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

Charles O'Neill [1] **Slava Chalnev** [2] **Chi Chi Zhao** [1] **Max Kirkby** [1] **Mudith Jayasekara** [1]

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 either (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 F_1 scores up to 0.99 on

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

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

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

[1] Readers 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 $\mathcal{F}$,

Table 1: Summary of datasets used for evaluating hallucination detection.

Dataset	Input	Output Type	Hallucination Type	Continuation Generator	Examples
CNN/DM	News article	Bullet-point summary	Fabricated detail	`gpt-4.1-mini`	1,000
XSum	BBC article	One-sentence summary	Fabricated detail	`gpt-4.1-mini`	1,000
ContraTales	Story prefix	Concluding sentence	Logical contradiction	`o4-mini`	2,000

and identify the index of the final token (typically a full stop) of the last sentence in the continuation, $t^* = T - 1$.

From a specific layer ℓ^* (selected via inner-fold validation on the training set), we extract the activation $\mathbf{h} = \mathbf{r}_{t^*}^{(\ell^*)}$. A logistic probe, parametrised by weights $\mathbf{w} \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$, then predicts the probability of hallucination:

$$\widehat{y} \;=\; \sigma(\mathbf{w}^\top \mathbf{h} + b), \quad \text{where} \quad \sigma(z) = \frac{1}{1+e^{-z}}.$$

The probe is trained using binary cross-entropy loss with L_2 regularisation on $\mathbf{w}$. At test time, the logit $s = \mathbf{w}^\top \mathbf{h} + b$ serves as the input for a hard decision threshold (typically $s > 0$ for hallucination) and as a continuous hallucination score.

3.3 Baselines

We compare our probe against several baselines. Each baseline generates a feature (or set of features) which is then used to train a logistic regression classifier, following the evaluation protocol described in §3.2.

Lexical Overlap Novelty is quantified as $\nu(Y, X) = 1 - \frac{|\text{n-grams}(Y) \cap \text{n-grams}(X)|}{|\text{n-grams}(Y)|}$ (Maynez et al., 2020). This is computed for $n \in \{1, 2, 3\}$, and the maximum score is used as the feature.

Entity Verification The entity novelty ratio is $\eta(Y, X) = \frac{|E_Y \setminus E_X|}{|E_Y|}$, where E_X and E_Y are sets of named entities extracted from input X and output Y using spaCy (Honnibal et al., 2020; Nan et al., 2021).

Semantic Similarity For the final sentence s_F in the continuation Y, we compute its maximum cosine similarity to any sentence in the input X: $\phi(s_F, X) = \max_{x_j \in X} \cos(\mathbf{e}(s_F), \mathbf{e}(x_j))$. Sentence embeddings $\mathbf{e}(\cdot)$ are from OpenAI's `text-embedding-3-small`. The resulting similarity score is used as a feature. (Note: The original paper mentioned flagging segments below a threshold τ; here, we use the score directly as a feature for the logistic regression, which learns the optimal threshold/weighting).

Lookback Lens Proposed by Chuang et al. (2024), this method analyses attention patterns. Given concatenated context X (length N) and continuation Y, for each head (l, h) at position t in Y, it computes average attention to context tokens $A_t^{l,h}(\text{context})$ and to preceding tokens in Y, $A_t^{l,h}(\text{new})$. The lookback ratio is $\text{LR}_t^{l,h} = \frac{A_t^{l,h}(\text{context})}{A_t^{l,h}(\text{context}) + A_t^{l,h}(\text{new})}$. For the final sentence of Y, these ratios are averaged over its tokens for each head, forming a feature vector $\overline{\mathbf{v}}$ by concatenating these values across all heads and layers. This vector $\overline{\mathbf{v}}$ is input to the logistic regression classifier. Our observer paradigm processes the concatenated text; N is the length of the original source context X.

3.4 Unit-Level Hallucination Attribution

To identify transformer sub-modules influencing the hallucination score, we use gradient-times-activation.

Notation. Let the probe score for a sequence, derived from the optimal layer ℓ^* and final token t^* as defined in §3.2, be $s = \mathbf{w}^\top \mathbf{r}_{t^*}^{(\ell^*)} + b$. For any layer $\ell \in \{0, \ldots, L-1\}$ and token position t:

$\mathbf{r}_t^{(\ell)} \in \mathbb{R}^d$	post-LN residual stream,
$\mathbf{z}_t^{(\ell,h)} \in \mathbb{R}^d$	output of head h in layer ℓ (after W^O),
$\mathbf{m}_t^{(\ell)} \in \mathbb{R}^d$	output of the MLP in layer ℓ (after W^{out}).

Gradient–times–activation. We compute gradients of the probe score s with respect to intermediate residual stream activations: $\mathbf{g}_t^{(\ell)} = \partial s / \partial \mathbf{r}_t^{(\ell)}$. The contribution $a_t(\mathbf{u})$ of a module's output vector $\mathbf{u}_t$ (where $\mathbf{u}_t \in \{\mathbf{z}_t^{(\ell,h)}, \mathbf{m}_t^{(\ell)}\}$ which contributes to a subsequent residual stream $\mathbf{r}_t^{(\ell')}$) is taken as its projection onto the gradient of that residual stream: $a_t(\mathbf{u}) = \langle \mathbf{g}_t^{(\ell')}, \mathbf{u}_t \rangle$. This value indicates if the module's output nudges the relevant residual stream in the direction that increases (positive) or decreases (negative) the hallucination score s. These contributions are averaged over the tokens $\mathcal{S}$ of the final sentence in the continuation:

$$A_{\text{head}}^{(\ell,h)} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} a_t(\mathbf{z}^{(\ell,h)})$$

$$A_{\text{mlp}}^{(\ell)} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} a_t(\mathbf{m}^{(\ell)}).$$

Dataset-level aggregation. To analyse general trends, we compute these attributions for the $N = 100$ examples with the highest hallucination scores (i.e., most confidently predicted as hallucinations by the probe) from each dataset and

report the mean $\bar{A} = \frac{1}{N}\sum_{n=1}^{N} A_{(n)}$. This focuses the analysis on mechanisms related to strong hallucination signals. Such gradient-based attributions offer first-order estimates of causal influence: rescaling a unit's output by $(1+\delta)$ would be expected to shift the probe logit by approximately $\delta\bar{A}$.

4 Results

4.1 Language models exhibit a robust *linear* representation of contextual hallucinations

To test the linear representation hypothesis—that a single direction in residual-stream activation space separates hallucinated from supported spans—we trained linear probes on this space. Figure 1 presents the F_1 scores of logistic regression probes, trained on each layer of Gemma-2 (2B, 9B, 27B), GPT-2-small, and a 4-layer GELU baseline.[2] Evaluations were performed on CNN/DM news summaries and the CONTRATALES synthetic contradiction dataset.

Probe performance typically rises in early layers, peaks in mid-to-late transformer blocks, and then plateaus. On CNN/DM, Gemma-2-9B achieved an F_1 of 0.98 by layer 17, maintaining > 0.95 through layer 37. GPT-2-small reached 0.78 at layer 11. On CONTRATALES, Gemma-2-9B reached 0.70 in its best layers, and Gemma-2-27B achieved 0.84. The consistent mid-layer performance plateau across models supports a single-direction explanation. This direction generally emerges by layers 8–12, with deeper layers offering marginal improvement in discriminatory power. The optimal layer for detection tends to be deeper in models with more parameters.

The extent to which these probes exploit superficial lexical cues is addressed by comparison to baselines in Section 4.2. The uniqueness of this linear direction is not established here, nor is causality, which is investigated in Section 4.4 through generation steering.

4.2 Residual-stream probes outperform heuristic detectors

Figure 2 shows F_1 scores for the residual-stream linear probe against four baseline detectors – lexical overlap, entity verification, semantic similarity, and Lookback Lens – across three datasets. On news benchmarks, the linear probe achieved 0.97 ± 0.01 F_1 on XSUM and 0.99 ± 0.01 on CNN/DM. This surpassed the strongest baseline, Lookback Lens, by 5–8 points and lexical measures by approximately 15 points.

[2] `gelu-4l` in TransformerLens: https://transformerlensorg.github.io/TransformerLens/generated/model_properties_table.html

On the CONTRATALES dataset, the performance gap increased. Lexical overlap and entity verification F_1 scores were 0.66 and 0.55, respectively. Lookback Lens scored 0.48 ± 0.11. The linear probe achieved 0.75 ± 0.04 F_1, outperforming these alternatives by 9–27 points. This dataset features contradictions that are primarily logical rather than lexically obvious, a characteristic that diminished the effectiveness of the baseline methods which target surface-level cues or specific attention shifts. The linear probe's F_1 score on CONTRATALES, while higher than baselines, did not reach its news-domain performance levels.

4.3 Hallucination representation transfers between news datasets

Cross-domain news transfer Figure 3 shows cross-domain generalisation performance. Detectors were trained on one news corpus (CNN/DM or XSUM) and evaluated on the other without re-tuning. All features were extracted from layer 20 of a Gemma-2-9B observer model. The linear probe exhibited minimal accuracy loss from domain shift. Lookback Lens also generalised to an extent. In contrast, surface cue-based methods showed substantial performance drops. For instance, lexical overlap F_1 decreased from 0.78 (in-domain CNN/DM) to 0.42 when transferred to XSUM. Semantic similarity performance was near chance levels out-of-domain.

MLP attributions identify a consistent sub-circuit Figure 4 presents aggregated MLP attributions, $\bar{A}^{(\ell)}_{\text{mlp}}$, for a linear probe trained on layer 10 of Gemma-2-9B and evaluated on CNN/DM, XSUM, and CONTRATALES.

Per-head attention attributions showed fluctuations around zero, lacking layer-consistent signs or overlap in top-ranked heads across datasets. This indicates that attention routing, in this observer setup, does not offer a stable attribution signal for contextual hallucination.

In contrast, MLP attributions were sparse and layer-consistent. For the layer 10 probe, layers 7 and 8 exhibited positive attribution towards the hallucination direction (with layer 8 being dominant), while layer 9 showed a strong negative attribution. Contributions from other layers were minimal ($|\bar{A}| < 0.02$). This specific {positive (layer 7) $\rightarrow$ strong-positive (layer 8) $\rightarrow$ negative (layer 9)} MLP attribution pattern was consistently observed across all three datasets.

Probe Activations on Unrelated Text The hallucination probe trained on Gemma-2-9B was applied to one million 256-token sequences from the Pile (Gao et al., 2020). Analysis of the sequences producing the highest and lowest probe activations revealed distinct patterns. No consistent pattern was identified in the highest-activating examples. However,

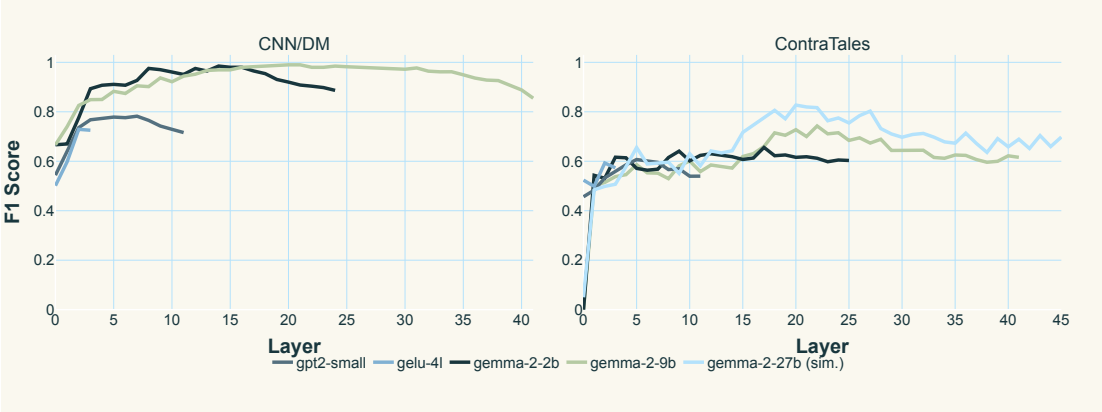

Figure 1: **Layer-wise detection performance of residual-stream linear probes.** Each curve shows the F_1 score (5-fold CV) of a logistic probe trained on a single transformer layer to classify the final sentence of a document as hallucinated or supported by context. **Left**: results on CNN/DM summarisation; **right**: results on the synthetic-contradiction CONTRATALES. The consistent mid-layer plateau across four observer models supports the hypothesis that contextual hallucinations are encoded along a common linear direction in activation space.

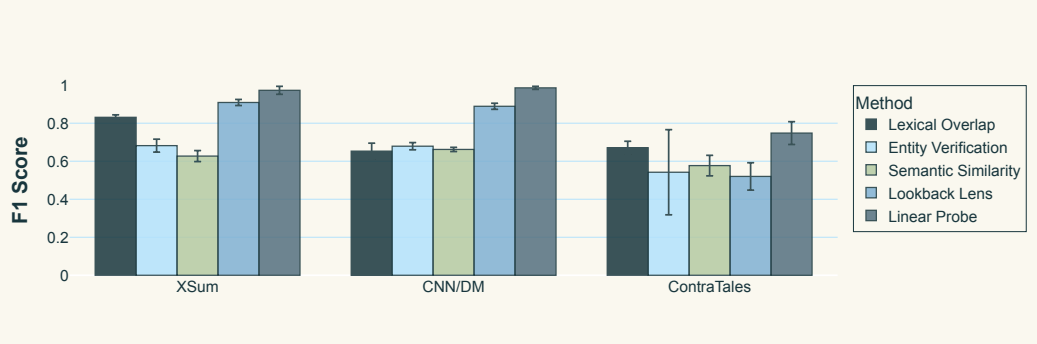

Figure 2: **Comparison of hallucination-detection methods.** Bars give mean F_1 over five cross-validation folds; whiskers show the 95% bootstrap confidence interval. The residual-stream *linear probe* (right-most bar in each group) consistently exceeds all baselines – lexical overlap, entity verification, semantic similarity, and Lookback Lens – across the news datasets (XSUM, CNN/DM) and the logically harder CONTRATALES.

the lowest-activating examples, detailed in Appendix D (Table 4), consistently featured textual repetition. This included exact phrase repetitions (e.g., in gaming or medical texts), quoted content (e.g., from forums or chat logs), and formulaic language found in technical or religious documents. The strongest negative activations (e.g., around -30) were associated with such repeated content.

4.4 Hallucination representation can be used to steer generation

To test the causal effect of the identified residual-stream direction, we patched the normalised probe vector $\mathbf{w}/\|\mathbf{w}\|$ (derived from a linear probe on layer 10 activations of the *same Gemma-2-2B model architecture* used for generation) into its layer 10 during CNN/DAILYMAIL summarization (50 new tokens, greedy decoding). This vector, specifically trained to distinguish hallucinated from faithful content for the Gemma-2-2B, was scaled by $\alpha \in \{-60, \ldots, +60\}$ and injected once at generation start. We generated 128 summaries per α, measuring hallucination rate (judged by GPT-4.1, prompt in App. A) and repetition rate (RapidFuzz, $\geq$ 5-grams, ratio $> 85\%$). The results, plotted in Fig. 5, demonstrate bidirectional control: positive scaling (e.g., $\alpha = +60$) increased hallucination to 0.86 while reducing repetition below 0.05, whereas negative scaling (e.g., $\alpha = -60$) increased repetition to 0.84 with a hallucination rate of 0.35. The unpatched model ($\alpha = 0$) served as a baseline for this trade-off.

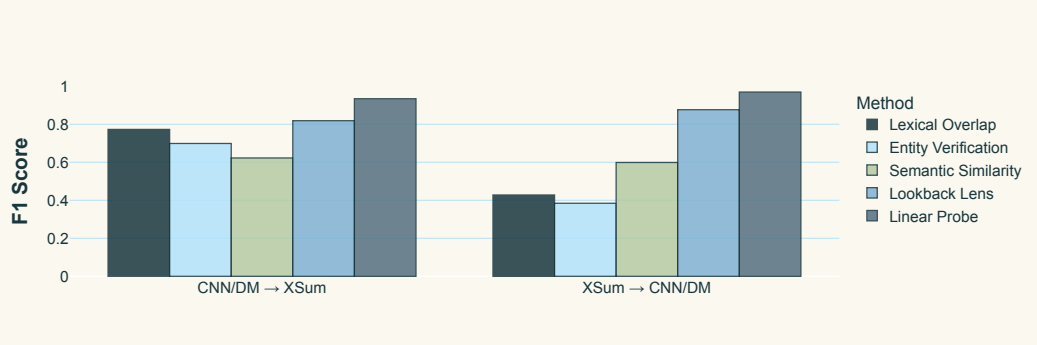

Figure 3: **Cross-domain transfer performance of hallucination detection methods.** F_1 scores for detectors trained on one news dataset (CNN/DM or XSUM) and evaluated on the other. Features were extracted from layer 20 of a Gemma-2-9B observer. The linear probe demonstrates high transferability compared to baseline methods.

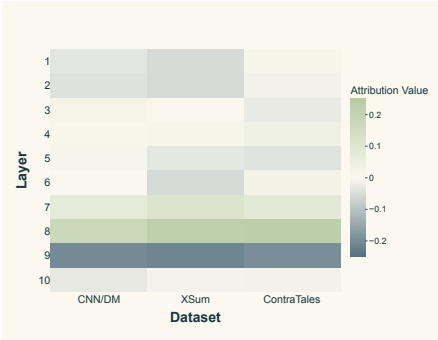

Figure 4: **Aggregated MLP layer attributions for the hallucination probe.** Mean MLP attributions ($\bar{A}^{(\ell)}_{\text{mlp}}$) per layer for a linear probe trained on layer 10 of Gemma-2-9B. Attributions are presented for evaluations on CNN/DM, XSUM, and CONTRATALES, revealing a consistent pattern across datasets in layers 7-9.

4.5 Finetuning improves internal hallucination indication

To investigate improving observer performance on a target domain without new hallucination labels, we performed unsupervised domain adaptation. Observer models (GPT-2-small, Gemma-2-2B, Gemma-2-9B) were further trained for two epochs on 1000 correct-only completions from the CONTRATALES corpus using standard SFT hyperparameters (AdamW, LR 1×10^{-5}, context length 512, batch size 8, 8xH200 GPUs, no dropout). Following this adaptation, the logistic residual-stream probe (as per §3) was retrained on the original labeled data and evaluated on an unseen test fold. As shown in Figure 6, this process improved F_1 scores for all models: by +0.10 for GPT-2-small, +0.17 for Gemma-2-2B, and +0.14 for Gemma-2-9B. For Gemma-2-9B on CONTRATALES, the F_1 score increased from 0.75 to 0.89.

5 Discussion

This work demonstrates a practical and actionable application of interpretability insights through a generator-agnostic observer paradigm: a linear probe on a transformer's residual-stream activations identifies contextual hallucinations in a single forward pass, achieving high F_1 scores. These results underscore the feasibility of leveraging internal model representations for addressing key AI challenges like hallucinations. While the F_1 scores on news benchmarks could be partially influenced by characteristics inherent in prompted synthetic hallucinations, the strong performance on CONTRATALES (a dataset designed to test unambiguous logical contradictions) mitigates this concern by showcasing the probe's ability to identify more fundamental contextual violations, arguably less susceptible to specific generation artifacts.

Our findings offer support for the linear representation hypothesis (LRH) in contextual understanding, showing how interpretability can move beyond correlation to causal intervention. The identified linear direction for hallucination's consistency across Gemma-2 model sizes (2B→9B) and transferability across news domains suggest a fundamental encoding. Crucially, its functional role is substantiated by causal interventions: a single, layer-local injection or ablation of the probe vector along this axis smoothly and monotonically modulates a generator's hallucination and repetition rates, demonstrating it as an actionable, low-dimensional, and causally effective axis for contextual-hallucination awareness. Mechanistically, gradient-times-activation attribution analyses refine this picture, pinpointing the signal for contextual inconsistency to a sparse, layer-consistent pattern of late-layer MLP activity (e.g., layers 7 and 8 positively, layer 9 negatively, for a probe on layer 10 of Gemma-2-9B), rather than diffuse attention patterns. This indicates the observer's awareness is canalised through a specific chain of late-layer feed-forward computations.

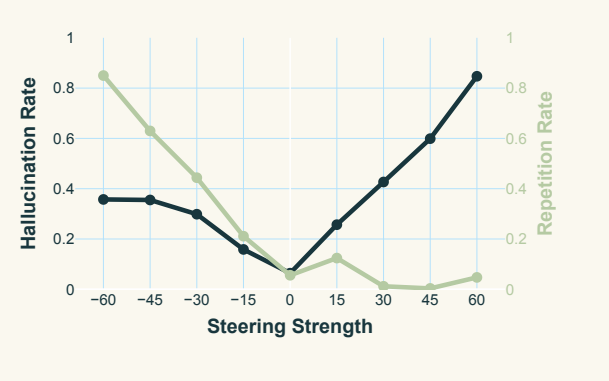

Figure 5: We use steering to generate outputs for CNNDM with the Gemma-2-2B model. We then use fuzzy string matching to determine the repetition rate, and gpt-4.1 to determine the hallucination rate.

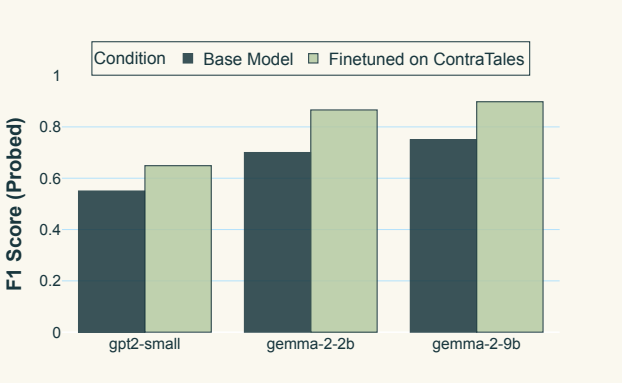

Figure 6: Unsupervised domain adaptation boosts probe accuracy. Bars report the F_1 of the residual-stream logistic probe before (base) and after (FT) two-epoch SFT on 1000 *correct* CONTRATALES continuations.

Furthermore, the practical utility of this internal representation is practically enhanced by unsupervised domain adaptation. Finetuning the observer model on in-domain, correct-only text from CONTRATALES (without new hallucination labels) significantly improved the logistic probe's F_1 score (e.g., for Gemma-2-9B, from 0.75 to 0.89). This implies that additional in-domain language modelling sharpens the internal distinction between logically supported and unsupported statements, a distinction the linear probe can then more effectively exploit, offering an inexpensive, label-free path to enhanced single-pass detection capabilities: important for real-world deployment.

Limitations A primary limitation of this work is the reliance on synthetically generated hallucinations for training and evaluating the majority of our detectors, particularly on the news and medical datasets. While necessary for creating labeled data at scale, continuations prompted from large language models may exhibit patterns or artifacts predictable to an observer model trained on similar data, which might not be representative of naturally occurring "in-the-wild" hallucinations generated by various models under different conditions. This could potentially lead to an overestimation of the detector's performance and generalisability beyond the specific generation methods used here. Although the CONTRATALES dataset offers a valuable benchmark for pure logical contradictions, it also represents a specific type of structured inconsistency. Further validation on datasets containing a broader spectrum of organically generated, human-verified hallucinations would strengthen claims regarding real-world applicability.

Beyond the nature of the training data, the evaluation of the steering experiments relies on a `gpt-4.1` judge to determine hallucination rates. While this is a common method, LLM-based evaluations can be subject to noise, bias, and variability, which may affect the precision of the quantitative steering results. Additionally, while the linear probe detector itself is lightweight and efficient, the observer paradigm necessitates deploying and running a sufficiently capable base transformer model, which still carries significant computational costs compared to purely surface-level detection heuristics. Finally, our study focuses specifically on **intrinsic** hallucinations that contradict or are unsupported by the provided source context; the applicability of the discovered hallucination direction and detection method to **extrinsic** hallucinations, which introduce novel but unverifiable information from outside sources, remains untested.

This research provides compelling evidence for a single, transferable, and causally effective linear direction within transformer activations that corresponds to contextual hallucination. This direction is primarily processed by a sparse and consistent MLP sub-circuit and can be leveraged for both lightweight detection and controlled generation, advancing interpretability by providing concrete methods for building more reliable AI systems. To facilitate further research, we release the CONTRATALES benchmark.

Impact Statement

This paper advances mechanistic interpretability by improving AI reliability and safety through evidence of generator-agnostic methods for detecting and controlling contextual hallucinations via mechanistic insights and causal steering. This offers potential for positive societal impact, enabling the deployment of more trustworthy AI in domains where accuracy is essential. However, this work also carries potential risks; the steering technique, while intended for mitigation, could be misused for generating misinformation, and the detector's performance may be influenced by biases present in its training data.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Azaria, A. and Mitchell, T. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, Dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., and Lim, S. K. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. URL https://distill.pub/2020/circuits.

Chuang, Y.-S., Qiu, L., Hsieh, C.-Y., Krishna, R., Kim, Y., and Glass, J. R. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps, 2024. URL https://arxiv.org/abs/2407.07071.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Dodds, Z. H., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. URL https://arxiv.org/abs/2209.10652.

Engels, J., Liao, I., Michaud, E. J., Gurnee, W., and Tegmark, M. Not all language model features are linear. *arXiv e-prints*, pp. arXiv–2405, 2024.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Ferrando, J., Obeso, O., Rajamanoharan, S., and Nanda, N. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models. *arXiv preprint arXiv:2411.14257*, 2024.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al. spacy: Industrial-strength natural language processing in python. 2020.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 1(1):1–58, 2024. doi: 10.1145/3703155.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., and Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Makhzani, A. and Frey, B. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Manakul, P., Liusie, A., and Gales, M. J. F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, Dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Minder, J., Du, K., Stoehr, N., Monea, G., Wendler, C., West, R., and Cotterell, R. Controllable Context Sensitivity and the Knob Behind It. *arXiv preprint arXiv:2411.07404*, 2024.

Mur, M., Bandettini, P. A., and Kriegeskorte, N. Revealing representational content with pattern-information fmri—an introductory guide. *Social cognitive and affective neuroscience*, 4(1):101–109, 2009.

Nan, F., Nallapati, R., Wang, Z., Santos, C. N. d., Zhu, H., Zhang, D., McKeown, K., and Xiang, B. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*, 2021.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Rai, D., Zhou, Y., Feng, S., Saparov, A., and Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. URL https://arxiv.org/abs/2407.02646.

Research, A. Hallushield: A mechanistic approach to hallucination–resistant models. https://apartresearch.com/project/hallushield-a-mechanistic-approach-to-hallucination-resistant-models, 2025. White paper.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Simhi, A., Herzig, J., Szpektor, I., and Belinkov, Y. Constructing benchmarks and interventions for combating hallucinations in llms. *arXiv preprint arXiv:2404.09971*, 2024.

Simhi, A., Itzhak, I., Barez, F., Stanovsky, G., and Belinkov, Y. Trust Me, I'm Wrong: High-Certainty Hallucinations in LLMs. *arXiv preprint arXiv:2502.12964*, 2025. URL https://arxiv.org/abs/2502.12964.

Slobodkin, A., Goldman, O., Caciularu, A., Dagan, I., and Ravfogel, S. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. *arXiv preprint arXiv:2310.11877*, 2023.

Sun, Z., Zang, X., Zheng, K., Song, Y., Xu, J., Zhang, X., Yu, W., and Li, H. Redeep: Detecting hallucination in retrieval–augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024. URL https://arxiv.org/abs/2410.11414.

Valentin, S., Fu, J., Detommaso, G., Xu, S., Zappella, G., and Wang, B. Cost-Effective Hallucination Detection for LLMs, 2024. URL https://arxiv.org/abs/2407.21424.

Yu, L., Cao, M., Cheung, J. C. K., and Dong, Y. Mechanistic understanding and mitigation of language model non–factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7943–7956, Miami, USA, 2024. doi: 10.18653/v1/2024.findings-emnlp.466. URL https://aclanthology.org/2024.findings-emnlp.466.

Yuksekgonul, M., Chandrasekaran, V., Jones, E., Gunasekar, S., Naik, R., Palangi, H., Kamar, E., and Nushi, B. Attention satisfies: A constraint-satisfaction lens on factual errors of language models, 2024. URL https://arxiv.org/abs/2309.15098.

A Generation Prompts

This section details the prompts used to generate hallucinated and factual continuations for our datasets. We used the OpenAI API with GPT-4.1-mini to create both hallucinated and factual continuations by providing specific instructions through system and user messages.

A.1 Hallucination Generation Prompts

For generating hallucinated continuations (introducing unsupported or contradictory information), we used the following system and user prompts:

```
# System Message
You are an expert summarization assistant helping to create a hallucination dataset.
Your task is to produce exactly ONE sentence that *appears* relevant to the given article
    but introduces at least one factual detail that cannot be inferred from the article.
The sentence must be grammatically correct, and <= 40 words.

# User Message
Original article:
<article>
{document}
</article>

Original correct summary:
<summary>
{original_summary}
</summary>

We are going to replace this sentence: <sentence_to_replace>{replaced_sentence}</
    sentence_to_replace> with a hallucinated sentence that you generate.

The new sentence must contain a made up factual detail that is not present in the original
     article.

Return JUST the new sentence, without quotation marks, xml tags, or any additional
    commentary.
```

A.2 Factual Generation Prompts

For generating factual continuations (containing only information supported by the source), we used the following system and user prompts:

```
# System Message
You are an expert summarization assistant helping to create a dataset.
You will be given an existing summary and asked to replace a single sentence with one
    written by you.
The sentence you write must be factually accurate, grammatically correct, and only contain
     information directly from the original article.
Your sentence should be concise and <= 40 words.

# User Message
Original article:
<article>
{document}
</article>

Original correct summary:
<summary>
{original_summary}
</summary>

We are going to replace this sentence: <sentence_to_replace>{replaced_sentence}</
    sentence_to_replace> with a new sentence that you generate.
```

```
The new sentence must be completely factual, only containing information directly from the
    original article.

Return JUST the new sentence, without quotation marks, xml tags, or any additional
    commentary.
```

A.3 Hallucination Evaluation Prompt

For evaluating hallucination rates during the steering experiments in Section 4.4, we used GPT-4.1 with the following prompt:

```
You are an expert at detecting hallucinations in summaries.

I will provide you with an article and a summary. Your task is to determine if the summary
    contains any hallucinations - information that is not supported by or contradicts the
    article.

A hallucination in a summary is when:
1. It contains information NOT present in the article
2. It contradicts information in the article
3. It makes claims that require outside knowledge not in the article

Answer with ONLY "hallucination" or "no hallucination".

Article:
{article}

Summary:
{summary}
```

These prompts were implemented in our data generation pipeline to create paired examples of source contexts and either factual or hallucinated continuations across all datasets described in Section 3.1.

B CONTRATALES Dataset Generation

The CONTRATALES dataset, comprising 2000 examples of stories with purely logical contradictions, was generated to test the contextual understanding of hallucination detectors. The generation process involved using a large language model, specifically `o4-mini`, guided by a detailed prompt that included instructional guidelines and few-shot examples.

B.1 Generation Process

Each example in CONTRATALES consists of three main parts:

1. **Story Prefix**: An initial narrative segment (typically 7-10 sentences). The first sentence of this prefix establishes an unambiguous constraint or fact about a character or situation (e.g., "Jack had been bald for 10 years," "Sarah was allergic to peanuts"). The subsequent sentences in the prefix develop a neutral, everyday scene, intentionally avoiding any direct reference to, or negation of, the established constraint.
2. **Correct Concluding Sentence**: A single sentence that provides a logical and coherent continuation of the story prefix, respecting the initial constraint.
3. **Contradictory Concluding Sentence**: A single sentence that is structurally and tonally similar to the correct concluding sentence but introduces a detail that subtly and logically contradicts the constraint established in the first sentence of the story prefix.

The generation was performed using the `o4-mini` model. The model was prompted with a set of instructions that specified the desired structure and characteristics of the CONTRATALES examples. To guide the model effectively, the prompt also included a random selection of 5 few-shot examples from a seed set of 24 pre-existing CONTRATALES. These examples

demonstrated the desired format and subtlety. The aim was to create contradictions that are primarily logical rather than lexically obvious, thereby challenging detection methods reliant on surface-level cues.

B.2 Example CONTRATALES

Table 2 showcases four examples from the CONTRATALES dataset, illustrating the structure and nature of the logical contradictions.

Table 2: Illustrative examples from the CONTRATALES dataset.

Story Prefix (Edited for brevity)	**Correct Concluding Sentence**	**Contradictory Concluding Sentence**
Jack had been bald for 10 years. Each morning, he stepped onto his small porch to water the potted herbs... By mid-morning, Jack was usually cycling through the neighborhood... Later, he settled on a bench in the park... In the afternoons, he volunteered at a literacy program... Tonight, he planned to check out the grand opening of an artisan soap store nearby.	Jack was going to the barber shop to get scalp oil.	Jack was going to the barber shop to get a haircut.
Sarah was allergic to peanuts. On Saturday morning, she and her friend packed their bags for a scenic hike... Sarah filled her water bottle and double-checked her map... At a shaded clearing, they paused... Sarah pulled out a granola bar from her pack... Just before dusk, they reached a picnic table overlooking the valley below.	Sarah refused the PB&J sandwich her friend offered.	Sarah ate the PB&J sandwich her friend offered.
Daniel can't swim. He spent Saturday mornings at the community center's lounge reading magazines... Daniel often joined the adult painting sessions instead of the pool activities... Afterwards, he sketched scenes from nature in his notebook. He left the lounge and walked toward the pool deck. There, he paused at the edge of the water, watching the ripples in the sunlight.	Daniel watched the swimmers from a bench by the pool.	Daniel swam laps to warm up before the race.
Laura didn't own a smartphone. Each day she followed her paper planner to keep track of appointments... At the office, she used the wall calendar to schedule meetings... During breaks, she called home from the lobby phone... For lunch, she read a paperback novel... Before dinner, she walked to the hotel and paused under its marquee.	Laura checked the directory in the lobby to find the restaurant's address.	Laura sent a text to reserve a table.

C Manual Features

C.1 Feature extraction

Let $X = (x_1, \ldots, x_{|X|})$ be the context and $Y = (y_1, \ldots, y_{|Y|})$ the model-generated continuation. We concatenate them and perform one forward pass through the 40-layer Gemma-2-27B model f_θ, storing the logits $L \in \mathbb{R}^{T \times V}$, the attention patterns $A^{(\ell)} \in [0,1]^{T \times T}$ for layers $\ell \in \{40, 42, 44\}$ (softmaxed over keys) and the mid-layer residual stream $R^{(m)} \in \mathbb{R}^{T \times d_{model}}$ at layer $m = 28$. The continuation is segmented into non-overlapping chunks $\mathcal{C}_Y = \{c_1^Y, \ldots, c_n^Y\}$ by splitting on full stops and newlines; the context is chunked analogously into $\mathcal{C}_X = \{c_1^X, \ldots, c_k^X\}$.

For each note chunk c_i^Y we derive a dense feature vector $\phi(c_i^Y)$ composed of nine orthogonal signal families described below; all computations reuse tensors already cached, thereby avoiding additional forward passes or temperature sampling.

Token-level uncertainty Given the gold tokens $y_t \in \mathbb{N}^V$ inside c_i^Y we compute the negative log-likelihood $\mathrm{NLL}\, t = -\log p_\theta(y_t)$ and the token entropy $H_t = -\sum_v p_\theta(v) \log p_\theta(v)$. We record the mean and maximum NLL, the proportion of tokens with ground-truth rank > 10 and the slope of a least-squares fit of H_t versus token position.

Cross-context attention Let $\bar{a}^{(\ell)}Y \to X(i,j)$ be the mean of $A^{(\ell)}$ over all query positions in chunk c_i^Y and key positions in c_j^X; similarly $\bar{a}^{(\ell)}Y \to Y(i)$ averages over keys in c_i^Y itself. We form the self-ratio $s^{(\ell)}i = \bar{a}^{(\ell)}Y \to Y(i) \big/ \big(\frac{1}{k}\sum j\bar{a}^{(\ell)}Y \to X(i,j) + 10^{-6}\big)$ and summary statistics of the top-15 values $\{\bar{a}^{(\ell)}Y \to X(i,j)\}_j$.

Residual-stream semantic alignment Using $\tilde{r}_i = \| \sum t \in c_i^Y R_t^{(m)} \|_2^{-1} \sum t \in c_i^Y R_t^{(m)}$ we measure its cosine similarity to (i) the residual average of the transcript chunk with maximal attention weight and (ii) the average residual of the entire transcript.

Embedding-based similarity Chunks are embedded with the OpenAI `text-embedding-3-small` model, yielding unit-norm vectors $\{e_i^Y\}$ and $\{e_j^X\}$. We store the maximum, mean top-3 and variance of $\langle e_i^Y, e_j^X \rangle$ across j, together with the gap between the two highest scores and global max/mean similarities.

Logit eigenspectrum For the slice $L_{c_i^Y} \in \mathbb{R}^{|c_i^Y| \times V}$ we take the top-32 singular values $\sigma_1 \geq \cdots \geq \sigma_{32}$ and include $\sum_r \log \sigma_r$ and the spectral gap $\sigma_1 - \sigma_2$.

Entity grounding Clinical entities are extracted using SciSpaCy. We compute the coverage ratio $|E_i \cap E_X|/|E_i|$ and the count of novel entities $|E_i \setminus E_X|$, where E_i and E_X are the entity sets of chunk i and the transcript respectively.

Surface-level heuristics Features include trigram novelty – $|\,\mathrm{Tri}(c_i^Y) \setminus \mathrm{Tri}(X)|/|\,\mathrm{Tri}(c_i^Y)|$ – numeric-token ratio and the z-score of mean sentence length.

Intra-chunk semantic-graph variance A $k{=}3$ cosine k-NN graph is built on sentence-level embeddings within the chunk; the feature is the variance of edge similarities, capturing semantic isolation.

Table 3: Chunk-level feature families used for hallucination classification. All quantities are derived from a single forward pass or from cached embeddings.

Family	*Description*	*Dim.*
Token uncertainty	NLL mean/max, rank, >10 fraction, entropy slope	4
Attention asymmetry	Self-ratio and top-k stats for three layers	3×3
Residual alignment	Cosine sim. to top-attn and whole-transcript residuals	2
Embedding similarity	Top-k statistics and global max/mean	6
Logit eigenspectrum	Log-sum singular values, spectral gap	2
Entity grounding	Coverage ratio, novel-entity count	2
Surface heuristics	Trigram novelty, numeric ratio, sentence-length z	3
Semantic-graph variance	Variance of k-NN intra-chunk sims	1
Total		**35**

Each note chunk is thus represented by a 35-dimensional feature vector $\phi(c_i^Y) \in \mathbb{R}^{35}$. We stream the vectors to a JSON-Lines file during extraction; the process is resumable, allowing the dataset to be generated incrementally on a single H200 GPU.

Feature Importance Analysis To understand which signals are most indicative of hallucinations, we calculated feature importances using mean decrease in impurity (MDI) from a Random Forest classifier trained on the 35-dimensional feature vectors. Figure 7 visualises the top 15 features for detecting hallucinations in the news summarisation datasets (XSUM and CNN/DM, averaged) versus the CONTRATALES logical contradiction dataset.

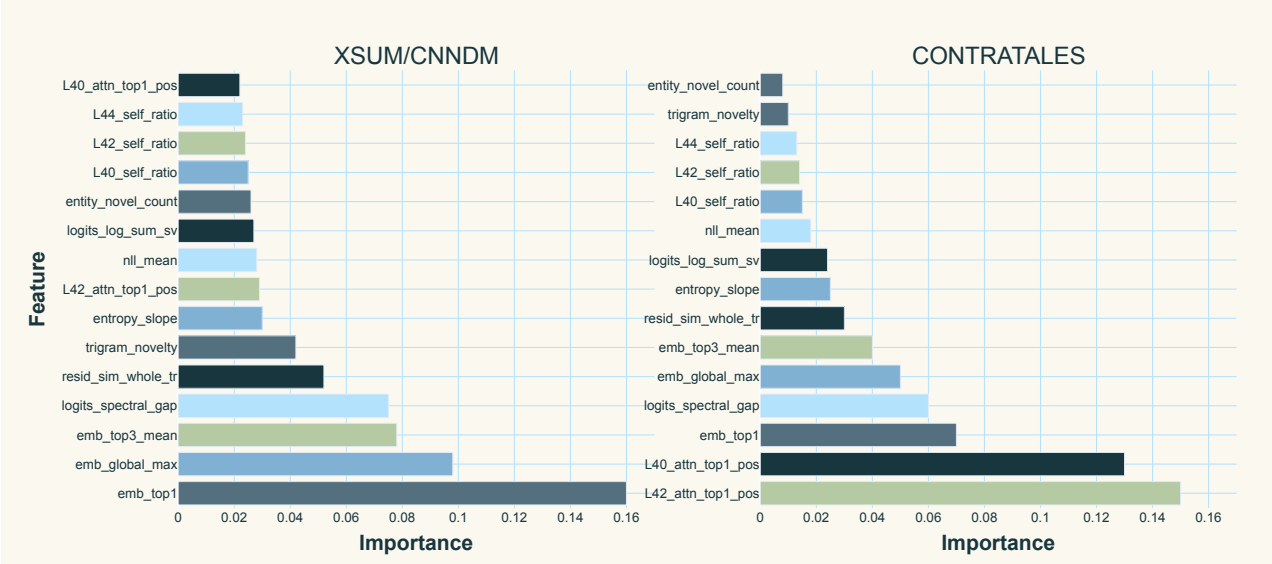

Figure 7: Comparison of top feature importances for hallucination detection across datasets. **Left**: Feature importances for detecting factual hallucinations in news summarisation datasets (XSUM/CNN/DM average). Features are ranked by their importance in this scenario. **Right**: Importances of the *same set of features* when applied to detecting logical contradictions in the CONTRATALES dataset. Note the significant re-ranking and change in relative importance values, highlighting how different feature types (e.g., attention-based vs. novelty-based) contribute variably to identifying distinct forms of contextual inconsistency.

Across both news and logical contradiction datasets, embedding-based similarity features, particularly the maximum similarity to any source chunk (`emb_top1`) and the global maximum similarity (`emb_global_max`), emerge as highly predictive. This suggests that even subtle hallucinations often exhibit a detectable semantic divergence from the source material when measured by robust embedding models.

However, the relative importance of features shifts notably between the two types of datasets. For news summarisation (Figure 7, left), features like the mean of the top-3 embedding similarities (`emb_top3_mean`) and the spectral gap of logits (`logits_spectral_gap`) also rank highly. These indicate that broader semantic disconnects and unusual logit distributions are characteristic of factual hallucinations in news summaries.

In the CONTRATALES dataset (Figure 7, right), which focuses on logical contradictions, attention-based features gain considerable prominence. Specifically, the mean attention from the current note chunk to the source chunk with the highest attention in layer 42 (`L42_attn_top1_pos`) and a similar feature for layer 40 (`L40_attn_top1_pos`) become top-tier predictors. This highlights that logical contradictions are often signalled by how the model attends to specific, relevant parts of the source text when generating the contradictory statement. While embedding similarities remain important, their dominance is reduced compared to the news datasets.

Conversely, features like trigram novelty (`trigram_novelty`) and the count of novel entities (`entity_novel_count`) show a marked decrease in importance for CONTRATALES. This is intuitive: a logical contradiction often reuses existing entities and trigrams from the source to construct the conflicting statement, making novelty a less reliable indicator than for more overt factual fabrications common in news hallucinations. The logit spectral gap also remains a strong feature for contradictions.

D Probe Minimal Activating Examples

In Table 4, we show the minimal activating examples from our hallucination probe on the Pile (Gao et al., 2020). Most contain repetitions, either exact repetitions or semantic ones.

Table 4: Lowest activating examples on the Pile from hallucination detection probe, trained on Gemma-2-9B. We used 1 million examples of sequence length 256 tokens from the Pile.

Ex #	First Instance	Repeated Instance	Activation
1	"Does anyone have suggestions for healy spells at low level? Apart from respeccing resto for levelling that is :)"	"Does anyone have suggestions for healy spells at low level? Apart from respeccing resto for levelling that is :)"	−30.6562
2	"For a year = continued use for a year starting from the initial prescription with a gap no greater than 30 days."	"For a year = continued use for a year starting from the initial prescription with a gap no greater than 30 days."	−29.8125
3	"Is it actually okay to treat nested std::arrays as a single flat C-style array by using .data()-¿data()?"	"Is it actually okay to treat nested std::arrays as a single flat C-style array by using .data()-¿data()?"	−28.4062
4	"Is it possible to fill the Combobox like this programmatically?"	"Is it possible to fill the Combobox like this programmatically?"	−26.8281
5	"Or, people in severe need of pain relief has the same need over a long time?"	"Or, people in severe need of pain relief has the same need over a long time?"	−26.8438
6	"A pseudonym of Jehovah's end-time servant, who personifies the light that dawns on Jehovah's people at the time Jehovah restores them..."	"A pseudonym of Jehovah's end-time servant, who personifies the light that dawns on Jehovah's people at the time Jehovah restores them..."	−26.8125
7	"[8:28 PM] cloppyhooves: They giggle, and you waifu tells you 'Oh, you won't *release* the answer? *Come* on, tell"	"[8:28 PM] cloppyhooves: They giggle, and you waifu tells you 'Oh, you won't *release* the answer? *Come* on, tell"	−26.5938
8	"I just see how we are too far to either the left side or the right side. If we do not get back to the middle, then we will end up like Greece. The USA does not have a 'Germany' to bail them out..."	"I just see how we are too far to either the left side or the right side. If we do not get back to the middle, then we will end up like Greece. The USA does not have a 'Germany' to bail them out..."	−26.1875
9	"Cardiff, Pembrokeshire & South Wales" tourism information with repeated formatting and structure	"Cardiff, Pembrokeshire & South Wales" tourism information with repeated sections	−26.1562
10	"As for Romulan Ale I've given it some thought and I think I've come up with an idea. Ok, they call it 'ale' but it's blue."	Similar brewing discussion repeating terms about Romulan Ale recipe	−25.7969
11	"It turns out Carney was being polite when he said the caution by Canadian CEOs might be excessive. It turns out that they are in fact scaredy cats. Chickens. Nervous Nellies. Cowards, even."	"It turns out Carney was being polite when he said the caution by Canadian CEOs might be excessive. It turns out that they are in fact scaredy cats. Chickens. Nervous Nellies. Cowards, even."	−25.7344
12	"It amazes me to see stuff on Amazon where the album on MP3 costs £7.99 and the cd can be bought 'new and used from' £2 or something ridiculous"	"It amazes me to see stuff on Amazon where the album on MP3 costs £7.99 and the cd can be bought 'new and used from' £2 or something ridiculous"	−25.6094
13	"A pseudonym of Jehovah's end-time servant, whom Jehovah appoints to confront his people with their hypocrisy..."	"A pseudonym of Jehovah's end-time servant, whom Jehovah appoints to confront his people with their hypocrisy..."	−25.3906
14	"The Bank of England has set up a research division looking at how it can get involved with digital currencies..."	Similar content about Bank of England and digital currencies repeated	−25.3750
15	"What's needed is not just yet another O/RM tool (which are tuppence a dozen anyhow - I personally have written three) but a tool which supports database programming using only the conceptual model..."	"What's needed is not just yet another O/RM tool (which are tuppence a dozen anyhow - I personally have written three) but a tool which supports database programming using only the conceptual model..."	−25.2656