

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  $X = (x_1, \dots, x_{|X|})$  be the context and  $Y = (y_1, \dots, y_{|Y|})$  the model-generated continuation. We concatenate them and perform one forward pass through the 40-layer Gemma-2-27B model  $f_\theta$ , 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  $\mathcal{C}_Y = \{c_1^Y, \dots, c_n^Y\}$  by splitting on full stops and newlines; the context is chunked analogously into  $\mathcal{C}_X = \{c_1^X, \dots, c_k^X\}$ .

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

**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  $\sigma_1 \geq \dots \geq \sigma_{32}$  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 –  $|\text{Tri}(c_i^Y) \setminus \text{Tri}(X)| / |\text{Tri}(c_i^Y)|$  – 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	$3 \times 3$
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
<b>Total</b>		<b>35</b>

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.

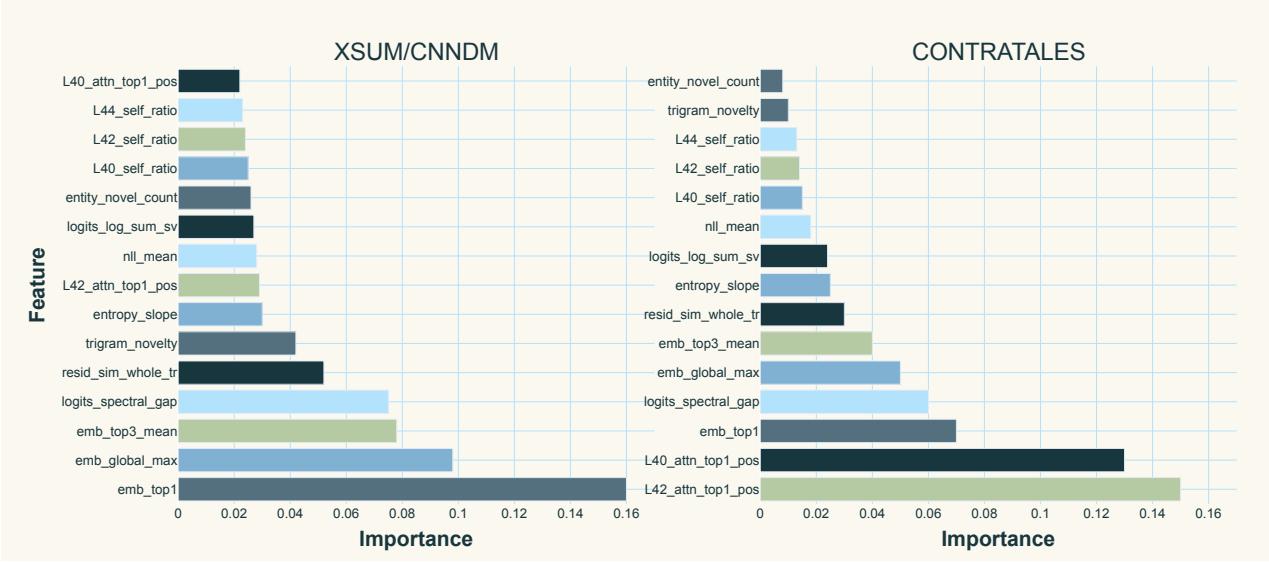


Figure 7: Comparison of top feature importances for hallucination detection across datasets. **Left:** Feature importances for detecting factual hallucinations in news summarisation datasets (XSUM/CNN/DM average). Features are ranked by their importance in this scenario. **Right:** Importances of the *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.