From Answers to Hypotheses

Internal Consensus and Its Limits in Large Language Models

Victor Lavrenko
victor@peacetech.vc

Abstract

Large language models (LLMs) now reach near-expert performance on medical QA and diagnostic benchmarks, yet this progress does not reliably translate into safe clinical decision support. We argue that a core limitation is not missing knowledge, but a failure to surface and reason over the *distribution of internal hypotheses* expressed across alternative decoding trajectories. Using controlled multi-branch sampling on a single medical QA benchmark with one model, one prompt, and a fixed sampling regime, we show that (i) majority-vote self-consistency does not yield statistically significant accuracy gains over greedy decoding in this setting; (ii) the correct answer is often present among sampled hypotheses even when the final prediction is wrong; and (iii) strong internal consensus can still coincide with incorrect answers. We formalize these effects with statistical tests and offer *errors of hypothesis space* as an empirically motivated interpretation—not a formally isolated mechanism—where models converge confidently on a systematically wrong explanation, motivating hypothesis-centric evaluation analogous to differential diagnosis.

1 Introduction

Large language models (LLMs) have recently achieved impressive results on medical examinations and diagnostic benchmarks, in some cases rivaling or exceeding average physician performance on structured question answering and exam-style tasks. This progress has fueled optimism about near-term clinical deployment, with several systems now approaching conversational history-taking, differential diagnosis, and longitudinal reasoning. However, randomized studies and post-deployment analyses increasingly suggest that benchmark gains do not reliably translate into improved clinical decision-making and may introduce new risks when model outputs are treated as definitive answers.

A dominant explanation for this gap attributes failures to hallucinations and overconfidence. While these phenomena are real, prior work paints a more nuanced picture of model uncertainty. Kadavath et al. (2022) show that language models often *know when they are likely to be wrong*, in the sense that internal signals correlate with error, but this knowledge is not reliably exposed through standard decoding. Similarly, Wang et al. (2022) demonstrate that sampling multiple reasoning traces and aggregating them via self-consistency can improve accuracy on some reasoning benchmarks, suggesting that alternative hypotheses already exist within the model's latent reasoning process.

Recent clinical and near-clinical evaluations further underscore the limits of treating improved benchmark accuracy as a proxy for dependable decision support. In a randomized controlled study, access to an LLM did not significantly improve physicians' diagnostic reasoning compared to conventional resources (Goh et al. 2024), despite strong standalone model performance. At the same time, frontier systems such as AMIE illustrate rapid progress toward conversational diagnostic assistance and differential diagnosis (Tu et al. 2025; McDuff et al. 2025), amplifying the importance of understanding when internally coherent reasoning trajectories nonetheless converge on incorrect conclusions.

Taken together, these findings suggest that uncertainty and alternative explanations are not absent from modern LLMs; rather, they are often *latent*. The open question is whether these internal structures are sufficiently reliable, diverse, and accessible to support high-stakes decision-making. In medicine, uncertainty is rarely resolved by selecting a single most likely hypothesis; instead, clinicians reason over *sets* of competing explanations, iteratively revising them as new evidence emerges. This work adopts that perspective and studies LLM reasoning as a *distribution over hypotheses*, rather than as a single chain of thought or a single decoded answer.

Operationally, we observe this distribution through repeated sampling of the same prompt under fixed conditions, yielding an empirical distribution over final answers and associated explanations. We do not assume access to internal symbolic representations or latent states; rather, by "hypotheses" we refer to the set of discrete answer-level explanations implicitly instantiated by independent decoding trajectories. This framing allows us to analyze internal agreement, diversity, and failure modes without introducing new architectures or training procedures.

Finally, we emphasize that this work is diagnostic rather than prescriptive. We do not propose a new aggregation rule, verification strategy, or decoding algorithm. Instead, our goal is to characterize when and why aggregation and internal agreement fail, even under idealized sampling conditions, and to identify a distinct failure mode—*errors of hypothesis space*—as an empirically motivated interpretation in which models converge confidently on a systematically wrong explanation. Understanding these failures is a necessary step toward evaluation paradigms that more closely resemble differential diagnosis in clinical practice.

In this work, we evaluate a sequence of hypotheses about ensemble reasoning. We begin by testing whether aggregation improves accuracy (H1), whether correct answers tend to appear among alternative hypotheses (H2), and whether internal consensus predicts correctness (H3). We then study the feasibility of selective leader override as a function of internal consensus, and conclude with implications and future directions. All claims are scoped to a single medical multiple-choice QA benchmark, one model and prompt configuration, and a fixed sampling regime; we do not claim generality beyond these conditions.

2 Problem Setup and Definitions

We study a medical QA benchmark where each prompt is evaluated via greedy decoding and repeated sampling, yielding an empirical distribution over answer hypotheses and associated explanations. We define the *leader* answer as the most frequent hypothesis in this distribution and quantify internal consensus using the maximum agreement fraction across sampled branches.

Throughout this paper, we avoid the notion of symmetric or heuristic "leader overrides" of predictions. Instead, we study *selective leader override*: rejecting the plurality (most frequent) hypothesis in favor of a lower-ranked alternative when uncertainty indicators suggest the leader is unreliable.

3 Experimental Setup

We evaluate a single open-weight, instruction-tuned Llama 3.1 family checkpoint, `HPAI-BSC/Llama3.1-Aloe-Beta-8B`, which has 8B parameters and is run in full precision (bfloat16, no quantization). The model is accessed via the Hugging Face checkpoint and used with its chat template.

Our evaluation set is drawn from MedMCQA (Pal et al. 2022) (`openlifescienceai/medmcqa`) using the official validation split. We filter to single-choice questions with explanation length between 20 and 500 characters, then shuffle with seed 42 and select 400 questions while enforcing a per-subject cap of 23 to prevent over-representation of subjects with very large question pools (e.g., dentistry). MedMCQA contains 20 subject areas, and this procedure yields a roughly balanced evaluation set across subjects. This yields a deterministic 400-question evaluation set.

For each question, we generate:

- a greedy prediction using deterministic decoding; and
- an ensemble of $N = 10$ independently sampled reasoning paths.

Let $a_{i,j}$ denote the final answer of branch j for question i. From the empirical distribution $P_i(a)$, we compute:

- leader answer: $\arg\max_a P_i(a)$;
- maximum agreement fraction:
$$\text{max\_frac}_i = \max_a P_i(a);$$
- Top-k coverage:
$$\text{Top-}k = \nVdash\{y_i \in \text{Top-}k(P_i)\}.$$

All results are computed over this fixed 400-question evaluation set using a single prompt and the model configuration above.

4 Hypotheses and Empirical Tests

Aggregation Improves Accuracy

H1. Majority-vote aggregation improves accuracy relative to greedy decoding.

Let $\hat{y}_i^{(g)}$ denote the greedy prediction and $\hat{y}_i^{(m)}$ the majority prediction. Accuracy is defined as:

$$\text{Acc} = \frac{1}{N}\sum_i \nVdash[\hat{y}_i = y_i].$$

Empirically:

- Greedy accuracy: 65.75%
- Majority accuracy: 66.75%

A two-sided binomial test with null hypothesis $H_0 : \pi = 0.6575$, where π denotes the true success probability equal to the greedy decoding accuracy, yields a p-value of approximately 0.63. We therefore fail to reject H_0, indicating that the observed difference between greedy and majority-vote accuracy is not statistically significant in this setting.

Interpretation. These results provide no statistical evidence in support of H1 in this experimental regime, suggesting that majority-vote aggregation does not reliably improve accuracy over greedy decoding under the tested conditions. While self-consistency can improve performance in some regimes (Wang et al. 2022), our results show that such gains are not evident in this single-model, single-dataset setting. Aggregation alone is therefore insufficient as a general reliability mechanism under these conditions.

Correct Answers Appear Among Alternatives

H2. When the final prediction is incorrect, the correct answer is often present among alternative hypotheses.

We measure Top-2 coverage:

$$\text{Top-2 coverage} = \frac{1}{N}\sum_i \nVdash\{y_i \in \text{Top-2}(P_i)\}.$$

Observed Top-2 coverage is 80.5%, compared to a greedy accuracy of 65.75%, corresponding to an absolute improvement of 14.75 percentage points. Using a binomial model with null hypothesis $H_0 : \pi = 0.6575$ (where π is the baseline success probability equal to the greedy accuracy), this difference is highly statistically significant (p-value $\ll 10^{-6}$).

Relation to prior work. This result is consistent with Kadavath et al. (2022), who show that models frequently possess internal signals of uncertainty. Our findings extend this by demonstrating that uncertainty manifests as *explicit alternative hypotheses*, not merely as reduced confidence.

Internal Consensus Implies Correctness

H3. Strong internal agreement implies correctness.

We analyze unanimous cases where $\text{max\_frac}_i = 1.0$. Out of 400 questions:

- Unanimous cases: 151
- Unanimous accuracy: 86.8%

We test the null hypothesis $H_0 : \pi \geq 0.95$, where π denotes the true accuracy of unanimous predictions. A one-sided binomial test yields a p-value below 0.01, allowing us to reject this hypothesis. Thus, even strong internal consensus does not guarantee near-perfect reliability.

Distributional analysis of consensus. Figure 1 illustrates how accuracy varies with internal consensus. Accuracy increases monotonically with agreement but saturates well below perfect reliability. Notably, near-unanimous cases (max_frac ≥ 0.9) still exhibit error rates above 15%.

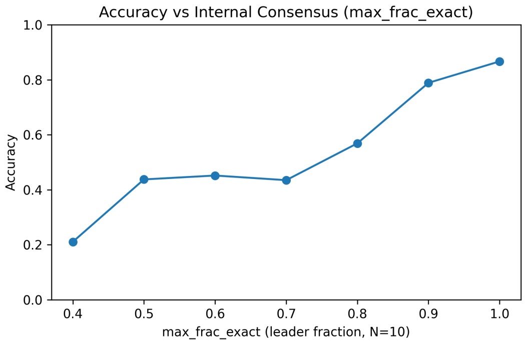

Figure 1: Accuracy as a function of internal consensus (max_frac). Higher branch agreement correlates with higher accuracy, but even near-unanimous predictions exhibit a non-zero error rate.

This behavior is inconsistent with the assumption that confidence or agreement can serve as a sufficient decision criterion.

Key insight. Internal consensus reflects *coherence*, not truth. Models can converge confidently on incorrect explanations, producing a dangerous illusion of reliability.

These results suggest that internal consensus is not only correlated with accuracy, but may serve as a control signal for selective intervention. These findings motivate evaluating whether selective leader override is practically feasible when restricted to low-consensus regimes.

Selective Leader Override Is Feasible

H4. Restricting leader override decisions to low-consensus regimes substantially reduces the precision required of an override algorithm while preserving access to a significant fraction of baseline errors.

Limits of naive top-2 leader override. A natural first idea is to apply leader override by switching predictions from the most frequent answer (top-1) to the second-most frequent answer (top-2) whenever uncertainty is detected. However, our results show that such naive leader override is generally unsafe. In regimes with strong consensus, correct top-1 predictions vastly outnumber correct top-2 predictions. As a result, even a small rate of incorrect leader overrides would outweigh any potential benefit. This imbalance is statistically significant in our data (binomial test, $p < 10^{-6}$ for high-consensus regimes), demonstrating that indiscriminate leader override would almost certainly degrade overall accuracy.

These findings indicate that top-2 coverage alone is insufficient: the feasibility of leader override depends critically on the relative frequencies of correct top-1 and top-2 hypotheses within a given uncertainty regime.

To make this trade-off explicit, we analyzed the feasibility of selective top-2 leader override across vote-consensus regimes defined by top-1 vote counts and top-1/top-2 gaps. For each regime, we quantified two key quantities (Figure 2).

First, we computed the maximum achievable overall accuracy under an idealized oracle that applies leader override in all cases where the second-most frequent hypothesis is correct (top-2 = gold) and never introduces new errors. This represents a strict upper bound, since errors outside top-2 coverage cannot be corrected by leader override.

Second, we measured the required false-override suppression, defined as the ratio between cases where the top-1 hypothesis is correct and cases where the top-2 hypothesis is correct within the regime. This ratio directly reflects how accurate a leader override decision policy must be in order to avoid degrading performance.

5 Feasibility of Selective Leader Override

The analysis reveals a sharp, regime-dependent trade-off. Broad regimes that include high-consensus predictions offer larger theoretical gains but require extremely high leader override precision, making them practically infeasible. In contrast, more conservative regimes focused on high-uncertainty cases—where top-1 and top-2 receive comparable support—exhibit substantially more favorable error budgets. In these regimes, even moderately accurate leader override policies yield meaningful improvements while incurring limited risk.

While the fixed 400-question evaluation set limits statistical power for fine-grained comparisons between neighboring regimes, the qualitative pattern persists within this dataset and sampling configuration. In particular, excluding high-consensus predictions (e.g., regimes with very high top-1 vote counts) leads to a statistically significant reduction in the precision required of a leader override algorithm, compared to regimes that include such cases.

We observe that moving the upper bound on top-1 vote counts from high-consensus regimes to uncertainty-focused regimes yields a more than twofold reduction in required false-override suppression (from ≈ 2.4 to ≈ 1.0), a difference that is statistically significant ($p < 10^{-6}$).

6 Discussion

The analysis presented in this work suggests that improving LLM performance in high-stakes medical settings is not primarily a matter of producing a single more accurate answer, but of reasoning over

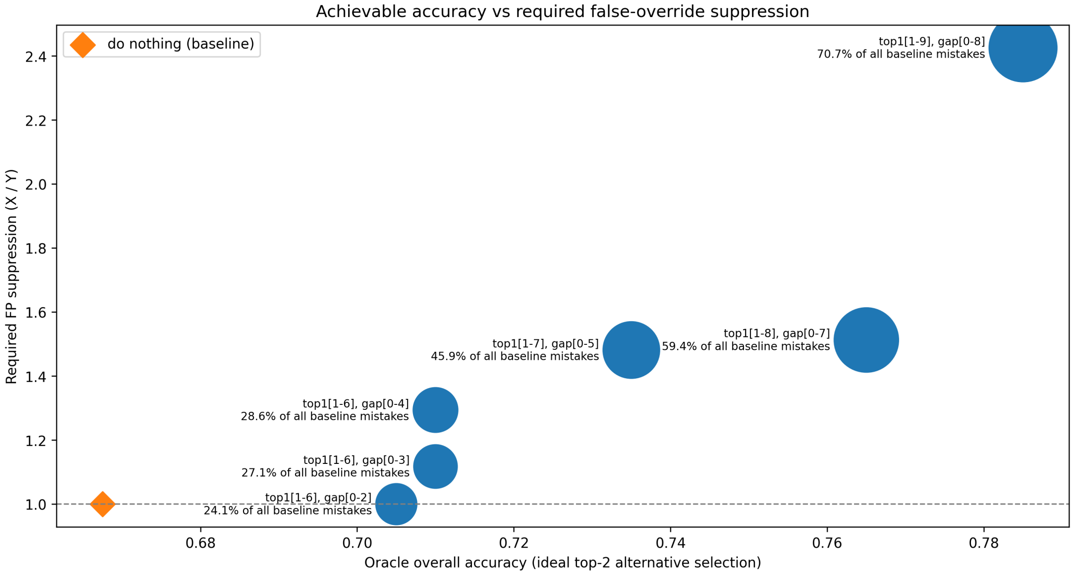

Figure 2: Feasibility of selective top-2 leader override across vote-consensus regimes. The x-axis shows the maximum overall accuracy achievable by an ideal oracle that corrects all top-2=gold cases within a regime. The y-axis indicates the required false-override suppression (top-1 correct vs. top-2 correct). Bubble size reflects the number of baseline errors covered by the regime.

structured sets of competing hypotheses. In particular, the presence of meaningful top-2 coverage opens a path toward correcting errors, but only under carefully constrained decision regimes. These conclusions are limited to the single medical multiple-choice QA benchmark, model, prompt, and sampling settings tested here; broader generalization remains an open question.

These findings suggest a staged and risk-aware strategy for selective leader override. Rather than attempting to correct all errors, override algorithms should initially target regimes with high uncertainty and favorable error budgets, where the cost of a mistaken override is relatively low and the likelihood of successful correction is higher. As decision accuracy improves, such algorithms could progressively expand into more challenging regimes, but this should be evaluated in additional datasets and model families before drawing general claims.

Importantly, this perspective reframes leader override not as a binary operation applied globally, but as a selective intervention governed by explicit uncertainty thresholds and error budgets. Even conservative strategies restricted to high-uncertainty regimes can already produce substantial gains in this setting, without requiring near-perfect override decisions.

6.1 Errors of Hypothesis Space

The failure to find evidence supporting H3 motivates an alternative *interpretive explanation* for these errors, consistent with the observed behavior but not yet empirically isolated as a distinct mechanism.

Interpretive hypothesis (H3′). One plausible interpretation is that some errors arise because the correct hypothesis is poorly represented or effectively absent from the model's hypothesis space.

Under this view, unanimous but incorrect predictions correspond to cases where the model's hypothesis space is locally coherent—yielding strong internal agreement—yet globally misaligned with the task. Importantly, the present analysis does not provide an operational test that cleanly separates missing hypotheses from low-probability selection among existing alternatives. We therefore do not claim to isolate this mechanism empirically.

Nevertheless, the observed behavior is inconsistent with an explanation based solely on stochastic selection noise. In these cases, increased sampling or stronger aggregation fails to recover the correct answer, suggesting a qualitative limitation of hypothesis representation rather than insufficient exploration.

This reframes aggregation not as a remedy for such errors, but as a *diagnostic probe*: a means of revealing when internal coherence reflects confidence without correctness.

Operational breakdown of failure modes. To make the taxonomy measurable, we operationalize three error categories using the $N{=}10$ sampled hypotheses per question. Among all evaluation items, selection errors—cases where the gold answer appears among the top-2 sampled hypotheses but is not selected by the leader—account for 13.8% of all questions and 41.4% of leader errors. In contrast, unsurfaced errors under sampling—cases where the gold answer does not appear in the top-2 hypotheses—account for 19.5% of all questions and 58.6% of leader errors. Finally, high-consensus errors highlight calibration limits: even in unanimous regimes, 13.2% of predictions are incorrect (20 out of 151 unanimous cases), representing 15.0% of all leader errors. We treat the second category as an *operational proxy* for hypothesis-space limitations under this sampling regime, without claiming mechanistic isolation. All percentages are reported as descriptive statistics for this fixed evaluation set and should be interpreted as approximate rather than precise estimates.

Taken together, these results suggest three qualitatively distinct failure modes in LLM reasoning. Items (1) and (3) form a partition of leader errors under our operational definition, while item (2) captures an orthogonal, non-exclusive diagnostic property related to calibration:

1. **Selection errors** (41.4% of errors): the correct hypothesis is present among alternatives (top-2) but not selected as the final answer.
2. **Calibration/confidence pathologies** (diagnostic; e.g., 13.2% wrong even when unanimous): internal agreement is misaligned with correctness.
3. **Hypothesis-space limitations** (58.6% of errors; proxied by gold unsurfaced in the top-2): the model converges on an internally coherent but incorrect explanation under the tested sampling regime.

Importantly, only the first class is addressable through improved selection or aggregation, while the latter two require changes to training objectives or hypothesis representation.

6.2 Relation to Prior Work

- **Self-consistency.** Wang et al. (2022) show that self-consistency can improve accuracy; we show its limits and failure modes in a medical QA setting.
- **Internal uncertainty.** Kadavath et al. (2022) demonstrate internal uncertainty awareness; we show how this uncertainty appears as alternative hypotheses at the answer level.
- **Iterative refinement.** Madaan et al. (2023) and STaR (Zelikman et al. 2023) focus on iterative correction and training; our work focuses on analysis rather than optimization.
- **Selective prediction / abstention.** Related questions arise in selective prediction and abstention for medical QA (Machcha et al. 2025), where the objective is to know when *not* to answer.

7 Conclusion

We evaluated a sequence of hypotheses about ensemble reasoning and selective decision-making. In this setting, majority-vote aggregation did not yield a statistically significant accuracy improvement over greedy decoding (H1), while correct answers frequently appeared among alternative hypotheses (H2).

Strong internal consensus did not guarantee correctness (H3), motivating the interpretive possibility of hypothesis-space errors (H3′), though we do not yet isolate that mechanism empirically. We also found partial support for H4: selective leader override appears feasible in low-consensus regimes with favorable error budgets, but the evidence is limited to the single dataset, model, and sampling configuration tested here. Together, these results underscore that benchmark accuracy can obscure qualitatively distinct failure modes that cannot be addressed by aggregation alone, calling for a distributional view of reasoning rather than reliance on a single trajectory.

8 Future Work

- learning non-oracle override decision policies that operate within predefined error budgets;
- richer uncertainty signals and hypothesis-set representations beyond vote counts;
- scaling evaluations to larger models, datasets, and longer reasoning horizons; and
- generalization of override policies to other tasks and clinical domains.

Reproducibility and Code Availability

All experiments reported in this paper are fully reproducible. A public, versioned snapshot of the complete codebase is available at `https://github.com/victorlavrenko/rofa/tree/paper/from-answers-to-hypotheses-v1`.

All results were obtained using the fixed model, dataset, prompt, and sampling configuration described in Section 3, evaluated on a deterministic 400-question subset of the MedMCQA validation split.

Reproduction is supported by two dedicated notebooks:

- Reproduction notebook (analysis only): `20_paper_reproduce.ipynb` [Colab ↗]
- Generation notebook (optional, stochastic): `10_colab_generate.ipynb` [Colab ↗]

References

Goh, Ethan, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen (Oct. 2024). “Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial”. In: *JAMA Network Open* 7.10, e2440969. DOI: `10.1001/jamanetworkopen.2024.40969`.

Kadavath, Saurav, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan (2022). *Language Models (Mostly) Know What They Know*. arXiv: `2207.05221 [cs.CL]`.

Machcha, Sravanthi, Sushrita Yerra, Sharmin Sultana, Hong Yu, and Zonghai Yao (Nov. 2025). “Do Large Language Models Know When Not to Answer in Medical QA?” In: *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*. Suzhou, China: Association for Computational Linguistics, pp. 27–35. DOI: `10.18653/v1/2025.uncertainlp-main.4`.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark (2023). *Self-Refine: Iterative Refinement with Self-Feedback*. arXiv: 2303.17651 [cs.CL].

McDuff, Daniel, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan (June 2025). "Towards Accurate Differential Diagnosis with Large Language Models". In: *Nature* 642.8067, pp. 451–457. doi: 10.1038/s41586-025-08869-4.

Pal, Ayush, L. Umapathi, and M. Sankarasubbu (2022). *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering*. arXiv: 2203.14371 [cs.CL].

Tu, Tao, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan (June 2025). "Towards Conversational Diagnostic Artificial Intelligence". In: *Nature* 642.8067, pp. 442–450. doi: 10.1038/s41586-025-08866-7.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou (2022). *Self-Consistency Improves Chain-of-Thought Reasoning in Language Models*. arXiv: 2203.11171 [cs.CL].

Zelikman, Eric, Yuhuai Wu, Jesse Mu, and Noah D. Goodman (Mar. 2023). "STaR: Bootstrapping Reasoning with Reasoning". In: arXiv: 2203.14465 [cs.CL].