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,

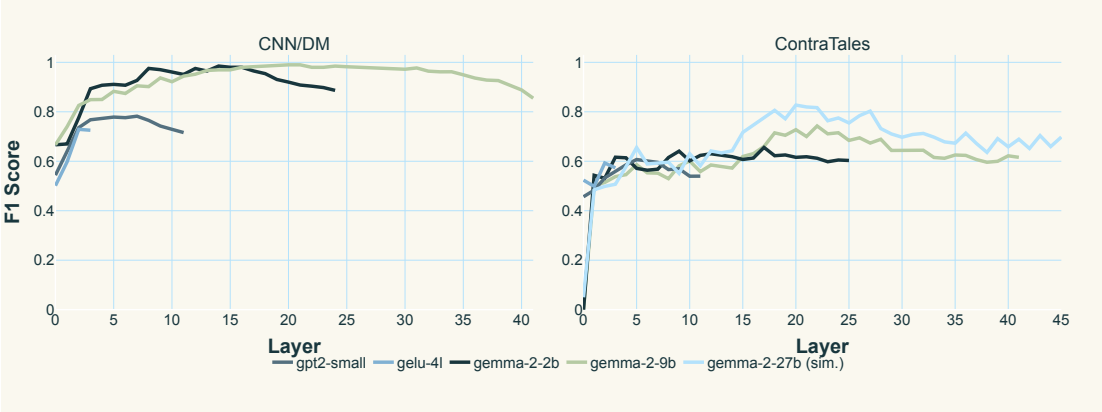

Figure 1: **Layer-wise detection performance of residual-stream linear probes.** Each curve shows the F_1 score (5-fold CV) of a logistic probe trained on a single transformer layer to classify the final sentence of a document as hallucinated or supported by context. **Left**: results on CNN/DM summarisation; **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.

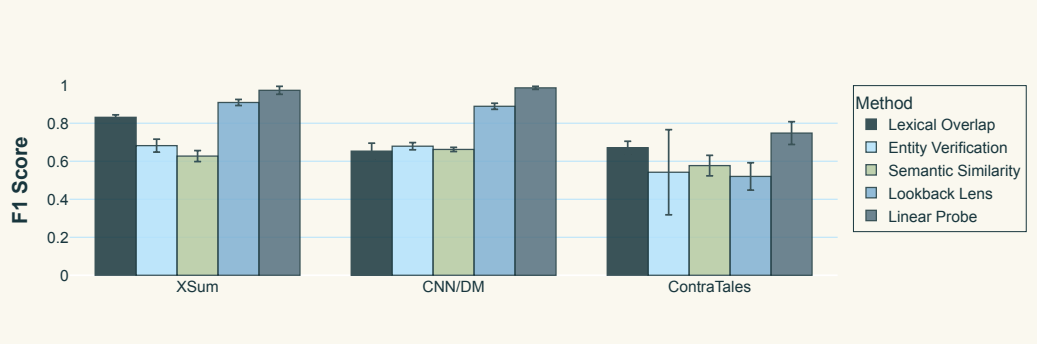

Figure 2: **Comparison of hallucination-detection methods.** Bars give mean F_1 over five cross-validation folds; whiskers show the 95% bootstrap confidence interval. The residual-stream *linear probe* (right-most bar in each group) consistently exceeds all baselines – lexical overlap, entity verification, 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.