

Hidden Signals: Hallucination Prediction in Medical QA

Stanford CS224N Custom Project

Christina Ba

Department of Computer Science
Stanford University
cba1@stanford.edu

Catherine Zhang

Department of Computer Science
Stanford University
czhang7@stanford.edu

Ina Chun

Department of Computer Science
Stanford University
inachun@stanford.edu

Abstract

Use of large language models in medical contexts have been steadily increasing, yet their tendency to hallucinate (generating claims unsupported by the provided evidence) poses serious risks in clinical contexts. We study hallucination detection in a retrieval-augmented generation (RAG) setting using PubMedQA, framing it as a supervised binary classification problem over signals extracted during a single forward pass of Qwen2.5-3B-Instruct. Rather than relying on expensive multi-sample inference or a separate model at deployment time, we extract token-level uncertainty signals, per-layer attention features, and mid-layer hidden states in a single greedy decoding pass, using an LLM judge (Gemini 2.5 Flash) to supply faithfulness labels. We train lightweight logistic regression classifiers on these features and systematically compare them against our baseline model based on Lookback Lens [1]. Results show that attention to-context ratios prove to be the weakest signal family - even underperforming simple 17-dimensional token probability features, suggesting that in biomedical QA, faithfulness failures arise from misinterpretation of retrieved context rather than inattention to it. Mid-layer hidden states emerge as the dominant signal, and compressing them to 32 principal components via PCA yields the best overall result (AUC-ROC = 0.717), demonstrating that hallucination is encoded compactly and linearly in the model's internal representations. Our results suggest that cheap, single-pass probing of hidden states is a viable and effective approach to hallucination detection even in high-stakes medical domains.

1 Key Information to include

- Mentor: Luke Bailey
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Large language models (LLMs) have demonstrated strong capabilities in biomedical question answering, yet they remain prone to hallucination, generating confident claims unsupported by the provided evidence [2]. In safety-critical domains like medicine, this is particularly dangerous, motivating the need for reliable and efficient hallucination detection.

Existing approaches are expensive. Post-hoc LLM judges [3, 4] require an additional inference call from a large proprietary model, while uncertainty-based methods such as semantic entropy [5] demand multiple sampled generations per query. Recent work suggests a cheaper alternative: signals already present within a single generation pass. Kossen et al. [6] show that linear probes on hidden states can approximate semantic entropy, and Lookback Lens [1] demonstrates that per-layer attention-to-context ratios alone are predictive of faithfulness failures in RAG settings.

We apply this insight to the medical domain. Using PubMedQA [7] in a RAG setting with Qwen2.5-3B-Instruct [8], we extract three complementary feature blocks during a single greedy decoding pass: 141 scalar token and attention signals, a 576-dimensional per-(layer, head) context concentration vector following Lookback Lens, and 2048-dimensional mid-layer hidden states - and train a logistic regression classifier labeled by Gemini 2.5 Flash as an LLM judge. We systematically compare these feature families to assess which signals matter most for hallucination detection in medical QA.

Our contributions are: (1) a single-pass feature extraction pipeline for hallucination detection in medical RAG; (2) a systematic comparison of uncertainty, attention, and hidden-state feature families for hallucination classification; (3) an analysis of whether the Lookback Lens finding transfers to a smaller open-weight generator on biomedical text; and (4) an empirical demonstration that mid-layer hidden states, compressed to 32 principal components, constitute the most effective single signal for hallucination detection in this setting.

3 Related Work

Hallucination in LLMs. Ji et al.[2] survey the landscape of hallucination in neural text generation, distinguishing faithfulness failures (outputs inconsistent with source context) from factuality failures (outputs inconsistent with world knowledge). Our work focuses on the former: detecting when a model’s answer is unfaithful to retrieved medical passages. LLM-as-judge frameworks such as RAGAS [3] and FActScore [4] operationalize faithfulness evaluation using a second model, providing a reliable but expensive label source. We use Gemini 2.5 Flash as an offline judge to generate training labels, rather than as a deployment-time detector.

Uncertainty-based detection. Farquhar et al.[5] introduce semantic entropy, which measures uncertainty over clusters of semantically equivalent outputs rather than raw token probabilities, and show it strongly predicts confabulation. However, computing semantic entropy requires multiple sampled generations per query-prohibitive in latency-sensitive or resource-constrained settings. SelfCheckGPT [9] similarly relies on sampling consistency across multiple generations, achieving zero-resource detection but at comparable cost. Our approach instead operates from a single greedy decoding pass given time and compute constraints.

Lightweight probing methods. Kossen et al.[6] show that linear probes trained on model hidden states can predict semantic entropy without sampling, dramatically reducing inference cost. Lookback Lens [1] takes a complementary approach, training logistic regression on per-layer attention-to-context ratios- the fraction of each layer’s attention directed at retrieved passages—and demonstrating strong hallucination detection in RAG settings. Quevedo et al.[10] show that just four scalar features derived from token probabilities (mean, minimum, standard deviation, and top-1 to top-2 gap) suffice for effective detection via logistic regression. Our feature set directly incorporates these four signals alongside attention-based features, testing whether combining both families improves over either alone.

4 Approach

Our approach detects hallucinations by training a lightweight linear classifier on internal signals extracted from a single instrumented forward pass of the generator, requiring no additional model calls at inference time. The pipeline has three stages: (1) instrumented generation, (2) feature extraction, and (3) linear classification.

Instrumented generation. We prompt Qwen2.5-3B-Instruct with each question and its associated context passages and generate a free-form answer using greedy decoding. We enable

output_scores=True, output_attentions=True, and output_hidden_states=True during generation, capturing per-step token logits, full attention tensors, and hidden states at every layer for every generated token—all within a single forward pass.

Feature extraction. From each instrumented generation, we extract three complementary feature blocks.

(i) Lookback Lens (LB, 576-dim)[1]: we compute per-(*layer, head*) context concentration—the fraction of each attention head’s weight falling on the context tokens—averaged across all generated steps, yielding a $36 \times 16 = 576$ -dimensional vector. This captures *where* the model is attending relative to the source passage.

(ii) Scalar features (141-dim): We compute 141 scalar signals grouped into six sub-blocks: token probability statistics (TP; 17-dim), context attention trajectory scalars (CAT; 10-dim), per-layer context concentration averaged over heads (PLC; 36-dim), per-layer attention entropy (PLE; 36-dim), per-layer head disagreement (PLH; 36-dim), and miscellaneous interaction and text features (6-dim). These capture generation confidence and the temporal and layer-wise dynamics of attention.

(iii) Hidden states (HS, 2048-dim): We mean-pool the hidden state at layer 18 (the middle layer of the 36-layer model) across all generated tokens, obtaining a single vector that summarizes the model’s internal semantic representation of its own output.

Classification. We concatenate the desired feature blocks and train an ℓ_1/ℓ_2 -regularized logistic regression classifier with balanced class weights. The three feature blocks are designed to capture complementary signals: LB reflects *attention allocation* to context, the scalar features capture *generation confidence and dynamics*, and HS reflects *what the model is internally representing*—the semantic content being computed. We hypothesize that hallucinated generations will exhibit lower context attention concentration, higher token-level uncertainty, and systematically different hidden-state activations compared to faithful ones. For our best-performing configuration, we additionally apply PCA to compress the 2048-dimensional hidden states to 32 principal components, fit on the training set only, which we find removes noisy dimensions and improves generalization.

5 Experiments

5.1 Data

We use PubMedQA (pqa_labeled), a biomedical question answering dataset of 1,000 expert-annotated examples, each consisting of a research question, a set of PubMed abstract passages as context, and a ground-truth yes/no/maybe label. We split the dataset into 700 training, 150 validation, and 150 test examples using a fixed random seed. Free-form answers are generated by Qwen2.5-3B-Instruct and labeled for context faithfulness by a Gemini 2.5 Flash LLM judge, producing binary hallucination labels. To validate label quality, we manually verified 100 sampled examples and achieved 92% agreement between human judgment and the LLM judge, achieved through iterative prompt engineering to refine the judge’s faithfulness criteria. The resulting dataset is class-imbalanced, with approximately 75% faithful and 25% hallucinated examples across all splits.

5.2 Evaluation method

Our primary metric is AUC-ROC (Area Under the Receiver Operating Characteristic Curve), which measures the probability that the classifier ranks a hallucinated example higher than a faithful one across all possible thresholds. AUC-ROC is threshold-independent and robust to class imbalance, making it well-suited to our setting where a naive majority-class classifier would achieve 75% accuracy while detecting zero hallucinations. We additionally report AUC-PR (Area Under the Precision-Recall Curve), which is more sensitive to performance on the minority (hallucinated) class, as well as F1, Precision, and Recall computed at a tuned threshold.

5.3 Experimental Details

All classifiers are logistic regression models trained on standardized features (zero mean and unit variance using `StandardScaler`) with balanced class weights to address the 3:1 class imbalance.

We perform a joint grid search over regularization type (L_1 or L_2) and regularization strength

$$C \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1.0, 10.0\},$$

selecting the setting that maximizes AUC-ROC on the validation set. We then tune the classification threshold separately on the validation set by sweeping values from 0.01 to 0.99 and choosing the threshold that maximizes F1. All reported metrics are computed only on the held-out test set.

We first establish a **Lookback Lens baseline** by training a classifier on the 576-dimensional per-(*layer, head*) context concentration vector alone, following Chuang et al [1]. We then evaluate each internal feature block *individually* to measure its standalone predictive power: token probability features (TP; 17-dim), per-layer context concentration (PLC; 36-dim), per-layer attention entropy (PLE; 36-dim), per-layer head disagreement (PLH; 36-dim), all per-layer attention features combined (PLA; 108-dim), and mid-layer hidden states (HS; 2048-dim).

Finally, we evaluate five *stacked* configurations that augment the Lookback Lens baseline with additional feature blocks: LB+TP, LB+CAT, LB+PLA, LB+HS, and a full system combining all features (LB + all scalar features + HS; 2765-dim). For PCA experiments, the projection matrix is fit exclusively on the training-set hidden states and then applied to the validation and test sets without refitting. All experiments use `random_state=42` for reproducibility.

5.4 Results

5.4.1 Feature Ablation

Model	Features	Dim	F1	Prec.	Recall	AUC-ROC	AUC-PR
<i>Standalone</i>							
LB	Lookback Lens (per-head ctx conc.)	576	0.316	0.227	0.517	0.604	0.293
TP	Token probability	17	0.360	0.267	0.552	0.637	0.308
PLC	Per-layer ctx concentration	36	0.315	0.215	0.586	0.576	0.224
PLE	Per-layer attention entropy	36	0.308	0.245	0.414	0.596	0.239
PLH	Per-layer head disagreement	36	0.330	0.235	0.552	0.601	0.233
PLA	All per-layer (PLC+PLE+PLH)	108	0.265	0.204	0.379	0.564	0.211
HS	Hidden-state probe (layer 18)	2048	0.364	0.385	0.345	0.703	0.319
<i>Stacked on Lookback Lens</i>							
LB + TP	Lookback Lens + token prob.	593	0.333	0.239	0.552	0.614	0.317
LB + CAT	Lookback Lens + attn trajectory	586	0.340	0.246	0.552	0.609	0.303
LB + PLA	Lookback Lens + all per-layer	684	0.292	0.217	0.448	0.603	0.287
LB + HS	Lookback Lens + hidden states	2624	0.418	0.306	0.655	0.683	0.307
Full	LB + all scalar + HS	2765	0.370	0.270	0.586	0.674	0.305

Table 1: Test set results across all feature configurations. **Bold** denotes the best value per metric column. All classifiers are logistic regression with hyperparameters tuned on the validation set.

As shown in Table 1, HS is the strongest standalone feature block, achieving the highest (bolded) AUC-ROC of 0.703. This is approx. 7 percentage points above the second-best standalone feature block (token probability, AUC-ROC = 0.637) and approx. 10 percentage points above our LB baseline (AUC-ROC = 0.604). For F1, stacking LB + HS achieves the best (with F1 = 0.418). Our data shows us that the remaining scalar features may introduce noise, particularly at the size of our training set, as introducing scalar to create a full system (LB + all scalar + HS) hurts performance (reduction in AUC-ROC from 0.683 to 0.674). Figure 1 clearly demonstrates that the hidden-state probe (HS) achieves the best performance at 0.703, outperforming all combinations.

Figure 2 reinforces this finding beyond a single operating point: the HS curve dominates across all false-positive rates, proving this advantage is not simply an artifact of threshold selection. Notably, the TP curve lies consistently above LB across all thresholds as well, corroborating that token-level uncertainty generalizes more reliably to this domain than attention-to-context patterns.

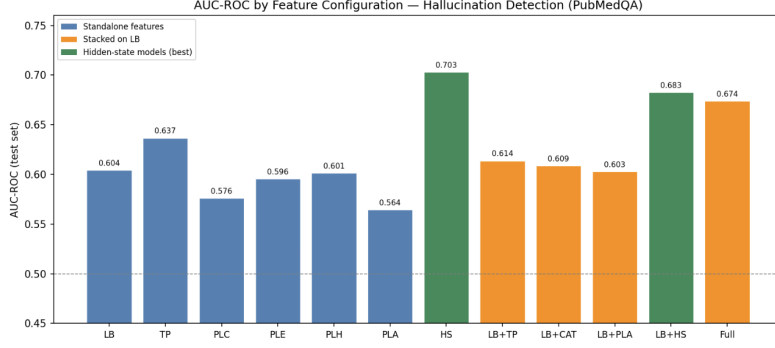


Figure 1: AUC-ROC by feature configuration on the PubMedQA test set. Blue bars show standalone feature sets, orange bars show features stacked on the Lookback Lens baseline (LB), and green bars show hidden-state models.

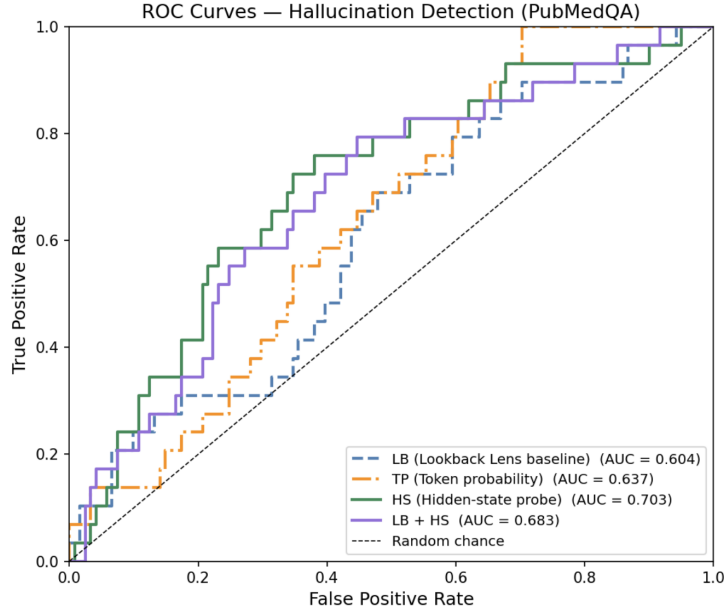


Figure 2: ROC curves for selected feature configurations. The hidden-state probe (HS, AUC=0.703) dominates across all thresholds. Token probability features (TP, AUC=0.637) outperform the Lookback Lens baseline (LB, AUC=0.604), suggesting uncertainty signals transfer more reliably than attention patterns to this domain.

Model	Dim	F1	Prec.	Recall	AUC-ROC	AUC-PR
HS raw	2048	0.364	0.385	0.345	0.703	0.319
HS-PCA-32	32	0.389	0.262	0.759	0.717	0.374
HS-PCA-64	64	0.382	0.283	0.586	0.690	0.339
HS-PCA-128	128	0.370	0.253	0.690	0.661	0.297
HS-PCA-256	256	0.359	0.286	0.483	0.611	0.259
HS-PCA-512	512	0.342	0.227	0.690	0.619	0.274

Table 2: Effect of PCA dimensionality k on hidden-state probe performance. HS raw (2048-dim) is included as a baseline. **Bold** denotes the best value in each column. PCA is fit on the training set only.

Table 2 shows our results upon applying PCA (fit only on training data), since we hypothesized that the hallucination-relevant signal is primarily concentrated in a small subspace of the hidden state. HS-PCA-32 achieves the best result across the entire study ($AUC-ROC = 0.717$, $AUC-PR = 0.374$, $F1 = 0.389$), performing higher than raw HS on all three metrics. As seen in Figure 3, as PCA dimensionality k increases, performance generally degrades, almost monotonically. By $k = 256$, $AUC-ROC$, $AUC-PR$, and $F1$ perform worse than the raw HS baseline ($AUC - ROC = 0.611$ vs. 0.703).

5.4.2 Discussion

Our results produced two unexpected surprises. First, the dominance of hidden states (and the impact of PCA compression to 32-dim on performance) was greater than expected. Based on prior probing work [6], we anticipated that hidden states would be informative, but the compactness of the signal (out of 2048 dimensions) suggests that hallucination is encoded as a tight geometric structure in activation space, not a diffuse property of the full representation. Second, the baseline Lookback Lens underperformed expectations. We expected LB to serve as a competitive baseline given its strong results on context-faithfulness tasks in its original research [1] but in our results, it underperformed even the 17-dimensional token probability features. We attribute this to a misalignment with the original LB measurement purpose and our task’s requirements. On PubMed QA, the model often looks towards the provided abstract and faithfulness errors arise from misinterpretation of the context, not from a lack of attention. Additionally, we retroactively recognized that our LB implementation slightly differs from the original in one respect, as we average attention to input tokens across steps and then compute context concentration ratios, unlike the original paper that computes per-step ratios and then averages. This means the denominator excludes attention mass directed at previously-generated tokens, inflating all concentration values; we note this as a limitation for future experiments.

6 Analysis

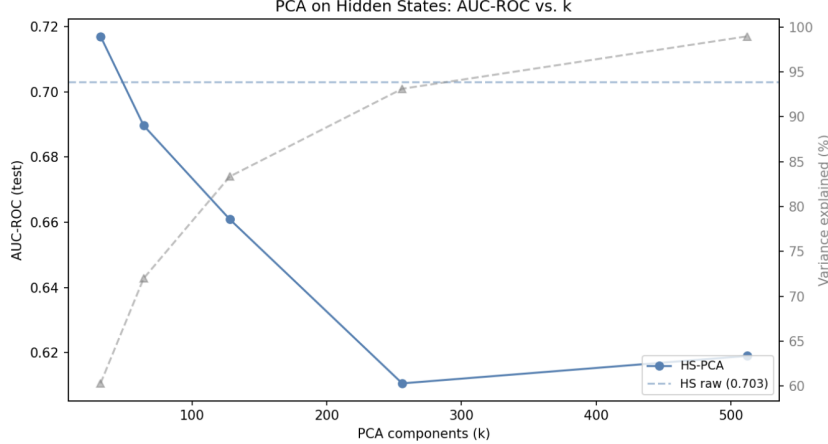


Figure 3: AUC-ROC as a function of PCA components k applied to the 2048-dimensional hidden state. Performance peaks at $k = 32$ ($AUC=0.717$), slightly exceeding the raw hidden-state probe (0.703 , dashed line), suggesting mild regularization benefits from dimensionality reduction.

6.1 Hidden States vs. Attention Features

The gap of approx. 7 percentage points in AUC-ROC between HS (0.703) and all attention-based features ($0.564 \sim 0.604$) identifies a fundamental gap in what these two signal families represent. Attention features reflect where the model attends, calculating the fraction of attention that is allocated to source-context tokens at each layer and head. It is thus a well-reasonable proxy particularly for faithfulness in settings where hallucination occurs from ignoring retrieved context. However, with our PubMedQA dataset, the generator generally attends to the provided abstract and the failure stems from misinterpretation of the context.

On the other hand, hidden states reflect the model’s internal semantic representation. We use exactly the midpoint (layer 18 of our 36-layer model) as middle layers are empirically where high-level semantic content is most linearly decodable. This representation is sensitive to comprehension failures that attention patterns are not able to capture.

6.2 The Compactness of the Hallucination Signal

Our PCA results indicate that hallucination-relevant information in the middle layer HS is concentrated in approximately 32 principal directions out of 2048, and hallucination occupies a low-dimensional subspace of the full representation. These results demonstrate direct implications for classifier conditioning. A logistic regression classifier trained on 32 features from 700 examples (in our case) can be well-conditioned, while the same classifier on 2048 features is highly underdetermined. PCA projects onto the principal subspace and increases the ratio of informative dimensions to training examples by almost two orders of magnitude, and thus accounts for the observed improvement in generalization.

6.3 Precision–Recall Tradeoffs

Table 2 reveals a pronounced tradeoff between precision and recall, with the raw HS probe operating conservatively (precision = 0.385 and recall = 0.345). HS-PCA-32 shifts significantly towards recall (with recall = 0.759 and precision = 0.262), reflecting the outcome of threshold tuning on the validation set. For use cases where safety takes precedent, the high-recall configuration is ideal. For applications where false-positives can risk user trust, the favored probe would be the higher-precision raw HS probe. In medical QA, the cost of a missed hallucination is much greater than that of a false alarm, and thus HS-PCA-32 (recall = 0.759) is the right operating point for our downstream purposes.

6.4 Failure Modes

Our experiments and resulting evaluation are restricted to a single model and relatively-small dataset. More broadly, even our best-performing model (HS-PCA-32, AUC-ROC = 0.717) represents only modest improvement over chance, which we attribute primarily to the limited size of our training set. With only 700 training examples, a linear classifier operating on high-dimensional internal signals has insufficient data to reliably learn the subtle patterns that distinguish faithful from hallucinated generations — a challenge compounded by the complexity of medical question answering, where faithfulness requires nuanced domain understanding rather than surface-level pattern matching. A larger dataset would likely yield substantially stronger results across all feature configurations. The PCA projection directions are also specific to the activation geometry of Qwen2.5-3B-Instruct on PubMedQA, and it remains uncertain whether the identified 32-dim subspace transfers to larger models or different architectures. Furthermore, as noted in our results section, our LB implementation slightly differs from the original paper: we average attention to input tokens across steps before computing context concentration ratios, whereas the original computes per-step ratios and then averages. Our approach effectively shifts the denominator as it excludes attention mass directed at previously-generated tokens. Without time constraints, we would aim to re-implement the Lookback Lens feature extraction to precisely match the original implementation, computing the context concentration ratios per step before averaging.

7 Conclusion

We presented a single-pass pipeline for hallucination detection in medical RAG, extracting token-level uncertainty signals, per-layer attention features, and mid-layer hidden states from Qwen2.5-3B-Instruct on PubMedQA. In clinical settings, where an undetected hallucination in a model-generated answer to a biomedical question can translate to misinformation in a high-stakes decision, a reliable and computationally cheap detection mechanism is essential. Our central finding is that mean-pooled hidden states at layer 18, compressed to 32 principal components via PCA, constitute the most effective signal for hallucination detection in this setting (AUC-ROC = 0.717), outperforming all other feature configurations including the Lookback Lens baseline. The dramatic performance degradation beyond $k=32$ PCA components suggests that hallucination is encoded along a small number of directions in the hidden state, rather than diffusely across the full representation.

Our results indicate that the Lookback Lens finding does not transfer directly to PubMedQA since nuanced clinical language in medical Q&A makes correct reasoning over context harder than simply attending to it. Simple 17-dimensional token probability features outperform the 576-dimensional attention baseline, suggesting uncertainty signals generalize more robustly. Future work should examine whether the identified 32-dimensional subspace transfers across model scales and biomedical domains, and whether these hidden-state directions correspond to interpretable properties of the generated text.

Team contributions (Required for multi-person team)

Catherine set up the data pipeline and led ablation experiments. Christina set up LLM judge and tuned human-judge alignment. Ina led feature extraction of the three feature blocks. We split up instrumented generation and all contributed to experimental design and report writeup.

References

- [1] Yung-Sung Chuang et al. Lookback lens: Detecting faithfulness failures in retrieval-augmented generation via attention to context. *arXiv preprint arXiv:2403.10112*, 2024.
- [2] Zhen Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- [3] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [4] Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [5] Sebastian Farquhar, Jan Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.
- [6] Jan Kossen, Sebastian Farquhar, Lorenz Kuhn, and Yarin Gal. Semantic entropy probes: Cheap hallucination detection from hidden states. *arXiv preprint arXiv:2401.17114*, 2024.
- [7] Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [8] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2501.15383*, 2025.
- [9] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.
- [10] Ernesto Quevedo et al. Detecting hallucinations in large language model generation: A token probability approach. *arXiv preprint arXiv:2405.19648*, 2024.