for evaluation of model responses, nor were they effective in discriminating top-ranking model responses from mediocre model responses. Hence, to complement the automated evaluation metrics we organized a human evaluation in which we required our annotators to rank the six strongest foundation models on our public leaderboard (Falcon-10B, Llama-3.3-70B, Llama-4 Scout, Mistral-Large-2411, Microsoft Phi-4 Base, Qwen-3-32B) together with the best participant submission, *Preceptor-AI* and the ClinIQLink ground-truth answers. For every question we shuffled these seven model responses plus the ClinIQLink dataset reference solution, and asked human annotators to rank them from best to worst. Each answer also received a coarse quality tag (good/okay/bad). The annotation portal that we built for this experiment is now open to accredited medical schools and hospitals who wish to contribute further judgments [8].

5 Baseline Systems

To provide a strong reference point for future work we evaluated a broad range of publicly–available large language models on the frozen CLINIQLINK test split. For transparency, it should be noted that the LLM utilised as the "neuro" component of our neurosymbolic pipeline for data generation was **Llama 3.3-70B–Instruct**. All baseline checkpoints were used as-is and therefore reflect their pre-training and instruction-tuning quality rather than any task-specific fine-tuning.

5.1 Llama family.

The *Meta Llama 3* (Grattafiori et al., 2024) decoder-only transformer was represented by four parameter scales; 1B, 3B, 8B and 70B weights as well as an intermediate commercial variant (`llama_4-scout`, $\approx$ 45B). All are dense models built with a 32-layer architecture (70B: 80 layers) and grouped-query attention; the instruction checkpoints add a supervised fine-tuning and reinforcement learning step to the base weights.

5.2 Mistral / Mixtral family.

We included the 7-billion-parameter **Mistral-7B** (Jiang et al., 2023) dense decoder and the **Mistral-Large-Instruct-2411** release (8×22B experts, two experts routed per token, giving 47B active parameters). The **Mixtral** series consisting of Mixtral-8×7B (Jiang et al., 2024) and Mixtral-8×22B tested share the same sparse Mixture-of-Experts (MoE) scaffold, however, only two of the eight experts are selected for each input token, keeping inference costs close to their 12–13 B dense peers while exposing $>$ 140B total capacity.

5.3 Qwen3 family.

Alibaba's *Qwen3* (Yang et al., 2025) decoder stack (RoPE positional encoding, grouped-query attention) was tested at five scales: 1.7B, 3B, 4B, 8B, and 32B parameters. All checkpoints were released under an open-source licence together with alignment ("`-Instruct`") variants that follow the Supervised Fine-Tuning (SFT) + Direct Preference Optimisation recipe.

5.4 Phi family.

We evaluated Microsoft's **Phi-4** (Abdin et al., 2024) ($\sim$ 14B dense decoder) and its lightweight derivatives (`phi-4-mini-instruct` and `phi-4-mini-reasoning` (Abdin et al., 2025), $\sim$ 3.8B). This family of LLMs was designed as "small-data curriculum models" whose pre-training is dominated by synthetic textbook-style content rather than filtered web corpora.

5.5 Falcon Family

For completeness, we benchmarked **Falcon-10B-Instruct** (Almazrouei et al., 2023), an Apache–2.0 decoder model trained on the RefinedWeb dataset and alignment-tuned with RLHF.

5.6 Google Flan family.

Encoder–decoder baselines were covered by **Flan-T5-XXL** (Chung et al., 2022) (11B parameters) and **Flan-UL2** (Tay et al., 2023) (20B). Both models extend the original T5/UL2 sequence-to-sequence architecture with instruction tuning on a curated

[7] https://github.com/Brandonio-c/ClinIQLink_CodaBench_docker-setup/blob/main/submission/evaluate.py

[8] https://bionlp.nlm.nih.gov/ClinIQLink2/NIHLogin

mixture of over one thousand NLP tasks; an additional `attrscore_flan_t5_xxl` (Yue et al., 2023) checkpoint was tested, which augments the T5-XXL Weights with token-level attribution heads for explanation capabilities.

6 Participants and Methods

The CLINIQLINK shared task was publicly released through the Codabench evaluation platform[9], with an accompanying containerized setup for local validation and submission via Docker and Apptainer[10]. Submissions of models/systems over 10GB in size and requiring more compute than what is offered via codabench were also enabled via direct submission to organizers to be run on the University of Maryland HPC Zaratan. In total, 43 participants registered for the challenge during the initial release window. The competition remains open for new submissions on Codabench for smaller models that can run via the Codabench platform.

6.1 Preceptor AI

Although forty-three teams registered, only PRECEPTOR AI submitted runnable systems. They provided three containerised runs, `v001`, `v002`, and `v003`, but discuss only `v001` in their participant paper.

v001 – VeReaFine (Verifier-augmented RAG). `v001` is an iterative, evidence-seeking pipeline that couples a Qwen-7B-Instruct generator with a separately fine-tuned Qwen-8B medical-reasoning verifier. For each question the system:

1. retrieves up to 20 passages from a ColBERT (Khattab and Zaharia, 2020) + BM25 (Robertson et al., 2009) hybrid index built over *PubMed* abstracts and *StatPearls*;
2. drafts an answer *with inline citations*;
3. scores every generated claim with the verifier's token-level entailment head;
4. if any claim falls below a 0.8 confidence threshold, expands the evidence pool and repeats steps (1)–(3) (max. four rounds).

The loop stops when all claims are verified or the round limit is reached, after which the final answer and citation list are emitted. This design yields strong gains on all four open-ended modalities (top-10 P75 recall) but was *not* tuned for the closed-ended formats, explaining its low rank on multiple-choice and true/false items (see Table 2).

v002 and v003. The team also submitted `v002` (a retrieval-free Qwen-32B classifier optimised for closed-ended questions) and an ablation run `v003`. Because their accompanying paper focuses on the verifier-augmented strategy, only `v001` is analysed in detail there; we include the headline numbers for all three runs in the leaderboard for completeness.

7 Results

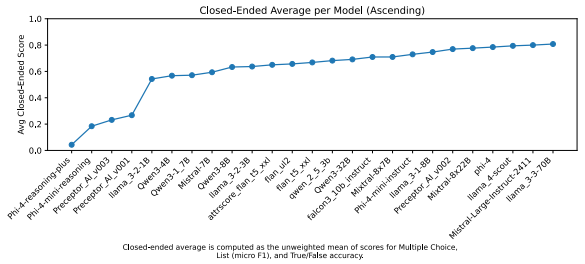

Figure 1: Average performance on closed-ended tasks (True/False accuracy, multiple-choice accuracy and list F_1).

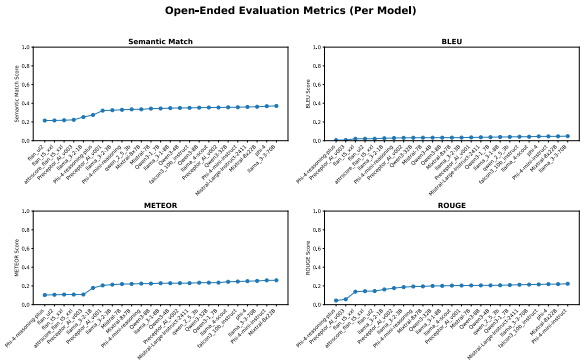

Figure 2: Distributions of individual n-gram scores (BLEU, ROUGE, METEOR) and semantic similarity for each open-ended question type.

Figure 1 summarised mean performance on the three closed-ended tasks. The spread between True/False, multiple-choice and list accuracy was modest, indicating that the leading models handled discrete answer formats with broadly comparable competence.

Open-ended behaviour was more nuanced. The per-task distributions in Figure 2 showed markedly heavier tails for semantic-similarity than for surface n-gram metrics, confirming that several systems produced answers that were lexically novel yet semantically similar. This pattern was especially

[9] https://www.codabench.org/competitions/5117/

[10] https://github.com/Brandonio-c/ClinIQLink_CodaBench_docker-setup

Table 2: ClinIQLink leaderboard snapshot (higher is better). Models were evaluated across all seven modalities of the ClinIQLink challenge. All models were retrieved from their public Hugging Face repositories, except for `Preceptor_AI`, which is private.

Rank	Model	Overall	MC Acc	TF Acc	List F1	Short	S-Inv	MHop	MH-Inv
1	`llama_3-3-70B`	0.541	0.796	0.822	0.682	0.235	0.488	0.313	0.450
2	`Mistral-Large-Instruct-2411`	0.530	0.797	0.822	0.645	0.260	0.472	0.313	0.398
3	`phi-4`	0.528	0.775	0.790	0.658	0.229	0.493	0.311	0.440
4	`llama_4-scout`	0.524	0.776	0.822	0.652	0.238	0.492	0.302	0.388
5	`Mixtral-8x22B`	0.521	0.752	0.800	0.643	0.227	0.491	0.306	0.428
6	`Preceptor_AI_v002`	0.512	0.762	0.817	0.583	0.213	0.479	0.298	0.430
7	`llama_3-1-8B`	0.499	0.720	0.765	0.613	0.223	0.479	0.293	0.396
8	`Phi-4-mini-instruct`	0.498	0.672	0.745	0.636	0.222	0.485	0.299	0.424
9	`falcon3_10b_instruct`	0.482	0.673	0.760	0.538	0.219	0.487	0.302	0.396
10	`Qwen3-32B`	0.477	0.737	0.803	0.373	0.233	0.474	0.307	0.415
11	`Mixtral-8x7B`	0.474	0.656	0.750	0.570	0.213	0.472	0.304	0.353
12	`qwen_2_5_3b`	0.461	0.629	0.726	0.535	0.216	0.484	0.292	0.347
13	`Qwen3-8B`	0.454	0.722	0.748	0.293	0.223	0.477	0.316	0.397
14	`llama_3-2-3B`	0.436	0.502	0.733	0.517	0.200	0.463	0.271	0.369
15	`Mistral-7B`	0.427	0.425	0.701	0.491	0.216	0.483	0.295	0.378
16	`Qwen3-4B`	0.423	0.515	0.752	0.294	0.212	0.470	0.310	0.408
17	`Qwen3-1_7B`	0.419	0.393	0.681	0.484	0.206	0.483	0.299	0.390
18	`flan_t5_xxl`	0.390	0.599	0.705	0.558	0.220	0.420	0.220	0.005
19	`flan_ul2`	0.383	0.567	0.695	0.556	0.205	0.430	0.223	0.003
20	`attrscore_flan_t5_xxl`	0.383	0.571	0.680	0.552	0.214	0.428	0.227	0.005
21	`llama_3-2-1B`	0.354	0.379	0.610	0.477	0.181	0.450	0.269	0.111
22	`Preceptor_AI_v001`	0.295	0.047	0.713	0.021	0.163	0.482	0.277	0.363
23	`Phi-4-mini-reasoning`	0.249	0.095	0.068	0.256	0.196	0.456	0.281	0.389
24	`Preceptor_AI_v003`	0.221	0.000	0.581	0.074	0.111	0.286	0.233	0.263
25	`Phi-4-reasoning-plus`	0.167	0.000	0.000	0.070	0.206	0.470	0.290	0.135

pronounced for the `multi-hop` and `multi-hop inverse` questions, where BLEU occasionally under-estimated quality relative to the embedding-based score.

To illustrate model-specific traits, Figures 3–5 present the full metric dashboards for three representative baselines. The FLAN-UL2 run exhibited tight clustering around mid-range similarity values and an extreme outlier for the multi-hop inverse modality.

LLAMA-3 70B displayed a broader inter-quartile range on semantic scores but maintained competitive n-gram fidelity, suggesting flexible paraphrasing capabilities.

Similarly, the PHI-4-REASONING-PLUS submission produced a long tail of semantically similar scores when evaluated with the CLinIQLInk semantic similarity metric, but low scoring across all the n-gram scoring metrics utilised; further inspection of the model responses revealed that, despite using the prescribed stop tokens and output template, the model frequently emitted extensive chain-of-thought traces capped by an ambiguous or missing "final answer" cue. Our automated evaluation script extracted only the required answers utilising pre-determined queues (i.e. the prompt templates used explicitly constrained models to provide list-type responses as comma-separated lists, etc.) and as such, the digressions observed from the Phi-4-Reasoning-Plus (amongst others) translated into poor task compliance rather than genuine comprehension deficits.

A consolidated leaderboard is provided in Table 2. The ranking served solely as an empirical reference from the evaluation metrics gathered from task 1 automated evaluation script.

8 Discussion

8.1 Closed-ended tasks (Figure 1).

Table 2 confirms what is visually apparent in the right-hand side of Figure 1, which is that single-labelled questions (e.g., t/f, MC, etc.) are close to saturation for modern LLMs. The top five systems tested (`llama_3-3-70B`, `Mistral-Large`, `phi-4`, `llama_4-scout`, and `Mixtral-8x22B`) all scored between 0.75 – 0.80 on **multiple-choice** and 0.79 – 0.82 on **True/False**. By contrast, **list** questions remained challenging with macro–micro F_1 not exceeding 0.68. List answers required both recognition of all correct options *and* rejection of distractors, and as such, the metric penalised even minor hallucinations; consequently, models whose generation style tended to "hedge" with ex-

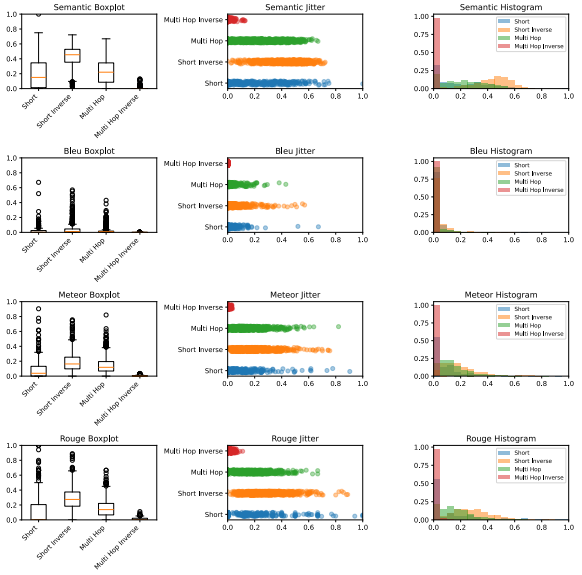

Figure 3: Comprehensive dashboard for FLAN-UL2 showing boxplots, jitter plots and histograms across semantic and n-gram metrics.

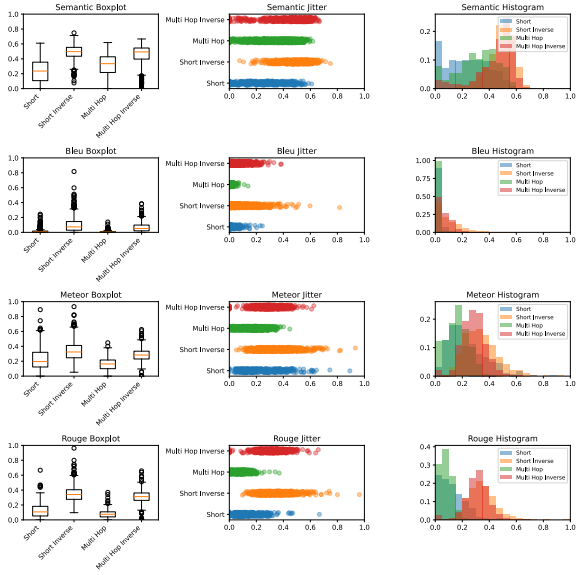

Figure 4: Comprehensive dashboard for LLaMA-70B showing boxplots, jitter plots and histograms across semantic and n-gram metrics.

tra choices (e.g. `falcon3_10b_instruct`) underperformed relative to their multiple-choice score. The tight inter-quartile ranges on True/False and multiple-choice further suggest that most contemporary LLMs share a common ceiling on purely factual one-shot classification, leaving little room for architectural distinctions to distinguish in these settings.

8.2 Aggregate open-ended behavior

Figure 2 shows that, across all models evaluated, the `short inverse` distributions peaked around 0.50 semantic similarity, while the forward `short` items clustered near 0.25, indicating that simply *critiquing* providing an answer was easier than generating an answer from scratch. The gap between semantic and n-gram scores widens for larger checkpoints. Mixtral-8×22B and LLaMA 3.3 70B frequently achieved high semantic similarity scores (above 0.60) despite very low BLEU scores (below 0.1), indicating that their correct answers were often paraphrased rather than copied verbatim, supporting the long-tailed distribution of paraphrastic responses seen in Figure 2. Inspection of the model responses for multi-hop inverse QA types also revealed answers that often diagnosed the wrong knowledge hop step, which in turn attracted the multiplicative penalties. Traditional n-gram metrics failed to flag these omissions, underscoring the necessity of the custom semantic evaluation platform.

8.3 Model-specific open-ended evaluations

Figures 3–5 illustrate how aggregate patterns materialised at the system level. The model-specific open-ended evaluations are shown for only the highest performing model across the board (llama-3.3 70B and the lowest performing model for closed and open-ended metrics (Phi-4-reasoning and FLAN-UL2, respectively).

- **FLAN-UL2** Figure 3 reveals that FLAN-UL2's outputs cluster tightly between 0.20 and 0.60 for the three forward-facing open-ended tasks, yet its `multi-hop inverse` scores collapse toward the origin on *all four* axes—semantic similarity, BLEU, ROUGE, and METEOR rarely rise above 0.05. The dashboard traces that floor effect to the model's habit of supplying only a step label (e.g., "Step 5") with no explanatory text, which earns minimal credit under the step-penalised rubric. Elsewhere, list questions are answered with bare option letters (e.g. "B, C, D"), boosting recall but cutting precision to roughly 0.33–0.50, while short prompts receive one or two-word noun phrases, driving n-gram metrics to zero even when the semantics are acceptable. These abrupt, template-bound behaviours keep variance low and prevent catastrophic errors, but they also cap the weighted open-ended average at 0.14 and hold FLAN-UL2 in 18$^{\text{th}}$ place despite competent

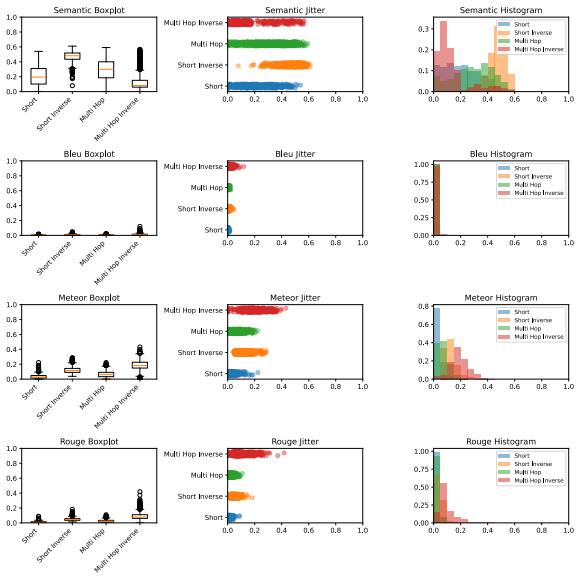

Figure 5: Comprehensive dashboard for Phi-4-reasoning-plus showing boxplots, jitter plots and histograms across semantic and n-gram metrics.

closed-ended performance.

- **LLaMA-3-3-70B** Figure 4 shows that LLaMA-3 70 B's open-ended answers cluster in the mid-range for every metric, not at the extremes. Its *semantic-similarity* box-plot sits roughly between 0.35 and 0.55, with whiskers reaching only the mid-0.70s; BLEU, ROUGE, and METEOR centre much lower (BLEU's median is barely above 0.04, ROUGE around 0.20, METEOR around 0.25). The small band of higher-value semantic outliers (around 0.65–0.75) is confined to `short inverse` and `multi-hop inverse` items in which the model repeated key medical terms but reordered the surrounding sequence of words, so n-gram overlap stayed muted. Conversely, many `short` replies are abrupt noun-phrases, depressing all four metrics and keeping the inter-quartile ranges tight.

- **Phi-4-Reasoning-Plus** (Figure 5). The cloud at the extreme lower-left of the dashboard mirrors the 624 malformed list entries and 813 invalid True/False lines produced by this model. Extensive "chain-of-thought" preambles obscured the required delimiters, so the automated evaluation script extracted empty or partial lines. BLEU/ROUGE medians (around 0.04) remained higher than the semantic median (around 0.02) because the responses still shared surface n-grams with the references.

8.4 Cross-metric contrasts.

1. The ClinIQLink Semantic similarity metric displayed higher variance than any n-gram metric across every model dashboard, reflecting sensitivity to both omissions *and* verbose digressions.

2. The gap between ClinIQLink Semantic similarity metric and BLEU was inversely correlated with parameter count; smaller Qwen checkpoints recycled reference wording, whereas 70-B LLaMAs paraphrased aggressively.

3. `Multi hop inverse` was the most discriminative sub-task; its step-penalty compressed medians for every system (lowest boxes in Figure 2), frequently reshuffling neighbouring ranks in Table 2.

8.5 Findings

- High closed-ended scores hide residual hallucinations. Even with a vocabulary capped at just true/false or four choice-letters, every model occasionally invented an out-of-range option, proving that 0.75–0.82 headline accuracies do not equal flawless control.

- List questions are the singular closed-ended format that is still able to effectively discriminate model effectiveness because they demand selecting all true items while rejecting distractors, and as such, macro–micro F_1 was found to be spread from 0.30 to 0.68. Those wider answer sets surface the hallucinated extras that multiple-choice and true/false conceal.

- "Critique" is easier than "generate". Across the board, `short inverse` prompts (spot the error) cluster around 0.50 semantic similarity which is roughly double the median for forward `short` prompts that require composing a fresh answer.

- Multi-hop-inverse is the most discriminative open-ended task. Its step-distance penalty drags every model's median to the bottom of Figure 2, reshuffling several adjacent leaderboard positions and exposing brittle reasoning chains.

- Embedding-level similarity scores for LLM evaluation tasks are now required as the minimum standard. High-ranked systems such

as Mixtral 22B and LLaMA-3.3 70B often score > 0.60 on the semantic metric while BLEU, ROUGE and METEOR sit < 0.05, confirming that lexically novel yet faithful paraphrases fool token-overlap measures. Conversely, runs that recycle reference text earn decent n-gram scores but remain low on the embedding metric, demonstrating that overlap alone no longer tracks answer fidelity.

- Note on open-ended evaluation challenges. Despite impressive progress in embedding-based metrics (e.g., BERTScore, SentenceMover, BLEURT, COMET etc.) and NLI-based metrics (e.g., MENLI, UniEval), no single method can yet (a) decide with high confidence that two free-form LLM responses convey the same meaning, while also (b) grounding that decision in consistent entity and relation alignment across passages. Embedding similarity captures distributional closeness but is blind to logical entailment; NLI classifiers reason over sentence-level entailment yet lack explicit entity grounding and scale poorly beyond short contexts, and recent surveys and benchmark studies conclude that integrating these complementary views into a robust, scalable metric remains an unsolved problem and a key direction for future work (Ito et al., 2025; Croxford et al., 2025).

9 Conclusion

The CLINIQLINK evaluation shows that modern LLMs reach impressive headline scores on tightly constrained *True/False* and single-letter *multiple-choice* items, yet every model evaluated still sporadically produces out-of-vocabulary or otherwise invalid answers; unordered *list* questions, with their wider response space, remain the only closed-ended format able to expose this fragility. On open-ended tasks, embedding-based semantic similarity distinguishes genuinely informative paraphrases from superficial n-gram overlap. Conventional n-gram indices systematically mis-score open responses, rewarding superficial token overlap while penalising lexically novel yet factually correct paraphrases; embedding-based similarity aligns far more closely with clinical accuracy and, through the step-penalised *multi-hop-inverse* task, reveals brittle reasoning chains. More work is required to produce an effective semantic similarity scoring metric with explicit reasoning validation into a composite metric that more rigorously captures factuality, logical coherence, entity relationship framing, and schema compliance. To support this goal, future iterations of CLINIQLINK will link each question–answer pair to a machine-readable knowledge graph for graph-based verification of multi-step rationales and will introduce multimodal variants that couple text queries with images, thereby challenging models to ground their answers in heterogeneous clinical evidence.

Acknowledgments

This research was supported in part by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov) and the University of Maryland's Zaratan cluster (https://hpcc.umd.edu/hpcc/zaratan.html). This research was also partly supported by the U.S. Fulbright program, enabling international collaborative efforts.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. Phi-4-reasoning technical report. *Preprint*, arXiv:2504.21318.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor,

Michigan. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. *Preprint*, arXiv:2202.07646.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. 2025. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2(1).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey. *arXiv preprint*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms. *Preprint*, arXiv:2205.05131.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *Preprint*, arXiv:2305.06311.

Hye Sun Yun and Timothy Bickmore. 2025. Online health information–seeking in the era of large language models: Cross-sectional web-based survey study. *J Med Internet Res*, 27:e68560.