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 medical QA benchmark, we show that (i) majority-vote self-consistency does not yield statistically significant accuracy gains over greedy decoding in our 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 characterize an alternative failure mode—*errors of hypothesis space*—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*—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.

2 Experimental Setup

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:**

$$\text{Top-}k = \mathbb{1}\{y_i \in \text{Top-}k(P_i)\}.$$

All results are computed over 400 questions using a fixed prompt and model configuration.

3 Hypothesis H1: 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:

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

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 robust in this setting. Aggregation alone is therefore insufficient as a general reliability mechanism.

4 Hypothesis H2: 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**:

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

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.

5 Hypothesis H3: 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.

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

6 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%.

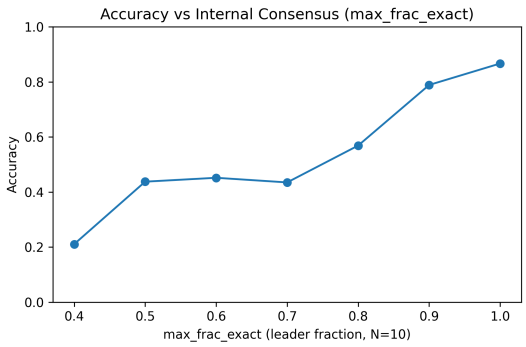

Figure 1: Accuracy as a function of internal consensus (max_frac). Even near-unanimous cases exhibit non-zero error rates.

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

7 An Alternative Explanation: Errors of Hypothesis Space

The failure to find evidence supporting H3 motivates an alternative explanation for these errors:

H3′. Some errors arise because the correct hypothesis is poorly represented or 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 but globally misaligned with the task. Such errors are not reducible to stochastic selection noise and therefore cannot be remedied by increased sampling or voting alone.

In such cases, increased sampling or stronger aggregation cannot recover the correct answer. This reframes aggregation as a *diagnostic probe* rather than a solution.

Taken together, our results suggest three qualitatively distinct failure modes in LLM reasoning:

1. **Selection errors**, where the correct hypothesis is present among alternatives but not selected as the final answer.
2. **Calibration or confidence errors**, where internal agreement or confidence is misaligned with correctness.
3. **Hypothesis space errors**, where the model converges on an internally coherent but systematically incorrect explanation.

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

8 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.

9 Implications and Future Work

These results suggest that safe deployment of LLMs in medicine requires decision rules that operate on *sets of hypotheses*, not single answers. Future work should focus on:

- explicit modeling of hypothesis spaces;
- training objectives that penalize confident convergence on incorrect hypotheses; and
- interfaces that expose structured uncertainty to clinicians.

Reproducibility and Code Availability

All experiments reported in this paper are fully reproducible. We release a public, versioned snapshot of the complete codebase at `https://github.com/victorlavrenko/rofa/releases/tag/paper/from-answers-to-hypotheses-v1`.

A dedicated reproduction notebook, `notebooks/20_paper_reproduce.ipynb`, reproduces all reported figures and statistical analyses using the released model outputs, without requiring rerunning model generation. For full end-to-end reproduction, including stochastic model generation, the generation notebook `notebooks/10_colab_generate.ipynb` can be executed prior to the reproduction notebook.

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.

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].