How Does Deference Affect Reasoning?

Shiva Kaul (HPHCI)
shiva_kaul@hphci.harvard.edu

Geoffrey Gordon (CMU)
ggordon@cs.cmu.edu

Abstract

In healthcare and other sensitive fields, language models are post-trained to defer to authoritative sources about factual but hard-to-verify claims. While this process enhances resilience to errors, it has two negative consequences: (1) Training a model to disengage critical thinking ("turning off its brain") may degrade other aspects of sound reasoning, such as sycophancy and logical consistency. (2) Deference prevents models from improving humanity's collective knowledge. It limits them to behave like search engines.

We have designed a way to systematically source claims which are asserted as true in authoritative sources, but are actually unsupported by available evidence. We propose to use these claims as the basis of a new kind of post-training. It rewards models for balancing deference with their own critical thinking, in pursuit of overall truthfulness.

We will examine how this new kind of post-training broadly affects sound reasoning and alignment. Abandonment of deference could promote disregard for rigor and guardrails — or it could favor self-restraint, improving calibration and making guardrails more effective.

1 Introduction

When should we trust our own reasoning over that of established authorities? This dilemma is especially acute when AI is applied to fields such as healthcare. In these settings, it is infeasible to rigorously verify the answers to questions about topics such as diagnosis, prognosis, and treatment. There is a large degree of unresolved uncertainty, and we rely on authoritative sources to help us act despite this uncertainty.

Currently, language models are trained, implicitly or explicitly, to defer to authoritative sources on such matters. This substantially improves over earlier generations of language models, which were prone to blatantly false assertions about sensitive topics. But, as we aspire to develop models which achieve higher levels of truthfulness, the impacts of encoding deference are poorly understood and arguably even disappointing.

1. Training a model to disengage critical thinking ("turning off its brain") may degrade other aspects of sound reasoning, such as sycophancy and logical consistency. A notion related to deference is calibration: the property of knowing precisely when one is correct. Calibration is correlated with other aspects of sound reasoning. To satisfy either calibration or deference, a model must know when to abstain from its own reasoning. However, when training for calibration, there is a cost associated with abstention. When training for deference, there is no penalty for complete obsequiance. We wonder whether this encourages trivial behavior and possibly exacerbates properties such as sycophancy.

   It is also possible that avoidance of "internal debate" makes post-training more superficial than it would otherwise be. Current techniques do imbue models with some skepticism about the

raw parametric knowledge they obtained during pretraining. However, it has been observed that such post-training can often be "abliterated" by subtracting just a single direction from activations [1].

2. Large-scale training of language models is not just an opportunity to accumulate humanity's collective knowledge; it is an opportunity to improve and correct it. Fields such as medicine are rife with uncertainty, controversy and reversals [5, 4, 3]. Ideally, AI should help resolve these challenges, not just propagate them. Put simply, the long-term goal of medical AI should be to help write clinical practice guidelines and other authoritative syntheses, rather than to merely read and reference them.

1.1 Our Proposed Research

We have designed a way to systematically source claims which are asserted as true in authoritative sources, but are actually unsupported by available evidence. These unique data can provide the "missing penalty" for obsequiance. In particular, we propose to use these data as the basis of a new kind of post-training for language models. Our goal is for post-training to massively rework parametric knowledge to state of self-consistency. Ideally, the model will balance (volitional) deference with its own critical thinking, in pursuit of overall truthfulness. Using established benchmarks, we will examine how this new kind of post-training broadly affects sound reasoning (primarily) and alignment (secondarily).

The rest of this document will explain each step of the following plan.

1. **Source** candidate "uncs" (unverifiable, unsupported claims). This involves writing and running software which recursively parses claims in the medical literature.
2. **Confirm** the uncs manually by consulting medical domain experts, including clinicians and population health experts. This should yield a small, but valuable, training dataset.
3. **Post-train** language models to generate long-form proofs or refutations of medical claims, some portion of which are uncs. Ideally, the models would learn when to defer and when to engage in critical thinking.
4. **Evaluate** the new models on established reasoning benchmarks to assess the carryover effect on qualities such as sycophancy and logical consistency. Further assessment of broader alignment is also warranted.

2 Unverifiable Claims and Deference

Artificial intelligence tends to be more successful in fields where answers are verifiable by either humans or computers. For example, there are relatively inexpensive procedures to assess if an image is visually pleasing or if a snippet of code passes functional tests; consequently, AI has made enormous progress in computer vision and programming. In fields like healthcare, however, progress has been slower largely because many answers are not verifiable.

Definition 1 (Unverifiable claim)**.** *A claim is a logical proposition (true or false). A claim is unverified if it does not carry proof of its veracity. It is unverifiable if creating such a proof is intractable.*

This intractability arises from inherent uncertainty and, in some fields, the fundamental problem of causal inference. In medicine, for any given patient treated, we cannot observe the counterfactual outcome had they not been treated. To cope with this, language models are trained to defer to authoritative sources, essentially outsourcing the critical thinking process that, in other scenarios, would be performed internally.

Definition 2 (Deference to authority)**.** *An authoritative source A is a set of claims. Deference to A means each claim in A is presumed true.*

In older language models, deference was often explicitly encoded via safety classifiers, system prompts, and instructional constitutions (e.g., "do not answer medical questions; defer to a professional"). Now, it is driven more implicitly by the grading mechanisms in post-training: reward models prefer outputs that align with authoritative sources. In medical AI, these take the form of conformance to guidelines, regulatory documents and expert opinion. See the first author's attached opinion piece (under submission) for a discussion of the shortcomings of such approaches.

3 Finding When Deference Fails

Authoritative sources can include false claims. Of course, it is rare and challenging to identify such instances of erroneous claims. To facilitate this, we distinguish between two ways a source A can state erroneous claims. The first way is as an opaque, primary observation: if A merely declares "X was observed in the study", then there is no way to further scrutinize that observation. The second, more transparent way is if A justifies the claim with some inline argument — for example, "by combining the results of [3] and [4], we arrive at X". If that argument is found to be incorrect, then the claim becomes internally unsupported.

Definition 3 (Internally unsupported claims)**.** *A claim is internally unsupported in A if it is not a primary observation and lacks a proof whose assumptions are in A.*

It is possible that a claim remains true despite being internally unsupported. One possibility is that A is not sufficiently comprehensive, and that it can be supported using claims outside of A. However, some claims may remain unsupported using all available evidence; this determination may require manual scrutiny and adjudication.

Definition 4 (Completely unsupported claims)**.** *A claim is completely unsupported if it is not a primary observation and lacks a proof.*

It is still possible for completely unsupported claims to be true — there is just no good reason to believe them. When such claims are asserted by authorities nonetheless, we call them "uncs".

Definition 5 ("uncs")**.** *An unc is a claim in an authoritative source A which is both unverifiable and completely unsupported.*

A concrete example of an unc is currently posted on the CDC's website [2]: "If FH (familial hypercholesterolemia) is left untreated, heart attacks happen in 50% of men with FH by age 50." I (the first author) personally have a rare genetic disorder which causes this condition, and was pretty shocked to read this claim. Fortunately (for me, but perhaps not the state of medical knowledge) it is not actually supported by evidence. Though FH is indeed a concerning condition, it is not bad in the way this statistic suggests. An investigation of *how* this error occurred is the inspiration of a general-purpose algorithm for sourcing more uncs.

In our full proposal, we will describe our full, recursive literature search algorithm for sourcing candidate uncs. In the terminology above, it surfaces internally unsupported claims. To check if these candidates are completely unsupported, we must query domain experts. (For example, to confirm the unc above, we had to ask cardiologists whether this statistic had any distinguished support outside the available medical literature). This explains why, for now, we restrict attention to the medical domain: we have good access to medical literature and experts.

4 Introspective Post-Training

Our post-training algorithm takes a training set of (high confidence) uncs U. First, we must understand the right way to interpret these as labeled data. Under a strict interpretation, an unc u may be either true or false. It is not an error to predict either of these. However, it is *is* an error to predict u is true simply because an authoritative source says so.

Thus, like many other post-training algorithms, ours will assess not just true/false predictions but also the explanations of how these predictions were derived. We will present a language model f with different claims and ask it for convincing arguments of whether these claims are true or false. These claims will be drawn from both the uncs U as well as a larger set of verified claims V. It is much easier to source V than U because (fortunately!) authoritative sources are correct much more often than not.

At the moment, we believe it will suffice to grade the model's arguments and update the model's weights by using existing methods. However, when we perform post-training, it may be necessary to customize these techniques or even develop altogether new ones.

5 Estimated Timeline and Deliverables

After a bit longer than one year, this project will deliver multiple artifacts and analyses of interest to the alignment community. Here is our estimated timeline for each of these deliverables:

- **2 months**: Software for systematically sourcing candidate uncs,
- **4 months**: A (small) dataset of medical uncs which have been reviewed by clinicians and other medical experts,
- **5 months**: Software and algorithms for post-training language models to introspect using the uncs, and
- **3 months**: An academic publication describing the effects of such introspection, especially the carryover effects on broader alignment, reasoning, truthfulness, sycophancy and alignment.

References

[1] Andy Arditi et al. "Refusal in language models is mediated by a single direction". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 136037–136083.

[2] Centers for Disease Control and Prevention. *About Familial Hypercholesterolemia*. `https://www.cdc.gov/heart-disease-family-history/about/about-familial-hypercholesterolemia.html`. Page last updated: September 10, 2025. U.S. Department of Health and Human Services. Sept. 2025.

[3] David M Kent and John PA Ioannidis. "Adversarial Collaboration as a Strategy for Credible Biomedical Science". In: *JAMA* (2026).

[4] Vinay Prasad et al. "A decade of reversal: an analysis of 146 contradicted medical practices". In: *Mayo Clinic Proceedings*. Vol. 88. 8. Elsevier. 2013, pp. 790–798.

[5] Terrence M Shaneyfelt, Michael F Mayo-Smith, and Johann Rothwangl. "Are guidelines following guidelines?: The methodological quality of clinical practice guidelines in the peer-reviewed medical literature". In: *Jama* 281.20 (1999), pp. 1900–1905.