Representational and Behavioral Stability of Truth in Large Language Models

Samantha Dies[1♯], **Courtney Maynard**[1], **Germans Savcisens**[1], **and Tina Eliassi-Rad**[1,2,3]

[1]Khoury College of Computer Sciences, Northeastern University, 440 Huntington Ave, #202, Boston, MA 02115 USA
[2]Network Science Institute, Northeastern University, 177 Huntington Ave, #1010, Boston, MA 02115 USA
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501 USA
[♯]dies.s@northeastern.edu

ABSTRACT

Large language models (LLMs) are increasingly used as information sources, yet small changes in semantic framing can destabilize their truth judgments. We propose `P-StaT` (Perturbation Stability of Truth), an evaluation framework for testing belief stability under controlled semantic perturbations in representational and behavioral settings via probing and zero-shot prompting. Across sixteen open-source LLMs and three domains, we compare perturbations involving epistemically familiar `Neither` statements drawn from well-known fictional contexts (`Fictional`) to those involving unfamiliar `Neither` statements not seen in training data (`Synthetic`). We find a consistent stability hierarchy: `Synthetic` content aligns closely with factual representations and induces the largest retractions of previously held beliefs, producing up to 32.7% retractions in representational evaluations and up to 36.3% in behavioral evaluations. By contrast, `Fictional` content is more representationally distinct and comparatively stable. Together, these results suggest that epistemic familiarity is a robust signal across instantiations of belief stability under semantic reframing, complementing accuracy-based factuality evaluation with a notion of epistemic robustness.

1 Introduction

Large language models (LLMs) are increasingly used as sources of information, but their behavior often blurs the line between knowledge and plausibility [1, 2]. While humans are expected to distinguish between `True`, `False`, and `Neither`-valued statements, it remains unclear whether LLMs form similarly structured internal representations, or whether such representations predict when truth judgments remain stable under changes in semantic framing [3, 4]. This gap is increasingly consequential as LLMs are deployed in high-stakes informational settings, where instability can undermine reliability and contribute to undesirable behaviors such as hallucinations [5, 6].

Existing work approaches this problem from two largely separate directions. Representation-based probing studies ask whether `True` and `False` statements occupy separable regions in activation space and whether "truth directions" transfer across LLMs or domains [7, 8, 9]. Behavioral studies document that small changes in prompting, context, or adversarial framing can substantially alter an LLM's outputs, including its apparent confidence and self-consistency [10, 11, 12, 13]. While both lines of work reveal important failure modes, they are rarely connected through a shared experimental lens. Thus, we lack a unified framework for testing whether a candidate stability signal in latent space corresponds to stability in LLM behavior [14, 15].

We address this gap by proposing the Perturbation Stability of Truth framework, `P-StaT`, which links internal representations and LLM behavior through the lens of belief stability (Fig. 1). `P-StaT` asks how stable LLM truth judgments remain when semantic assumptions are systematically varied. This perspective is motivated by formal epistemology: under Leitgeb's notion of P-stability, a belief system is epistemically well-formed only if established beliefs are preserved under small, principled changes in evidential context [16].

We operationalize semantic variation through perturbations Θ that modify which `Neither` statements are treated as compatible with truth. We distinguish epistemically *familiar* `Neither` statements (`Fictional`) from *unfamiliar* `Neither` statements (`Synthetic`) across three domains. Crucially, the same perturbation is instantiated both representationally, by retraining probes, and behaviorally, via zero-shot prompting.

We observe a consistent pattern: Unfamiliar `Synthetic` content induces the largest epistemic retractions in both representational and behavioral settings. By contrast, familiar `Fictional` content occupies more distinct regions of activation space and produces substantially fewer retractions. Together, these results identify *epistemic familiarity* as a cross-instantiation signal for when LLM truth judgments are most likely to destabilize under semantic reframing.

Contributions

1. We introduce a dataset of fictional statements across three domains (City Locations, Medical Indications, and Word Definitions) to enable controlled compar-

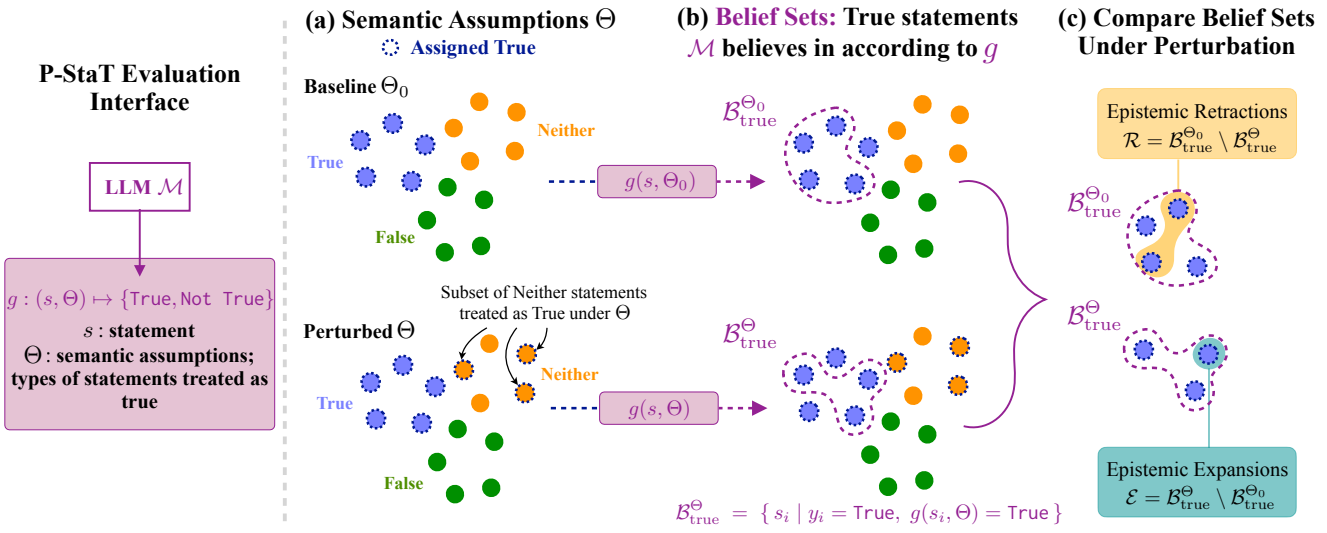

Figure 1. Overview of the P-StaT framework. P-StaT evaluates the stability of an LLM $\mathcal{M}$'s beliefs by using a decision function g to map statements s to `True` or `Not True` under semantic assumptions Θ. **(a)** In the baseline case Θ_0, only `True` statements (purple) are labeled `True`, while `False` (green) and `Neither` (orange) statements are labeled `Not True`. Under a perturbed assumption set Θ, a subset of `Neither` statements is treated as `True`. **(b)** The sets of `True` statements that $\mathcal{M}$ believes to be `True` in the baseline ($\mathcal{B}^{\Theta_0}_{\text{true}}$) and perturbed ($\mathcal{B}^{\Theta}_{\text{true}}$) cases are identified via g. **(c)** Comparing these belief sets reveals unstable beliefs that are either retracted ($\mathcal{R}$) or expanded ($\mathcal{E}$). Crucially, g can be instantiated using either representation-based probing or behavioral zero-shot prompting, enabling direct connections between instability in latent space and behavioral instability.

isons between epistemically familiar and unfamiliar `Neither` content.

2. We propose P-StaT, a perturbation-based stability framework grounded in epistemic notions of belief stability that supports both probing and zero-shot prompting.

3. Across sixteen open-source LLMs, we show that unfamiliar `Synthetic` statements induce the largest epistemic retractions both representationally and behaviorally, identifying epistemic familiarity as a key signal of stability.

2 Related Work

Our work lies at the intersection of three research threads: (i) representation-based probing of veracity, (ii) behavioral sensitivity to semantic context, and (iii) epistemic notions of belief stability.

Representational Probing of Veracity. Probing examines which properties are recoverable from hidden representations of LLMs, providing insights into what LLMs encode beyond observable behavior [17, 18, 19]. Recent work has shown that `True` and `False` statements form separable clusters across LLMs and domains in activation space [7, 8]. The `sAwMIL` framework [9] extends this work by modeling `True`, `False`, and `Neither` statements as distinct representational directions. Related work on hallucination detection similarly suggests that hidden representations encode strong veracity signals even when outputs are wrong [2].

Behavioral Stability and In-Context Effects. A separate line of research documents small changes in wording or framing can lead to large shifts in LLM behavior, including jailbreak vulnerabilities [13], sycophancy [20], and instability under paraphrasing [11]. Recent work has framed this brittleness in terms of behavioral consistency across interactions. For example, Li et al. introduce benchmarks for evaluating response stability in multi-turn settings [12]. Related work on in-context learning suggests that larger LLMs can override pretraining-induced semantics when presented with conflicting in-context labels, while smaller LLMs rely more heavily on prior knowledge [21]. Complementarily, Bigelow et al. propose a Bayesian account in which both prompting and activation steering modify latent beliefs [22].

Epistemic Belief Stability. LLMs also exhibit systematic difficulties in tracking epistemic distinctions such as belief, knowledge, and factuality [3, 4]. Herrmann and Levinstein argue that the study of LLM beliefs lacks unified criteria for when internal representations should count as belief-like [15]. Formal epistemology offers a complementary perspective: Leitgeb's notion of *P*-

stability characterizes rational belief systems as those that preserve established beliefs under small, *justified* changes in evidential context [16].

By combining representational probing with behavioral perturbations through `P-StaT`, our work bridges prior studies of latent veracity structure and in-context sensitivity. We assess how beliefs respond to principled shifts in semantic assumptions, providing a unified view of epistemic stability across representations and behavior.

3 Methodology

We study how LLMs organize statements in activation space and whether this organization predicts the stability of beliefs under semantic reframing using `P-StaT` (Perturbation Stability of Truth), which applies the same perturbations in both probing and zero-shot evaluations.

3.1 Epistemic Familiarity and `Neither` Statements

Let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ denote declarative statements drawn from factual domains with labels $y_i \in \{\texttt{True}, \texttt{False}, \texttt{Neither}\}$. Let $\mathcal{N} = \{s_i \in \mathcal{S} \mid y_i = \texttt{Neither}\}$ denote the set of `Neither` statements. While all statements in $\mathcal{N}$ lack real-world truth value, $\mathcal{N}$ is heterogeneous and can be partitioned into familiar (`Fictional`) and unfamiliar (`Synthetic`) subsets:

$$\mathcal{N} = \mathcal{N}_{\text{fam}} \cup \mathcal{N}_{\text{unf}}, \qquad \mathcal{N}_{\text{fam}} \cap \mathcal{N}_{\text{unf}} = \emptyset.$$

Here, $\mathcal{N}_{\text{fam}}$ contains well-known fictional entities likely present in training corpora, while $\mathcal{N}_{\text{unf}}$ contains constructed entities intended to be absent from training data. For `Fictional`, we additionally distinguish canonically true and false subsets, $\mathcal{N}_{\text{fam}}^{(T)}$ and $\mathcal{N}_{\text{fam}}^{(F)}$.

We evaluate three domains that differ in how sharply truth and falsehood are delineated (City Locations, Medical Indications, Word Definitions; Table 1) [7, 8, 9]. `True`, `False`, and `Synthetic` statements originate in [9], while we introduce the new `Fictional` dataset.[1] Details on data construction and validation are in Supplementary Section B.

3.2 Veracity Representations

For an LLM $\mathcal{M}$, we define a representation map $\phi_{\mathcal{M},l} : s_i \mapsto \mathbf{z}_i^{(l)}$ between a statement s_i and its token-level hidden representations at layer l. For each ⟨dataset, LLM⟩ pair we select l to maximize linear separability between `True` and `Not True` statements [7, 8, 9]. We refer to $\mathbf{z}^{(l)}$ as *veracity representations*. Together with statements and labels, they define the dataset $\mathcal{D} = \{(s_i, \mathbf{z}_i^{(l)}, y_i)\}_{i=1}^{N}$, which serves as the input to representational and behavioral experiments.

[1]The new fictional dataset can be accessed at https://huggingface.co/datasets/samanthadies/representational_stability.

3.3 `P-StaT`: Perturbation Stability of Truth

`P-StaT` evaluates stability for a fixed LLM $\mathcal{M}$ by assuming the evaluation procedure

$$g : (s_i, \Theta) \mapsto \{\texttt{True}, \texttt{Not True}\},$$

where Θ encodes semantic assumptions about which statements should be treated as compatible with truth. The $g(\cdot, \Theta)$ interface is the core abstraction of `P-StaT`, enabling the same semantic perturbation to be instantiated in distinct evaluation settings.

Representational vs. Behavioral Stability. In representational experiments, g is implemented as $g_{\text{p}}(s, \Theta) = h_\Theta(\phi_{\mathcal{M},l}(s))$, where h_Θ is a linear probe trained with labels induced by Θ. In behavioral experiments, g is implemented as $g_{\text{zs}}(s, \Theta)$ via zero-shot prompting and a *belief context* C_Θ. Both settings support the same perturbations, enabling direct comparison.

Baseline evaluation. Under a baseline semantic interpretation Θ_0, g identifies the set of ground-truth `True` statements that are assigned `True`:

$$\mathcal{B}_{\text{true}}^{\Theta_0} = \{\, s_i \mid y_i = \texttt{True},\ g(s_i, \Theta_0) = \texttt{True} \,\}.$$

Perturbations. A perturbation modifies the semantic assumptions encoded in Θ. Let $\mathcal{N}_\Theta \subseteq \mathcal{N}$ denote the subset of `Neither` statements treated as compatible with truth under a perturbed interpretation. The perturbed interpretation Θ differs from Θ_0 only in this reassignment and yields the perturbed belief set:

$$\mathcal{B}_{\text{true}}^{\Theta} = \{\, s_i \mid y_i = \texttt{True},\ g(s_i, \Theta) = \texttt{True} \,\}.$$

Thus, only the semantic interpretation of `Neither` content is altered.

Epistemic retractions. Within `P-StaT`, we quantify stability by comparing the baseline and perturbed belief sets. Epistemic retractions, $\mathcal{R} = \mathcal{B}_{\text{true}}^{\Theta_0} \setminus \mathcal{B}_{\text{true}}^{\Theta}$, capture ground-truth `True` statements that lose belief status under perturbation. While we also consider epistemic expansions, $\mathcal{E} = \mathcal{B}_{\text{true}}^{\Theta} \setminus \mathcal{B}_{\text{true}}^{\Theta_0}$, for completeness, epistemic retractions are a stronger signal of belief instability than expansions because they withdraw previously held beliefs [16].

4 Experiments

We use `P-StaT` to apply the same perturbation Θ to both (i) probe-based evaluations over $\phi_{\mathcal{M},l}$ and (ii) zero-shot evaluations via belief context. We implement all experiments in Python using PyTorch [23], NumPy, SciPy, and scikit-learn [24], and rely on the HuggingFace Transformers ecosystem for LLM access and data handling [25]. All experiments were run on a university HPC cluster with NVIDIA `H200` GPUs and required approximately 36 GPU-hours (activation extraction and zero-shot evaluations), with probe training performed on CPU.