

A Single Direction of Truth: An Observer Model’s Linear Residual Probe Exposes and Steers Contextual Hallucinations

Charles O’Neill¹ Slava Chalnev² Chi Chi Zhao¹ Max Kirkby¹ Mudith Jayasekara¹

Abstract

Contextual hallucinations — statements unsupported by given context — remain a significant challenge in AI. We demonstrate a practical interpretability insight: a generator-agnostic observer model detects hallucinations via a single forward pass and a linear probe on its residual stream. This probe isolates a single, transferable linear direction separating hallucinated from faithful text, outperforming baselines by 5–27 points and showing robust mid-layer performance across Gemma-2 models (2B→27B). Gradient-times-activation localises this signal to sparse, late-layer MLP activity. Critically, manipulating this direction causally steers generator hallucination rates, proving its actionability. Our results offer novel evidence of internal, low-dimensional hallucination tracking linked to specific MLP sub-circuits, exploitable for detection and mitigation. We release the 2000-example CONTRATALES benchmark for realistic assessment of such solutions.

1 Introduction

Large language models (LLMs) have made striking progress in open-ended generation, yet they continue to produce statements that are *plausible but unsupported* by their given context – a failure mode broadly labelled *hallucination* (Huang et al., 2024; Ji et al., 2023). When these systems are deployed in medicine, finance, or law, even an isolated hallucination can trigger costly or harmful downstream actions, eroding user trust and slowing adoption. Detecting hallucinations *after* text has been produced is therefore a key practical requirement, both for automated self-monitoring and for post-hoc auditing workflows (Valentin et al., 2024).

Unfortunately, most high-performing detectors to date ei-

¹Parsed, London, UK ²Independent. Correspondence to: Charles O’Neill <charles@parsed.com>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

ther (i) require privileged access to the **generator model**’s parameters and logits – an unrealistic assumption for commercial APIs and safety-critical settings – or (ii) fall back on brittle surface cues such as lexical overlap, unseen named entities, or low embedding similarity between source and output. These heuristics falter whenever hallucinations manifest as subtle logical contradictions rather than obvious factual novelties, a gap that is starkly revealed by synthetic logic benchmarks such as our CONTRATALES. This highlights the need for robust, interpretability-based approaches that offer deeper insights beyond surface cues for more reliable detection

Recent work in mechanistic interpretability argues that transformers internally encode rich world knowledge along *approximately linear* directions in activation space (Burns et al., 2023; Park et al., 2023). Linear or sparse probes applied to hidden states have uncovered directions for factual truth, policy compliance, and other high-level properties (Alain & Bengio, 2016; Marks et al., 2024). Yet, existing studies either rely on signals *inside* the generator or do not test whether the discovered directions are (i) *consistent* across model sizes and domains, (ii) *causal*—that is, capable of steering generation when manipulated—and (iii) sufficiently salient to be decoded by small observer models, limiting their direct actionability in practical, generator-agnostic scenarios.

We ask: does a transformer *notice* when a span of text contradicts its earlier context, and if so, is that realisation encoded along a single, linearly accessible axis that can be read out by another model? We cast this as a mechanistic hypothesis and test it directly. Concretely, we introduce the **observer paradigm**: a *separate, frozen* language model ingests an arbitrary document consisting of a source passage followed by a candidate continuation and, in a single forward pass, predicts whether the continuation is contextually supported. Our detector uses nothing more than a linear probe on the observer’s residual-stream activations at the full stop of the final sentence.

Across standard news-domain summarisation sets (CNN/DAILYMAIL, XSUM) and an out-of-domain corpora (our introduced dataset CONTRATALES) our linear probe approach achieves F₁ scores up to 0.99 on

news and 0.84 on CONTRATALES, significantly beating lexical-overlap, entity-verification, semantic-similarity, and attention-pattern baselines by 5–27 points. Layer sweeps reveal a broad mid-layer performance plateau that is *shared* across model sizes (Gemma-2 2B→27B) and datasets. Attribution analysis based on gradient-times-activation localises the signal for contextual inconsistency to a sparse pattern of late-layer MLP activity, whose characteristics are notably stable across diverse datasets. Most importantly, causal interventions that ablate or inject the learned direction in a generator modulate hallucination rates, demonstrating that the axis captures a functionally meaningful representation.

Contributions.

1. We provide the first *generator-agnostic* detector that identifies contextual hallucinations with a *single forward pass*, requiring no knowledge of the text’s origin and no sampling overhead, demonstrating a practical application of interpretability for efficient AI monitoring.
2. We show that a *single, highly linear* residual-stream direction robustly separates hallucinated from supported spans across domains and remains readable by comparatively small observer models, offering an actionable insight into model representations.
3. Through gradient-times-activation attribution, we map the representation of contextual inconsistency to a compact and consistent pattern of late-layer MLP activity, and demonstrate its sparsity and stability across key datasets.
4. We establish *causality*: manipulating activity along this axis in a generator steerably reduces or amplifies hallucination prevalence.
5. To contribute to developing realistic benchmarking methods for actionable interpretability, we release CONTRATALES, a 2000-example benchmark of purely logical contradictions tailored to stress-test contextual hallucination detectors.

2 Background and Related Work

Large language models (LLMs) frequently generate plausible yet factually incorrect content that contradicts their input or world knowledge, a phenomenon known as hallucination (Huang et al., 2024; Ji et al., 2023). Despite remarkable capabilities in text generation and comprehension, this tendency to hallucinate significantly undermines model reliability in high-stakes domains (Ji et al., 2023). We focus on *intrinsic hallucinations*, which directly contradict

provided source content, as opposed to *extrinsic hallucinations*, which introduce unverifiable external information (Ji et al., 2023; Huang et al., 2024).¹

2.1 Hallucination Detection Methodologies

Existing detection methods can be categorised by their access requirements to the generating model and by underlying techniques.

Generator-internal approaches require access to the generator’s internals, encompassing token-probability analysis, calibrated confidence estimation, and internal-state probing (Azaria & Mitchell, 2023). Recent MI-based work extends this line, with Yu et al. (2024) identifying drifting sub-modules for non-factual outputs, and Sun et al. (2024) tracing retrieval versus parametric knowledge to flag hallucinations in RAG systems. However, such methods are inapplicable to black-box APIs (Valentin et al., 2024). Post-hoc external methods operate on generated text alone, and include checking claims against knowledge bases, using entailment models, or enlisting an LLM judge (Huang et al., 2024; Valentin et al., 2024). Sampling-based consistency, such as SelfCheckGPT (Manakul et al., 2023), generates and compares multiple outputs for agreement but is computationally intensive.

Embedding/representation-based methods exploit vector representations or internal states without needing generator weight access. Examples include semantic similarity checks of source and output embeddings (Ji et al., 2023), internal-state distribution analysis for hidden-state drift (Farquhar et al., 2024), and efficient linear classifiers like Semantic Entropy Probes (SEPs) (Kossen et al., 2024). Hallucination-resistant finetuning along “hallucination directions” has also been demonstrated (Research, 2025). The ongoing challenge is to design detectors that are simultaneously efficient, post-hoc, and generator-agnostic.

2.2 Utilising LLM Internals

Linear Representation Hypothesis The *linear representation hypothesis* (LRH) posits that a transformer’s activation space can be approximated as a sparse sum of *linearly separable feature directions*, first articulated for toy models by Elhage et al. (2022) and formalised for language models by Park et al. (2023). Empirical support includes sparse-autoencoder (SAE) studies showing that few orthogonal, near-monosemantic directions can reconstruct most hidden-state variance (Makhzani & Frey, 2013; Cunningham et al., 2023), and linear or low-rank probes isolating causal directions for truthfulness, gender bias, and chain-of-thought features (Alain & Bengio, 2016; Nanda et al., 2023; Marks

¹Readers new to mechanistic interpretability (MI) will find a concise overview in Rai et al. (2024).

et al., 2024). Nonetheless, competing evidence, such as observations of multidimensional toroidal embeddings in Llama and Mistral (Engels et al., 2024), suggests the LRH may be incomplete.

Probing Probing operationalises the LRH by training lightweight decoders on hidden states to predict an external label. Originating in visual-cortex studies using linear classifiers on biological neurons (Mur et al., 2009), it was adapted to ANNs by Alain & Bengio (2016) and is now an interpretability research staple. A growing body of work demonstrates that concepts like factual truth (Burns et al., 2023), policy compliance (Azaria & Mitchell, 2023), and even sleeper-agent triggers (Hubinger et al., 2024) are linearly decodable, especially from middle-to-late layers, consistent with transformer-circuits analyses (Cammarata et al., 2020). In the context of hallucination, Azaria & Mitchell (2023) also investigated internal states for related properties. Simhi et al. (2024) demonstrated probe use across most layers and components in a 7B model, though with limited transferability and a narrow focus on unambiguous answers. While many approaches link hallucination to model uncertainty (Farquhar et al., 2024), models can also exhibit high-certainty hallucinations (Simhi et al., 2025). Other relevant research has shown LLMs internally encode question-answerability (Slobodkin et al., 2023), or has emphasised latent-knowledge awareness (Ferrando et al., 2024) and controllable context reliance (Minder et al., 2024).

Attention Attention patterns, which reveal information flow, can also detect contextual hallucinations. For example, Lookback Lens quantifies the balance between context-focused and self-focused heads to detect contextual hallucinations (Chuang et al., 2024). Similar attention-based mechanisms underpin causal editing efforts by Ferrando et al. (2024), controllable context-sensitivity work by Minder et al. (2024), and Yuksekgonul et al. (2024) found a strong positive correlation between an LLM’s attention to relevant prior tokens and the factual accuracy of its generations.

3 Methods

This section details the datasets, our proposed linear probing methodology for detecting contextual hallucinations, the baseline methods used for comparison, and the unit-level attribution technique employed to interpret the probe.

3.1 Datasets

Our core task is detecting contextual inconsistencies. In our observer paradigm, this involves a model identifying text spans within a concatenated input-output sequence that either contradict or lack support from the preceding context.

The observer model processes this unified document to detect such unsupported claims. We evaluate this capability using four datasets, summarised in Table 1.

News Summarisation Datasets We use CNN/Daily Mail (CNN/DM) (See et al., 2017) and XSum (Narayan et al., 2018), standard abstractive summarisation benchmarks. CNN/DM contains news articles with multi-sentence summaries. XSum features BBC articles with single-sentence summaries, often requiring higher abstraction.

Synthetic Contradictions (ContraTales) Story prefixes were initially generated using Claude Opus. Then, for these prefixes, factual and hallucinated concluding sentences (outputs) were generated by o4-mini . Hallucinated examples feature a concluding sentence that logically contradicts information established in the prefix (e.g., stating a character who is bald is going for a haircut). This dataset provides unambiguous logical contradictions. See Appendix B for full details of constructing this dataset.

Data Preparation For each dataset, we prepared paired examples of a source context and a continuation (output text). Continuations were generated to be either: (i) factual, containing only information directly inferable from the source, or (ii) hallucinated. Hallucinated continuations were produced by prompting gpt-4.1 to introduce a plausible but unsupported or contradictory factual detail, while maintaining grammaticality and a sentence length under 40 words. Factual continuations were generated with prompts emphasising strict adherence to the source material, again using gpt-4.1 . See Appendix A for the prompts used to generate continuations.

During evaluation, the source context and its corresponding continuation are concatenated and processed as a single sequence by the observer model.

3.2 Residual-Stream Linear Probe Methodology

Evaluation Protocol Unless otherwise specified, all detection methods, including our proposed probe and the baselines, are evaluated using a logistic regression classifier trained to distinguish between factual and hallucinated continuations. Performance is assessed via 5-fold cross-validation. A fixed random seed was used for data splitting and sampling across all experiments to ensure reproducibility.

Residual-stream linear probe Given a frozen observer transformer \mathcal{F} with L decoder blocks and model dimension d , let $\mathbf{r}_t^{(\ell)} \in \mathbb{R}^d$ denote the post-layer-norm residual stream at token position t after block ℓ . For each example, we concatenate the source context X and its candidate continuation Y , feed the resulting sequence $(x_{0:T-1})$ through \mathcal{F} ,