From GPT-4 to GPT-5: Measuring Progress in Medical Language Understanding Through MedHELM

Fernando Trevino

August 11, 2025

Abstract

This work integrates GPT-5 into the public, deterministically scored subset of MedHELM, a medically focused subsuite of HELM spanning quantitative calculation, factual question answering, evidence grounding, hallucination resistance, fairness probes, and text-to-SQL generation. The integration is append-only and configuration driven (parity prompts, fixed seeds, temperature 0.0), preserving longitudinal comparability with prior GPT-4 era baselines and external model entries without altering scenario semantics. The process provides a transparent path to track incremental frontier model progress in medically relevant capabilities while exposing residual risk surfaces.

Results show selective capability gains—stronger numerically grounded reasoning (MedCalc-Bench tie) and broad factual recall (HeadQA, Medbullets new highs)— alongside regressions or plateaus in schema-constrained generation (EHRSQL), fairness-sensitive reasoning (RaceBias), and full hallucination suppression (MedHallu shortfall vs leader). Efficiency is heterogeneous: some longer reasoning traces run faster, while short structured queries incur latency penalties without accuracy benefit.

Keywords: MedHELM, GPT-5, Medical AI Evaluation, Clinical Reasoning, Benchmarking, Safety

1 Introduction

Large language models (LLMs) have advanced clinical reasoning, structured information extraction, and medical knowledge retrieval. MedHELM [1] is a comprehensive benchmark suite that evaluates LLM performance across medical domains through standardized scenarios covering factuality, multi-step reasoning, safety constraints, and clinical ambiguity. The benchmark tracks progress across multiple vendors and architectures in the foundation model landscape.

However, benchmark coverage lags behind new model releases. The recently released GPT-5 represents a major release not yet systematically evaluated within MedHELM's medical scenarios. This gap limits our ability to (i) quantify longitudinal progress from the latest model generation, (ii) identify which capability clusters (medical calculations, EHR text-to-SQL, differential diagnosis, hallucination resistance) benefit most from recent innovations, and (iii) surface persistent high-risk failure modes.

Evaluating GPT-5 within MedHELM is essential for both scientific measurement and evidence-based deployment risk assessment. Clinically meaningful progress requires more than aggregate accuracy gains—it depends on calibration under uncertainty, reduced hallucinations, correct structured outputs, and reliable tool-integrated reasoning. By systematically adding GPT-5 to existing MedHELM scenarios, this work enables temporal comparison against GPT-4 era baselines and other leading models.

This work reports (1) methodology for reproducibly integrating a post-release model into MedHELM without contaminating prior scenario distributions, (2) quantitative deltas between GPT-4 and GPT-5 across diverse medical task families, and (3) analysis of error typologies that

remain clinically salient. The results offer an updated empirical snapshot of frontier medical LLM capability and highlight evaluation design principles needed to ensure benchmarks remain discriminative as model performance improves.

2 Results

2.1 Selected Scenarios

MedHELM categorizes benchmarks into three access levels: (i) **Public benchmarks** that are fully open and freely available (e.g., HuggingFace Datasets) with no access requirements, (ii) **Gated benchmarks** that require special permissions, credentials, or data use agreements (e.g., PhysioNet datasets), and (iii) **Private benchmarks** based on proprietary datasets available only to specific organizations. Only public benchmarks enable full reproducibility for any researcher.

This evaluation prioritized public scenarios satisfying three criteria: (i) **public accessibility**, (ii) **objective, automatable scoring** (exact match, execution accuracy, or deterministic classification without LLM jury adjudication), and (iii) **reproducibility** (fixed seeds, deterministic sampling, versioned evaluation scripts). This avoids reliance on proprietary data or LLM jury scoring pipelines that introduce variability and limit longitudinal comparability through judge model selection, prompt sensitivity, and biases that shift over time. Table 1 summarizes the selected scenarios and their clinical rationales.

Scenario	Capability Axis	Description
MedCalc-Bench	Quantitative calculations	Ensures precise dosing and acid-base computations; arithmetic drift risks clinical mis-dosing or misinterpretation.
Medec	Error detection	Surfaces charting errors early to prevent propagation into downstream decision tools.
HeadQA	Factual reasoning	Emulates board-style multi-domain reasoning requiring cross-specialty integration.
Medbullets	Factual recall	Gauges breadth of core clinical knowledge foundational for higher-order reasoning.
PubMedQA	Evidence-based QA	Tests alignment of answers to limited evidence and discourages unsupported speculation.
EHRSQL	Text-to-SQL generation	Assesses faithful structured data retrieval; failures risk incorrect patient data extraction and analytic bias.
RaceBias	Fairness evaluation	Probes avoidance of race-based inappropriate differentials to ensure equitable recommendations.
MedHallu	Hallucination detection	Evaluates resistance to confident fabrication of unsupported clinical claims.

Table 1: Selected MedHELM scenarios with deterministic scoring. Numeric scoring formula is scenario-specific and represents variants of accuracy or exact match (EM).

This subset balances breadth across medical reasoning tasks while maintaining reproducible evaluation for tracking longitudinal model progress.

2.2 Evaluation Results and Comparative Analysis

To quantify GPT-5 progress, this work compares GPT-5 against (i) established GPT-4 era baselines (GPT-4o and reasoning-focused o3-mini) to measure generational improvements, and (ii) the highest-performing model ("Leader") per scenario to assess competitive positioning. All comparisons use identical evaluation conditions (temperature 0.0, same prompts and metrics).

These comparisons isolate generational improvements within the GPT lineage while assessing competitive positioning across major vendors.

To get started with the analysis, we first present the current MedHELM leaderboard standings prior to GPT-5 evaluation in Table 2. This table summarizes the highest reported performance per scenario across all evaluated models to date, including GPT-4o, o3-mini, and other leading models from the MedHELM leaderboard [2].

Scenario	Current Leader Model	Score
MedCalc-Bench	DeepSeek R1	0.35
Medec	o3-mini	0.69
HeadQA	Claude 3.7 Sonnet / Claude 3.5 Sonnet / GPT-4o	0.91
Medbullets	o3-mini	0.81
PubMedQA	o3-mini / DeepSeek R1 / Claude 3.5 Sonnet	0.74
EHRSQL	GPT-4o	0.32
RaceBias	DeepSeek R1	0.92
MedHallu	Claude 3.5 Sonnet	0.93

Table 2: Current MedHELM leaderboard standings prior to GPT-5 evaluation.

With the current leaderboard table established, we now report GPT-5 results on the selected MedHELM scenarios. Table 3 lists scores for GPT-4o, o3-mini, GPT-5, and the current leader per scenario. All runs share identical settings (temperature 0.0, shared prompts, fixed seeds). Metrics are scenario-defined (exact match, execution accuracy, or numerical accuracy) and values are shown as proportions.

Scenario	Metric	GPT-4o	o3-mini	GPT-5	Current Leader Model
MedCalc-Bench	MedCalc Acc	0.19	0.34	**0.35**	0.35
Medec	Flag Acc	0.58	0.69	**0.66**	0.69
HeadQA	EM	0.91	0.89	**0.93**	0.91
Medbullets	EM	0.71	0.81	**0.89**	0.81
PubMedQA	EM	0.70	0.74	**0.67**	0.74
EHRSQL	ExeAcc	0.32	0.27	**0.18**	0.32
RaceBias	EM	0.90	0.87	**0.72**	0.92
MedHallu	EM	0.85	0.90	**0.88**	0.93

Table 3: MedHELM scenario results for GPT-4 era models, GPT-5 and current leader model.

To contextualize performance we compute per-scenario deltas: GPT-5 minus the best GPT-4 baseline (max of GPT-4o, o3-mini) and GPT-5 minus the current leader (which may be an external model). Table 4 reports these values; positive indicates improvement, negative regression.

GPT-5 establishes new highs in HeadQA (+0.02) and Medbullets (+0.08) and ties for the lead in MedCalc-Bench (+0.01 over prior best GPT-4). Regressions are largest in EHRSQL (-0.14) and RaceBias (-0.18), with smaller declines in PubMedQA (-0.07) and hallucination resistance (MedHallu -0.02). The negative mean delta (-0.04 vs best GPT-4, -0.05 vs leader) is driven by a minority of scenarios with substantial deficits rather than broad underperformance. In this regard, the high priority remediation targets are: (i) schema grounding + constrained decoding for EHRSQL; (ii) fairness robustness (RaceBias); (iii) calibration on evidence- constrained QA (PubMedQA); (iv) residual hallucination suppression.

Scenario	Δ GPT-5 – GPT-4	Δ GPT-5 – Current Leader	Δ GPT-4 – Current Leader
MedCalc-Bench	+0.01	0.00	-0.01
Medec	-0.03	-0.03	0.00
HeadQA	+0.02	+0.02	0.00
Medbullets	+0.08	+0.08	0.00
PubMedQA	-0.07	-0.07	0.00
EHRSQL	-0.14	-0.14	0.00
RaceBias	-0.18	-0.20	-0.02
MedHallu	-0.02	-0.05	-0.03
Mean	**-0.04**	**-0.05**	**-0.01**

Table 4: GPT-5 performance deltas relative to best GPT-4 model and current leaders. Positive values indicate improvement; negative values indicate regression.

Scenario	New Leader Model	Score
MedCalc-Bench	**GPT-5** / DeepSeek R1	0.35
Medec	o3-mini	0.69
HeadQA	**GPT-5**	0.93
Medbullets	**GPT-5**	0.89
PubMedQA	o3-mini / DeepSeek R1 / Claude 3.5 Sonnet	0.74
EHRSQL	GPT-4o	0.32
RaceBias	DeepSeek R1	0.92
MedHallu	Claude 3.5 Sonnet	0.93

Table 5: Updated MedHELM leaderboard standings after GPT-5 evaluation. GPT-5 achieves new leadership in HeadQA and Medbullets, and ties for leadership in MedCalc-Bench, demonstrating strengths in quantitative reasoning and factual knowledge tasks.

GPT-5 attains new leadership in HeadQA and Medbullets and ties MedCalc-Bench, indicating strong multi-domain factual recall, reasoning, and quantitative calculation gains. Regressions in EHRSQL and RaceBias show unresolved weaknesses in schema-grounded structured generation and fairness-sensitive reasoning. Overall performance is mixed: targeted advances with persistent gaps. Mean deltas (-0.04 vs best GPT-4, -0.05 vs leader) indicate uneven progress and clear remediation needs. Table 5 reflects the post-integration leaderboard, updating scenario leadership to include GPT-5's new or tied first-place scores.

2.3 Inference Latency Analysis

To complement accuracy-centric comparisons, this work reports per-scenario inference latency for GPT-5 versus the Leader model in Table 6, using the per-instance mean (single request, no batching, temperature 0.0) extracted from the HELM package generated reports. Because only GPT-5 and (when different) the Leader model were logged in the current run configuration, earlier GPT-4 family baselines (GPT-4o, o3-mini) appear only when they serve as Leader (EHRSQL, Medbullets). A simple ratio (GPT-5 / Leader) is included. Latency should be interpreted jointly with accuracy: a slower model may justify overhead if it provides clinically relevant quality gains, while regressions with higher latency (e.g., EHRSQL) highlight priority optimization targets.

GPT-5 averages 15.05s per instance vs 13.56s for current leaders (mean ratio 1.11). It is faster where leader runs are longer: MedCalc-Bench (0.50×) and Medec (0.67×). Large ratio

Scenario	GPT-5 (s)	Current Leader Model (s)	GPT-5 / Leader
MedCalc-Bench	22.06	43.75	0.50
Medec	28.27	41.88	0.67
HeadQA	5.87	0.36	16.31
Medbullets	13.65	7.29	1.87
PubMedQA	4.88	1.49	3.28
EHRSQL	30.94	3.83	8.08
RaceBias	6.47	7.51	0.86
MedHallu	8.23	2.39	3.44
Mean	**15.05**	**13.56**	**1.11**

Table 6: Mean per-scenario inference latency. Ratio <1: GPT-5 faster than the accuracy leader model (Table 2). Times include network overhead and are dependent on deployment conditions. In leader score ties, the fastest (minimum) leader model among tied models is reported.

outliers (HeadQA 16.31×, EHRSQL 8.08×) stem from very small leader baselines (0.36s, 3.83s) rather than extreme absolute times. EHRSQL combines slowdown and accuracy regression, making it the top efficiency + quality remediation target; RaceBias is near parity (0.86×).

Overall latency is heterogeneous: GPT-5's fixed overhead dominates short retrieval / classification prompts but amortizes on multi-step arithmetic and error-detection traces. Future releases should add per-token latency with output length, variance / p95 statistics, and energy or cost-normalized efficiency metrics to refine these comparisons.

3 Methods

3.1 Evaluation Framework

MedHELM is a curated subsuite within CRFM HELM that reuses HELM's runners, metrics, and artifact schema [4]. Adding GPT-5 therefore required an append-only configuration plus a custom client; task semantics and deterministic settings (seed, temperature 0.0) were held fixed so new outputs align with prior model–scenario pairs and preserve comparability.

We restrict to public, deterministically scored scenarios to avoid opacity from judge models, subjective adjudication and stochastic drift.

A local HELM configuration layer requires:

- A run entries list enumerating scenario–model tuples (including GPT-5) for reproducible invocation.
- A deployment specification mapping a logical GPT-5 id to the custom client and decoding defaults.
- A credentials file or environment variable reference for secure key loading.
- A custom client that calls the OpenAI GPT-5 API using the desired deployment.

The existing OpenAI client targets GPT-4 era APIs and could not be reused (parameter schema differences; unsupported fields like max tokens). Therefore a new wrapper that implements the HELM interface was added.

The released packaged used is `crfm-helm 0.5.6` to guarantee metric and normalization parity with public leaderboard entries. Environment setup (recommended: Conda) isolates Python + binary dependencies; no core HELM source was modified, preserving an auditable append-only change surface.

The post-installation evaluation workflow is as follows:

1. `helm-run`: resolve datasets (e.g., HuggingFace), then run inference for the configured scenario–model tuples under fixed seeds.
2. `helm-summarize`: aggregate raw outputs into normalized artifacts for visualization and comparison.
3. `helm-server`: launch a local UI to inspect summarized results interactively.

This process balances engineering overhead with auditability, enabling incorporation of new frontier models while maintaining longitudinal integrity of MedHELM results.

3.2 Integrating GPT-5 into HELM

Documentation gaps affected model extension: existing HELM docs describe installation and scenario structure but leave addition of a new models largely implicit. The framework is intentionally modular, so integration strategy followed established extension patterns without modifying core source.

The existing OpenAI client did not yet expose GPT-5, so a lightweight Python module was added to wrap the OpenAI API with deterministic decoding while reusing HELM's adapter + normalization layers. This preserves metric semantics and artifact schema.

Undocumented but required steps were confirmed empirically: dynamic resolution of a new deployment name, alignment between the registered key and run entry, and credential loading via environment variable or a token file. The general implementation steps were:

1. Install HELM editable.
2. Add custom client exposing model id `openai/gpt-5`.
3. Register deployment in the models config.
4. Add run list entry.
5. Dry-run a small HeadQA batch for schema validation.
6. Execute full suite and collect JSON + aggregates.

Sampling, prompting, and metrics follow earlier described methodology.

The exact implementation artifacts for integrating GPT-5 into HELM alongside an unmodified HELM installation are the following:

- **Custom OpenAI Client Module (`custom_client.py`):** Wraps the OpenAI API exposing model id `openai/gpt-5` while reusing HELM adapter + normalization layers to apply existing metric implementations.
- **Deployment Configuration (`model_deployments.yaml`):** Registers the GPT-5 deployment (logical id, provider, decoding defaults) enabling dynamic resolution by the runner.
- **Model Metadata (`model_metadata.yaml`):** Supplies descriptive attributes (context window, modality flags, intended use tags) consumed by reporting/aggregation scripts.
- **Credentials File (`credentials.conf`):** Supports loading the OpenAI API key via environment variable or token file path referenced by the deployment entry.
- **Public Run Entries List (`run_entries_medhelm_public.conf`):** Enumerates scenario–model tuples (including GPT-5) for reproducible invocation of the selected public MedHELM scenarios.

These configuration artifacts mirror the schemas and field conventions used by existing model entries within the upstream HELM package (e.g., prior OpenAI and Anthropic model definitions), enabling drop-in comparability and minimizing maintenance divergence.

Code Availability. All configuration files, custom client code, and run entry lists used to generate the results in this paper are publicly available at `https://github.com/fertrevino/medhelm_gpt_5`. The repository includes the exact `run_entries` specification, deployment metadata, and wrapper client needed to reproduce the GPT-5 MedHELM runs under the deterministic settings described.

4 Discussion

4.1 Performance Interpretation

GPT-5 exhibits selective advances rather than uniform superiority across medical task families. Largest relative gains appear in numerically grounded and broad mixed-domain knowledge scenarios (MedCalc-Bench tie for lead, HeadQA +0.02 over prior leader, Medbullets +0.08 absolute over best GPT-4 model), consistent with scaling benefits for multi-step arithmetic and semantic retrieval without explicit chain-of-thought prompting. These improvements likely reflect increased routed capacity and latent reasoning reliability, enabling accurate internal decomposition under deterministic decoding. For each of the evaluated categories, these are the result interpretations:

- **MedCalc-Bench (0.35 tie for lead).** +0.16 over GPT-4o and +0.01 over o3-mini indicates consolidated arithmetic reliability; residual errors cluster in multi-variable acid–base edge cases.
- **Medec (0.66 $<$ 0.69 leader).** +0.08 vs GPT-4o but −0.03 vs o3-mini shows partial progress in error flagging yet ceiling remains below clinically desirable threshold ($\geq$0.75) for low-risk deployment.
- **HeadQA (0.93 new leader).** +0.02 over prior 0.91 plateau signals incremental but consistent cross-specialty reasoning lift under deterministic decoding.
- **Medbullets (0.89 new leader).** +0.08 over best GPT-4 baseline reflects broadened factual recall; misses concentrate in low-frequency subspecialty items.
- **PubMedQA (0.67 $<$ 0.74 leader).** −0.03 vs GPT-4o and −0.07 vs o3-mini suggests reduced calibration on abstract-style evidence questions; likely sensitivity to implicit ternary answer priors.
- **EHRSQL (0.18 $<$ 0.32 leader).** Largest regression: −0.14 vs GPT-4o driven by column hallucinations and partial predicate omission; indicates schema grounding gap despite broader reasoning gains.
- **RaceBias (0.72 $<$ 0.92 leader).** Substantial fairness regression (−0.18 vs GPT-4o) raising concern of emergent bias behaviors with scaling; prioritizes bias-aware fine-tuning.
- **MedHallu (0.88 $<$ 0.93 leader).** +0.03 vs GPT-4o but trailing o3-mini (−0.02) and leader (−0.05); improvements incomplete for high-stakes factuality demands.

Performance regressions in EHRSQL (−0.14 vs GPT-4o) highlight that generic reasoning and factual recall gains do not automatically translate to schema-constrained structured generation. Failure analyses (not shown) indicate a mix of column name hallucinations and incomplete

logical condition coverage—suggesting a need for improved schema grounding and compositional planning. RaceBias degradation (-0.18 vs best GPT-4 baseline, -0.20 vs external leader) underscores that fairness-sensitive behaviors can backslide even as aggregate reasoning improves, reinforcing multi-axis evaluation requirements for clinical adoption.

Safety and reliability signals are mixed: modest hallucination resistance improvement over GPT-4o (MedHallu $+0.03$ absolute) still trails the external leader, while medical error flagging (Medec) remains below 0.70 accuracy for all models, leaving clinically material residual risk. These plateaus indicate that scaling alone has not closed gaps in subtle factuality and documentation error detection. Latency profiling shows heterogeneous efficiency: GPT-5 is faster on longer arithmetic/error-detection traces (e.g., $0.50\times$ leader latency on MedCalc-Bench) yet markedly slower on short retrieval or structured generation tasks (HeadQA $16.31\times$, EHRSQL $8.08\times$), compounding the cost of quality regressions in the latter.

Limitations temper interpretation: (i) reduced sample counts (100) in some scenarios inflate variance; (ii) exclusion of tasks requiring LLM or human judge scoring prioritizes determinism but omits nuanced subjective safety assessments; (iii) potential public data leakage cannot be definitively excluded; (iv) some external leaderboard references rely on published, not locally replicated, scores; (v) latency measurements include deployment overhead and may not isolate pure model inference.

In aggregate, GPT-5 advances numerically grounded and broad factual recall capabilities while leaving structured querying, fairness, and certain safety behaviors as primary remediation targets.

4.2 Future Work

Planned benchmark extensions focus on converting identified weaknesses into discriminative, automatable evaluations: (a) enrich structured data tasks (expanded EHRSQL schemas, FHIR query and temporal aggregation benchmarks) to pressure-test schema grounding; (b) introduce calibrated probability elicitation for differential diagnosis and ambiguity-heavy cases to assess overconfidence; (c) release fine-grained error taxonomies (hallucination subtype, bias mechanism, SQL failure class) for open auditing; (d) add longitudinal drift tracking comparing monthly checkpoint deltas to surface silent regressions; (e) integrate fairness stress tests spanning intersectional attributes beyond race; (f) pilot semi-automated factuality adjudication pipelines (retrieval-augmented citation verification) while retaining deterministic scoring paths.

Model-side remediation avenues include targeted schema grounding adapters, constrained decoding or executable-in-the-loop SQL validation, curriculum-style fairness fine-tuning, and hallucination-focused contrastive supervision. Public release of reproducible configurations (run entries, deployment metadata) will continue to support longitudinal comparability and external replication.

Overall, maintaining MedHELM as an append-only, transparency-first evaluation substrate remains critical for distinguishing genuine clinical reasoning progress from redistribution of errors across capability axes.

References

[1] Stanford CRFM. *MedHELM: Holistic Evaluation of Large Language Models for Medical Applications*. arXiv preprint arXiv:2505.23802, 2025.

[2] Stanford Center for Research on Foundation Models. *MedHELM v2.0.0 Leaderboard*. Available at: `https://crfm.stanford.edu/helm/medhelm/v2.0.0/#/`, 2024.

[3] Stanford CRFM. *HELM Installation Documentation*. Available at: `https://crfm-helm.readthedocs.io/en/latest/installation/`, 2024.

[4] Liang, Percy and Bommasani, Rishi and Lee, Tony and others. *Holistic Evaluation of Language Models (HELM)*. GitHub repository, 2023. Available at: `https://github.com/stanford-crfm/helm`

[5] Chen, Justin Chih-Yao and Yun, Sukwon and Stengel-Eskin, Elias and Chen, Tianlong and Bansal, Mohit. *Symbolic Mixture-of-Experts: Adaptive Skill-based Routing for Heterogeneous Reasoning*. arXiv preprint arXiv:2503.05641, 2025.

[6] Chen, Justin Chih-Yao and Yun, Sukwon and Stengel-Eskin, Elias and Chen, Tianlong and Bansal, Mohit. *Symbolic-MoE: Implementation Repository*. GitHub repository, 2025. Available at: `https://github.com/dinobby/Symbolic-MoE/`

[7] Lee, Gyubok and Hwang, Hyeonji and Bae, Seongsu and others. *EHRSQL: A Practical Text-to-SQL Benchmark for Electronic Health Records*. Advances in Neural Information Processing Systems, volume 35, pages 15589–15601, 2022. arXiv:2301.07695.

[8] Chen, Zixiang and Deng, Yihe and Wu, Yue and Gu, Quanquan and Li, Yuanzhi. *Towards Understanding Mixture of Experts in Deep Learning*. arXiv preprint arXiv:2208.02813, 2022.

[9] Yao, Shunyu and Zhao, Jeffrey and Yu, Dian and Du, Nan and Shafran, Izhak and Narasimhan, Karthik and Cao, Yuan. *ReAct: Synergizing Reasoning and Acting in Language Models*. arXiv preprint arXiv:2210.03629, 2022.

[10] Lee, Gyubok and Hwang, Hyeonji and others. *EHRSQL: Implementation Repository*. GitHub repository, 2022. Available at: `https://github.com/glee4810/EHRSQL/tree/main`

[11] Khandekar, Nikhil and Jin, Qiao and Xiong, Guangzhi and others. *MedCalc-Bench: Evaluating Large Language Models for Medical Calculations*. arXiv preprint arXiv:2406.12036, 2024.

[12] Stanford CRFM. *MedCalc-Bench Scenario Implementation for HELM*. GitHub, 2024. Available at: `https://github.com/stanford-crfm/helm/blob/main/src/helm/benchmark/scenarios/medcalc_bench_scenario.py`

[13] Khandekar, Nikhil and Jin, Qiao and others. *MedCalc-Bench: Implementation Repository*. GitHub, 2024. Available at: `https://github.com/ncbi-nlp/MedCalc-Bench`

[14] Vilares, David and Gómez-Rodríguez, Carlos. *HEAD-QA: A Healthcare Dataset for Complex Reasoning*. Proceedings of ACL 2019 (Short Papers), 2019. arXiv:1906.04701.

[15] Jin, Qiao and Dhingra, Bhuwan and Liu, Zhengping and Cohen, William W. and Lu, Xinghua. *PubMedQA: A Dataset for Biomedical Research Question Answering*. Proceedings of EMNLP 2019, 2019. arXiv:1909.06146.

[16] Chen, Hanjie and Fang, Zhouxiang and Singla, Yash and Dredze, Mark. *Benchmarking Large Language Models on Answering and Explaining Challenging Medical Questions*. arXiv preprint arXiv:2402.18060, 2024.

[17] Ben Abacha, Asma and Yim, Wen-wai and Fu, Yujuan and Sun, Zhaoyi and Yetisgen, Meliha and Xia, Fei and Lin, Thomas. *MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes*. arXiv preprint arXiv:2412.19260, 2024.

[18] Pandit, Shrey and Xu, Jiawei and Hong, Junyuan and Wang, Zhangyang and Chen, Tianlong and Xu, Kaidi and Ding, Ying. *MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models*. arXiv preprint arXiv:2502.14302, 2025.

[19] [Title unavailable due to access overlay]. npj Digital Medicine, 2023. URL: https://www.nature.com/articles/s41746-023-00939-z. (Please provide exact title for update.)

[20] OpenAI. *GPT-5 System Card and Technical Report*. 2025. (Placeholder citation; update upon official release.)

[21] Stanford CRFM. *Holistic Evaluation of Language Models (HELM) Documentation*. Available at: `https://crfm-helm.readthedocs.io/en/latest/`, 2025.