Do Hallucination Neurons Generalize? Evidence from Cross-Domain Transfer in LLMs

Anonymous ACL submission

Abstract

Recent work identifies a sparse set of "hallucination neurons" (H-neurons), less than 0.1% of feed-forward network neurons, that reliably predict when large language models will hallucinate. These neurons are identified on general-knowledge question answering and shown to generalize to new evaluation instances. We ask a natural follow-up question: do H-neurons generalize across *knowledge domains*? Using a systematic cross-domain transfer protocol across 6 domains (general QA, legal, financial, science, moral reasoning, and code vulnerability) and 5 open-weight models (3B–8B parameters), we find they do not. Classifiers trained on one domain's H-neurons achieve AUROC 0.783 within-domain but only 0.563 when transferred to a different domain ($\Delta = 0.220$, $p < 0.001$), a degradation consistent across all models tested. Our results suggest that hallucination is not a single mechanism with a universal neural signature, but rather involves domain-specific neuron populations that differ depending on the knowledge type being queried. This finding has direct implications for the deployment of neuron-level hallucination detectors, which must be calibrated per domain rather than trained once and applied universally.

1 Introduction

Large language models (LLMs) hallucinate: they produce fluent, confident text that is factually incorrect, logically inconsistent, or unsupported by any training evidence (Ji et al., 2023; Huang et al., 2023; Rawte et al., 2023). As LLMs are deployed in high-stakes domains such as legal analysis (Guha et al., 2024), financial reasoning (Chen et al., 2024), and medical decision support (Pal et al., 2024), detecting and mitigating hallucination has become a central challenge for the field.

A promising line of mechanistic interpretability research has begun to localize hallucination within model internals. Gao et al. (2025) identify "H-neurons," a sparse set of fewer than 0.1% of feed-forward network (FFN) neurons, whose activation patterns reliably predict whether a model will hallucinate. Using the Contribution Estimation Through Token-level Traits (CETT) metric, they train lightweight classifiers on neuron activation features that achieve strong hallucination detection performance. They further demonstrate that these neurons have a causal relationship with hallucination behavior through activation scaling experiments.

This work opens a natural follow-up question that Gao et al. (2025) do not systematically address: **do H-neurons discovered in one knowledge domain generalize to others?** The original study identifies H-neurons on TriviaQA and validates on related QA benchmarks, but never trains and tests across fundamentally different knowledge domains. If H-neurons encode a universal "hallucination signal," then a classifier trained on general QA should also detect hallucination in legal, financial, or scientific contexts. If hallucination involves domain-specific neural pathways, then cross-domain transfer should fail.

This distinction matters for deployment. Current practice implicitly assumes that hallucination detection methods can be developed on convenient benchmarks and applied broadly. If H-neurons are domain-specific, this assumption is false, and practitioners need domain-specific detection infrastructure.

We present, to our knowledge, the first systematic cross-domain transfer analysis of hallucination neurons across 6 knowledge domains and 5 open-weight models (3B–8B parameters). Our key contributions are:

1. **A cross-domain transfer protocol for H-neurons.** We train H-neuron classifiers on each domain and evaluate every classifier on

every domain, yielding 180 train-test domain pairs across 5 models.

2. **Evidence of domain-specific hallucination pathways.** We find a large and statistically significant gap between within-domain detection (AUROC = 0.783) and cross-domain transfer (AUROC = 0.563), with $\Delta = 0.220$ ($p < 0.001$), consistent across all models.

3. **Analysis of domain and model factors.** We characterize which domain pairs exhibit partial transfer and which are completely non-transferable, and examine how model architecture and scale modulate domain specificity.

Our findings reframe hallucination not as a single mechanism with a universal neural signature, but as a family of domain-conditioned processes, with implications for both detection and intervention.

2 Related Work

2.1 Hallucination Detection Methods

Detecting hallucination in LLM outputs has been approached from multiple angles. **Output-based methods** operate on generated text without accessing internal states: SelfCheckGPT (Manakul et al., 2023) detects hallucination by measuring consistency across sampled responses, while Mündler et al. (2024) prompt models to check their own outputs for contradictions. **Uncertainty-based methods** leverage confidence signals: semantic entropy (Kuhn et al., 2023) computes entropy over meaning rather than tokens, and Semantic Entropy Probes (Kossen et al., 2024) train linear probes on hidden states to predict semantic entropy with a single forward pass. **Activation-space methods** analyze internal representations: HaloScope (Du et al., 2024) identifies a "truth subspace" via SVD, and Marks and Tegmark (2024) demonstrate that truth is linearly represented in LLM activations, though their transfer experiments remain within similar knowledge categories. Our work differs by operating at the level of *individual neurons* and specifically testing cross-domain transferability.

2.2 Mechanistic Interpretability and Neuron-Level Analysis

In the context of LLMs, **knowledge neurons** (Dai et al., 2022) are neurons that, when suppressed, prevent expression of specific factual knowledge. Meng et al. (2022) develop ROME based on the finding that factual associations are localized in specific MLP layers. **H-neurons** (Gao et al., 2025) extend this paradigm to hallucination using the CETT metric, identifying fewer than 0.1% of neurons as hallucination-predictive with a causal role via activation scaling. Yu and Ananiadou (2024) attribute different knowledge types to distinct neuron populations, and Su et al. (2025) identify "value neurons" encoding social and ethical values. **Sparse autoencoders** (Bricken et al., 2023) decompose polysemantic neurons into interpretable units; Lan et al. (2024) examine whether SAE features transfer across models but do not test cross-domain transfer within the same model. Our work tests whether H-neuron functional specialization extends across domain boundaries.

2.3 Domain-Specific Hallucination Evaluation

A growing body of work documents that hallucination patterns differ across domains. MedHallu (Pal et al., 2024) shows medical hallucination has distinct characteristics, FinanceBench (Islam et al., 2024) reveals plausible but incorrect numerical claims, and CodeHalu (Tian et al., 2024) taxonomizes code hallucinations into execution, planning, and API-level errors. LegalBench (Guha et al., 2024) reveals systematic failure patterns differing from general QA errors. These benchmarks document *behavioral* differences but none investigate whether they correspond to *mechanistic* differences. Our work bridges this gap.

2.4 Chain-of-Thought and Hallucination

Chain-of-thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022) improves reasoning but has a complex relationship with hallucination. Most relevant, Cheng et al. (2025) demonstrate that CoT *obscures* hallucination detection signals, attributing this to CoT redistributing information across longer generations. Their analysis operates at the behavioral level; our study includes both direct and CoT conditions, enabling us to test whether CoT affects *neuron-level* hallucination signatures.

3 Methodology

Figure 1 provides an overview of our experimental pipeline.

3.1 H-Neuron Identification via CETT

We follow the neuron identification procedure of Gao et al. (2025). For a given input prompt x, the

Figure 1: Overview of the cross-domain H-neuron transfer protocol. For each model, we extract CETT-based neuron activation features across 6 domains, train per-domain L1-regularized logistic regression classifiers, and evaluate every classifier on every domain to construct a 6×6 transfer matrix.

FFN in each transformer layer ℓ computes:

$$\mathrm{FFN}^{(\ell)}(h) = W_{\text{down}}^{(\ell)} \cdot \sigma\left(W_{\text{up}}^{(\ell)} h + W_{\text{gate}}^{(\ell)} h\right) \quad (1)$$

where h is the hidden state and σ is SiLU. The **CETT** metric quantifies neuron j in layer ℓ's contribution to the output:

$$\mathrm{CETT}_j^{(\ell)} = a_j^{(\ell)} \cdot \left\| w_j^{(\ell)} \right\|_2 \quad (2)$$

where $a_j^{(\ell)}$ is the post-activation value and $w_j^{(\ell)}$ is the j-th column of $W_{\text{down}}^{(\ell)}$. For each input, we hook into the `down_proj` layers and record CETT scores for all FFN neurons across all layers, producing feature vectors of 322K–917K dimensions depending on architecture. Each input is labeled as hallucinating or not based on exact-match comparison with ground truth.

3.2 Experimental Setup

Table 1 summarizes our models and domains. We evaluate 5 open-weight instruction-tuned models spanning 3B to 8B parameters across 6 domains representing fundamentally different knowledge types. For each domain, we sample 1,000 questions from the respective benchmark. We evaluate two prompting conditions: direct prompting and chain-of-thought (CoT), where the model is instructed to "think step by step" (Kojima et al., 2022).

3.3 Cross-Domain Transfer Protocol

For each model m and prompting strategy $s \in \{\text{direct}, \text{CoT}\}$, we construct a 6×6 transfer matrix T_m^s:

Model	Params	Layers
Qwen2.5-3B-Instruct	3B	36
Nemotron-Mini-4B-Instruct	4B	32
Phi-3.5-mini-instruct	3.8B	32
Mistral-7B-Instruct-v0.3	7B	32
Llama-3.1-8B-Instruct	8B	32
Domain	**Dataset**	**Type**
General QA	TriviaQA	Open QA
Legal	CaseHOLD	5-way MC
Financial	FinancialPhraseBank	3-way cls.
Science	ARC-Challenge	4-way MC
Moral	ETHICS	Binary cls.
Code	Devign	Binary cls.

Table 1: Models and domains used in our study. MC = multiple choice, cls. = classification. Full dataset details in Appendix A.

1. **Neuron extraction.** For each domain d_i, we extract CETT features and hallucination labels, producing dataset $\mathcal{D}_i = \{(\mathbf{f}_k, y_k)\}_{k=1}^{n_i}$.

2. **Classifier training.** For each source domain d_i, we train an L1-regularized logistic regression classifier C_i using 5-fold cross-validation to select $C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$, maximizing cross-validated AUROC.

3. **Cross-domain evaluation.** Each C_i is evaluated on every target domain d_j (including d_i for the within-domain baseline), reporting AUROC.

The **transfer gap** is defined as:

$$\Delta = \frac{1}{D}\sum_{i=1}^{D} T[i,i] - \frac{1}{D(D-1)}\sum_{i \neq j} T[i,j] \quad (3)$$

where $D = 6$. A large positive Δ indicates domain specificity. We validate all results with bootstrap CIs (1,000 resamples), permutation testing (10,000 iterations for the aggregate gap), and multi-seed stability analysis. Details are in Appendix F.

4 Results

4.1 Within-Domain H-Neuron Detection

Before examining transfer, we establish that H-neuron classifiers achieve meaningful within-domain performance. Table 2 reports the diagonal entries of the transfer matrix.

Within-domain AUROC ranges from 0.542 (Llama-3.1-8B on code) to 0.929 (Mistral-7B on science), with a grand mean of 0.783. Five of six domains achieve mean AUROC above 0.79. The

Model	Gen.	Leg.	Fin.	Sci.	Mor.	Code
Qwen-3B	.896	.795	.882	.662	.836	.597
Nemotron-4B	.876	.807	.677	.753	.872	.563
Phi-3.5	.902	.768	.838	.871	.899	.557
Mistral-7B	.824	.862	.767	.929	.818	.666
Llama-8B	.805	.747	.890	.797	.805	.542
Mean	**.861**	**.796**	**.811**	**.802**	**.846**	**.585**

Table 2: Within-domain AUROC for H-neuron classifiers (direct prompting). All values except code substantially exceed chance (0.5). Bootstrap CIs in Appendix F.

Train↓ Test→	Gen.	Leg.	Fin.	Sci.	Mor.	Code
General	.861	.582	.617	.604	.634	.510
Legal	.525	.796	.556	.698	.552	.517
Financial	.526	.595	.811	.560	.602	.494
Science	.559	.722	.542	.802	.572	.523
Moral	.601	.578	.618	.547	.846	.530
Code	.470	.507	.508	.542	.508	.585

Table 3: Mean cross-domain AUROC across 5 models (direct prompting). Shaded diagonal entries are within-domain; off-diagonal entries are cross-domain transfers.

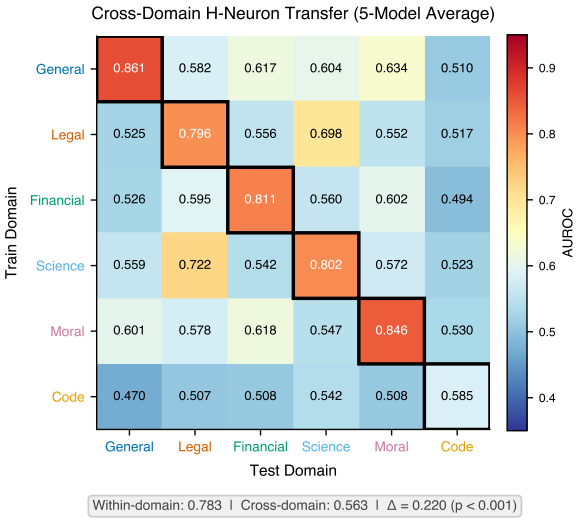

Figure 2: Cross-domain transfer matrix (mean AUROC across 5 models, direct prompting). Diagonal entries (within-domain) are consistently higher than off-diagonal entries (cross-domain). The transfer gap $\Delta = 0.220$ is statistically significant ($p < 0.001$, permutation test).

exception is code vulnerability detection (Devign), averaging only 0.585, likely reflecting the fundamentally different nature of code vulnerability detection compared to factual recall.

4.2 Cross-Domain Transfer Analysis

The central finding of this paper is that H-neuron classifiers fail to transfer across domains. Figure 2 shows the full transfer matrix averaged across models.

Aggregate transfer gap. The mean within-domain AUROC is 0.783 (95% CI: [0.742, 0.821]) and the mean cross-domain AUROC is 0.563 (95% CI: [0.545, 0.581]), yielding $\Delta = 0.220$ ($p < 0.001$, permutation test with 10,000 iterations; Cohen's $d = 1.99$). A transferred classifier loses approximately 28% of its discriminative power. After Benjamini-Hochberg FDR correction across all 60 within-domain experiments, 57 remain significant at $\alpha = 0.05$.

Per-model consistency. The transfer gap is consistent across all five models: Llama-8B ($\Delta = 0.160$), Nemotron-4B ($\Delta = 0.198$), Qwen-3B ($\Delta = 0.201$), Mistral-7B ($\Delta = 0.247$), and Phi-3.5 ($\Delta = 0.294$). No model achieves a gap below 0.16.

Domain-pair analysis. Table 3 reports the mean cross-domain AUROC for each source→target pair. Several patterns emerge:

1. **Legal-science partial transfer.** The legal→science and science→legal transfers show relatively high AUROC (0.698 and 0.722 respectively), substantially above the off-diagonal mean of 0.563. On Mistral-7B, legal→science achieves 0.873 and science→legal 0.845, nearly matching within-domain performance. This suggests shared hallucination-related neural substrates, possibly because both involve structured logical reasoning.

2. **Code as an outlier.** Classifiers trained on code transfer poorly to all other domains (mean off-diagonal AUROC = 0.507) and other domains transfer poorly to code (0.515). Code vulnerability detection involves maximally distinct neural pathways.

3. **Below-chance transfers.** Several transfers produce AUROC below 0.5, meaning the classifier is *anti-predictive*. Notable cases include Nemotron-4B code→moral (0.210) and Phi-3.5 legal→moral (0.256). These indicate not merely absence of shared signal but active interference: neurons that predict hallucination in one domain predict *correct* answers in another (see Section 5.3).

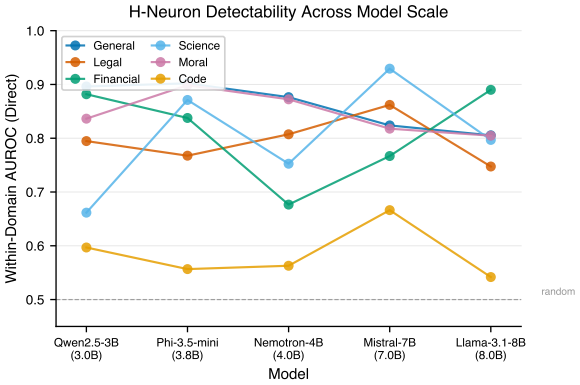

Figure 3: Domain specificity (transfer gap Δ) by model size. Domain specificity does not consistently increase or decrease with scale in the 3B–8B range.

4.3 Chain-of-Thought Effects

We examine how CoT prompting affects H-neuron detection. Across all 30 model-domain pairs, mean within-domain AUROC under CoT is 0.806 compared to 0.783 under direct prompting ($\Delta = +0.023$). However, 6 of 30 pairs produced identical results under both strategies (response caching). Among the 24 genuine pairs, CoT improves detection in 18 cases and degrades it in 6. The largest improvements occur on general QA (Mistral-7B: $0.824 \rightarrow 0.995$), while the largest degradations are Llama-8B on general QA ($\Delta = -0.107$) and Mistral-7B on science ($\Delta = -0.107$). Crucially, CoT substantially changes the number and identity of selected H-neurons (shifts ranging from -76 to $+100$), providing neuron-level evidence for the behavioral findings of Cheng et al. (2025): CoT does not merely suppress or enhance the hallucination signal but *reorganizes* it across different neuron populations. Full per-model CoT results are in Appendix D.

4.4 Model Scale Analysis

The relationship between model size and domain specificity is not monotonic in our 3B–8B range (Figure 3). The smallest model (Qwen-3B) shows $\Delta = 0.201$, while the largest (Llama-8B) shows the *smallest* gap ($\Delta = 0.160$) and Phi-3.5 (3.8B) shows the *largest* ($\Delta = 0.294$). This suggests that architecture-specific factors (attention-to-FFN ratio, training data, instruction tuning) play a larger role than raw parameter count, though qualitatively different behavior might emerge at larger scales (70B+).

5 Analysis and Discussion

5.1 Mechanistic Explanations

Our transfer results demonstrate that the neurons predictive of hallucination differ across knowledge domains. We consider three mechanistic explanations: (1) *domain-specific knowledge storage*, where H-neurons are "broken knowledge neurons" (Dai et al., 2022) that differ because knowledge itself is stored in different FFN regions; (2) *domain-specific reasoning circuits*, where hallucination arises from malfunctions in domain-specific computation pathways; and (3) *domain-specific calibration*, where H-neurons capture domain-varying miscalibration rather than hallucination per se. Our data favor Hypotheses 1–2 over 3: the below-chance transfers suggest active interference (more consistent with conflicting pathways than independent calibration errors), and the partial legal-science transfer aligns with shared reasoning circuits rather than calibration artifacts.

5.2 The Legal-Science Bridge

The partial transfer between legal and science domains (legal→science = 0.722, science→legal = 0.698 averaged across models; on Mistral-7B: 0.873 and 0.845) suggests shared hallucination-predictive neurons. Both legal reasoning (applying rules to facts) and scientific reasoning (applying principles to observations) involve structured logical inference. If hallucination in both domains arises from failures in the same logical-inference circuits, the corresponding H-neurons would overlap. This is distinct from moral reasoning or code analysis, which involve different cognitive processes and show no such bridge. This finding suggests H-neurons may be organized along *reasoning-type* boundaries rather than content-domain boundaries.

5.3 Below-Chance Transfers

When a classifier trained on code achieves AUROC = 0.210 on moral reasoning (Nemotron-4B), the neurons that predict hallucination in code predict *correct answers* in moral reasoning. This is evidence of *anti-correlated* hallucination mechanisms. One interpretation, consistent with the polysemanticity literature (Bricken et al., 2023), is that certain neurons serve dual roles: contributing to faithful processing in one domain while being co-opted during processing in another. This has practical

consequences: suppressing H-neurons from one domain may *increase* hallucination in another.

5.4 Deployment Implications

Our results challenge the implicit assumption that hallucination detectors developed on one benchmark generalize broadly. Concretely: (1) deploying neuron-level detectors in a new domain requires domain-specific hallucination labels and retraining; (2) multi-domain systems need separate detection per domain or a domain-routing mechanism; and (3) neuron-level interventions must target domain-specific neuron sets, as our below-chance transfers show that suppressing the wrong neurons can be counterproductive. We additionally conducted causal intervention experiments via activation scaling of H-neurons (Appendix E), finding no significant effect on hallucination rates, suggesting H-neurons are reliable diagnostic markers but not sufficient causal levers at this sparsity level.

6 Conclusion

We have presented the first systematic cross-domain transfer analysis of hallucination neurons in large language models. Across 6 domains and 5 models, we find a consistent and statistically significant transfer gap: H-neuron classifiers achieve AUROC 0.783 within their training domain but only 0.563 when transferred to a different domain ($\Delta = 0.220$, $p < 0.001$). This finding holds across all five models tested, spanning 3B to 8B parameters.

Our results establish that hallucination is not a monolithic phenomenon with a single neural signature. Instead, different knowledge domains activate different hallucination-associated neuron populations. We further find that some domain pairs (particularly legal and science) exhibit partial transfer, suggesting that reasoning-type similarity may be the organizing principle for H-neuron specialization.

Future directions. Testing at larger scales (70B+) would reveal whether domain specificity diminishes with capacity. Finer-grained domain taxonomies could test whether the legal-science bridge generalizes to a broader "structured reasoning" category. Extending the analysis to sparse autoencoders (Bricken et al., 2023) could test whether domain specificity persists at the level of interpretable features rather than raw neurons.

Limitations

Our study has several limitations that qualify the scope of our conclusions.

Model scale. We test models ranging from 3B to 8B parameters. It remains an open question whether the domain specificity of H-neurons persists at larger scales (e.g., 70B+), where models may develop more universal internal representations due to greater capacity.

Domain granularity. Our six domains (general QA, legal, financial, science, moral reasoning, code) represent broad categories. Finer-grained domain distinctions (e.g., contract law vs. constitutional law within legal) may reveal additional structure in H-neuron specialization that our current experimental design does not capture.

Sample sizes. While we use 1,000 questions per domain for neuron identification, the evaluation sample sizes vary across domains (26–264 samples per transfer cell). Some domain-model combinations, particularly moral reasoning on Llama-3.1-8B ($n = 26$), have limited statistical power.

Correlational evidence. Our cross-domain transfer analysis demonstrates that H-neuron classifiers fail to generalize, which is consistent with domain-specific neural pathways but does not constitute causal proof. Our activation-scaling experiments (Appendix E) found no significant causal effect of H-neuron modulation on hallucination rates, suggesting that the diagnostic signal captured by CETT does not concentrate sufficient causal influence for intervention at this sparsity level.

CoT response caching. Six of 30 model-domain pairs produced identical outputs under direct and CoT prompting, limiting our CoT analysis to 24 genuinely distinct pairs. This reduces our statistical power for CoT-related conclusions.

CETT dependence. Our findings are specific to the CETT attribution metric for neuron identification. Other attribution methods (e.g., integrated gradients, activation patching) might yield different neuron sets with potentially different transfer properties.

Ethics Statement

This work investigates the internal mechanisms of hallucination in language models. We do not develop new capabilities for generating misinforma-

tion; rather, our findings contribute to understanding *why* models hallucinate and the limitations of current detection approaches. All datasets used are publicly available research benchmarks. We use open-weight models and release our code and analysis scripts to support reproducibility. Our finding that hallucination detectors trained on one domain do not generalize has direct implications for responsible AI deployment: practitioners should not assume that a single hallucination detection system will work across all application domains.

Reproducibility Statement

All code, configurations, datasets, and analysis scripts required to reproduce our experiments are available at https://anonymous.4open.science/r/h-neurons-cross-domain-709C/. The repository includes: (i) dataset loading scripts for all six benchmarks (TriviaQA, CaseHOLD, Financial PhraseBank, ARC-Challenge, ETHICS commonsense, Devign); (ii) the full CETT-based H-neuron extraction pipeline with hooks into `down_proj` layers; (iii) classifier training, cross-domain transfer, and robustness analysis scripts; (iv) the experiment configuration file (`experiment_config.yaml`) specifying all hyperparameters, sample sizes (1,000 questions per domain), and random seeds. We use publicly available model checkpoints from Hugging Face for all five models (Qwen2.5-3B, Nemotron-Mini-4B, Phi-3.5-mini, Mistral-7B-Instruct-v0.3, and Meta-Llama-3.1-8B-Instruct via the NousResearch ungated mirror). Upon acceptance, the anonymous repository will be released as a public GitHub repository under the MIT license.

Use of AI Assistance

In accordance with the ACL/EMNLP policy on the use of large language models, we disclose that we used Claude (Anthropic) to assist with paper drafting, prose refinement, code review, and copy-editing throughout this project. All research contributions — including the cross-domain transfer protocol design, experimental execution, statistical analysis, result interpretation, and all scientific claims — are the authors' own. No AI-generated content was included in the paper without authorial verification, and all references, numerical results, and experimental procedures were manually validated against source materials and experiment outputs.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Qianqian Chen and 1 others. 2024. FinBen: A holistic financial benchmark for large language models. *Proceedings of NeurIPS Datasets and Benchmarks*.

Canyu Cheng and 1 others. 2025. Can LLM-generated misinformation be detected? In *Findings of EMNLP*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of ACL*, pages 8493–8502.

Xuefeng Du, Chaowei Xue, Yifei Li, and Yixuan Li. 2024. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Proceedings of NeurIPS*.

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. 2025. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in LLMs. *arXiv preprint arXiv:2512.01797*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Arvind Narayanan, Brandon Choi, Catalin Cui, Felix Daumler, Amit Deshpande, and 1 others. 2024. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Proceedings of NeurIPS Datasets and Benchmarks*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Pranab Islam and 1 others. 2024. FinanceBench: A new benchmark for financial question answering. *arXiv preprint*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of NeurIPS*.

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of ICLR*.

Michael Lan, Philip Goldowsky-Dill, and Max Tegmark. 2024. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, pages 9004–9017.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Proceedings of NeurIPS*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *Proceedings of ICLR*.

Vibhor Pal and 1 others. 2024. MedHalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in "large" foundation models. *arXiv preprint arXiv:2309.05922*.

Hangyu Su and 1 others. 2025. Value neurons: Discovering and locating value representations in language models. In *Findings of EMNLP*.

Yuchen Tian and 1 others. 2024. CodeHalu: Code hallucinations in LLMs driven by execution-based verification. *arXiv preprint*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Zeping Yu and Sophia Ananiadou. 2024. Neuron-level knowledge attribution in large language models. In *Proceedings of EMNLP*, pages 3267–3280.

A Dataset Details

Table 4 provides detailed statistics for each domain's benchmark dataset.

B Per-Model Transfer Matrices

Tables 5–9 present the full 6×6 transfer matrices for each model under direct prompting.

C Sample Size Details

Table 10 reports the evaluation sample sizes for each model-domain combination.

D Chain-of-Thought Detailed Results

Table 11 reports within-domain AUROC under both prompting strategies for all model-domain pairs.

E Causal Intervention via Activation Scaling

Following the activation-scaling methodology of Gao et al. (2025), we tested whether modulating H-neuron activations at inference time can increase or suppress domain-specific hallucination rates. We scaled the activations of the top-k H-neurons (identified per domain via CETT) by factors $\alpha \in \{0.0, 0.5, 1.5, 2.0, 3.0\}$ during forward passes on 200 held-out questions per domain across three models (Qwen2.5-3B, Mistral-7B, Llama-3.1-8B), using greedy decoding. A random-neuron control applied identical scaling to neurons selected uniformly at random.

The result is negative. Zeroing H-neurons entirely ($\alpha = 0.0$) produced a within-domain hallucination rate change of only $\Delta = +0.0019$ (paired t-test $p = 0.316$); cross-domain effects were negligible ($\Delta = +0.0001$). Amplification at $\alpha = 3.0$ yielded comparably small shifts ($\Delta_{\text{within}} = -0.0006$, $\Delta_{\text{cross}} = -0.0042$, $p = 0.535$). The random-neuron control showed $\Delta \approx 0.000$ at all scales. No condition reached statistical significance, and effect sizes were uniformly below one percentage point.

We interpret this null result as evidence that H-neurons are reliable *diagnostic markers* of hallucination but not sufficient *causal levers* at the sparsity level we operate (<0.1% of FFN neurons). This is consistent with distributed, redundant hallucination circuits: the model's tendency to hallucinate is spread across many neurons, and the small subset captured by CETT carries predictive signal without concentrating causal influence. We note that Gao et al. (2025) reported scaling effects on a different behavioral measure (over-compliance rather than factual accuracy), suggesting that the causal role of H-neurons may be behavior-dependent.

Domain	Dataset	Split	Samples Used	Task Type	Answer Format
General QA	TriviaQA	test	1000	Open-ended factual QA	Free-form text
Legal	CaseHOLD	test	1000	Holding identification	5-way multiple choice
Financial	FinancialPhraseBank	train	1000	Sentiment classification	3-way classification
Science	ARC-Challenge	test	1000	Science reasoning	4-way multiple choice
Moral	ETHICS (commonsense)	test	1000	Moral judgment	Binary classification
Code	Devign	test	1000	Vulnerability detection	Binary classification

Table 4: Dataset details for each domain. Samples Used refers to questions presented to each model; see Table 10 for evaluation counts after filtering.

Train ↓ / Test →	General	Legal	Financial	Science	Moral	Code
General	**.805**	.571	.690	.563	.876	.529
Legal	.607	**.747**	.707	.734	.604	.588
Financial	.593	.659	**.890**	.606	.763	.534
Science	.593	.754	.519	**.797**	.574	.520
Moral	.602	.561	.641	.522	**.805**	.498
Code	.489	.539	.558	.463	.651	**.542**

Table 5: Transfer matrix for Llama-3.1-8B-Instruct (direct prompting). Within-domain diagonal in bold.

Train ↓ / Test →	General	Legal	Financial	Science	Moral	Code
General	**.824**	.576	.685	.742	.725	.481
Legal	.351	**.862**	.607	.873	.374	.448
Financial	.642	.579	**.767**	.731	.586	.514
Science	.393	.845	.627	**.929**	.513	.475
Moral	.651	.631	.559	.384	**.818**	.527
Code	.411	.436	.423	.593	.529	**.666**

Table 6: Transfer matrix for Mistral-7B-Instruct-v0.3 (direct prompting).

Train ↓ / Test →	General	Legal	Financial	Science	Moral	Code
General	**.876**	.536	.618	.552	.444	.586
Legal	.579	**.807**	.406	.582	.779	.516
Financial	.498	.535	**.677**	.391	.737	.498
Science	.592	.738	.590	**.753**	.718	.647
Moral	.524	.563	.663	.669	**.872**	.512
Code	.488	.644	.513	.470	.210	**.563**

Table 7: Transfer matrix for Nemotron-Mini-4B-Instruct (direct prompting).

Train ↓ / Test →	General	Legal	Financial	Science	Moral	Code
General	**.902**	.597	.451	.665	.511	.474
Legal	.467	**.768**	.491	.642	.256	.548
Financial	.516	.645	**.838**	.532	.461	.404
Science	.604	.605	.374	**.871**	.332	.423
Moral	.605	.490	.558	.625	**.899**	.588
Code	.462	.465	.449	.631	.488	**.557**

Table 8: Transfer matrix for Phi-3.5-mini-instruct (direct prompting).

Train ↓ / Test →	General	Legal	Financial	Science	Moral	Code
General	**.896**	.633	.641	.498	.614	.479
Legal	.619	**.795**	.567	.660	.747	.485
Financial	.383	.555	**.882**	.541	.460	.519
Science	.614	.668	.598	**.662**	.724	.548
Moral	.624	.642	.667	.534	**.836**	.524
Code	.500	.450	.597	.552	.662	**.597**

Table 9: Transfer matrix for Qwen2.5-3B-Instruct (direct prompting).

Domain	Qwen	Nem.	Phi	Mis.	Llama
General	160	200	212	142	90
Legal	230	104	180	82	88
Financial	128	86	110	112	80
Science	100	106	56	76	64
Moral	124	66	192	90	26
Code	264	240	256	103	182

Table 10: Evaluation sample sizes per model and domain.

Domain	Qwen	Nem.	Phi	Mis.	Llama
Direct prompting					
General	.896	.876	.902	.824	.805
Legal	.795	.807	.768	.862	.747
Financial	.882	.677	.838	.767	.890
Science	.662	.753	.871	.929	.797
Moral	.836	.872	.899	.818	.805
Code	.597	.563	.557	.666	.542
Chain-of-thought prompting					
General	.945	.982	.993	.995	.698
Legal	.787	.808	.770	.924	.738
Financial	.901	.677†	.838†	.785	.890†
Science	.702	.786	.906	.823	.796
Moral	.874	.872†	.878	.898	.852
Code	.639	.563†	.603	.726	.542†

Table 11: Within-domain AUROC under direct vs. CoT prompting. † indicates pairs where the model produced identical responses under both strategies, yielding identical classifier results.

F Robustness Analysis Details

Table 12 and Table 13 report the robustness analysis results for direct and CoT prompting respectively. Each experiment includes a bootstrap 95% confidence interval (1,000 resamples), 5-fold cross-validation mean and standard deviation, permutation test p-value (20 permutations), and an overall verdict. Experiments are classified as ROBUST if the bootstrap CI excludes 0.5 and the permutation test yields $p < 0.05$, and WEAK otherwise. Overall, 54 of 60 experiments are classified as ROBUST; the 6 WEAK experiments involve the code domain (5 cases) or have borderline CIs (1 case: Qwen2.5-3B science under CoT). After Benjamini-Hochberg FDR correction across all 60 experiments, 57 remain statistically significant at $\alpha = 0.05$.

For per-experiment significance testing, we use 20 permutations per classifier due to the computational cost of retraining L1-regularized classifiers on high-dimensional feature spaces. For multi-seed stability, we retrain each classifier with 5 different random seeds and measure the Jaccard similarity of selected neuron sets across seeds. The mean Jaccard stability is 0.467 ($\sigma = 0.078$, range 0.351–0.640), indicating moderate agreement in neuron selection.

Model	Domain	AUROC [95% CI]	CV (mean ± std)	Perm. p	Verdict
Qwen2.5-3B	General	.925 [.882, .963]	.920 ± .013	.000	ROBUST
	Legal	.788 [.725, .843]	.794 ± .035	.000	ROBUST
	Financial	.749 [.655, .839]	.806 ± .026	.000	ROBUST
	Science	.783 [.691, .866]	.763 ± .076	.000	ROBUST
	Moral	.818 [.746, .889]	.811 ± .020	.000	ROBUST
	Code	.604 [.538, .673]	.621 ± .040	.000	ROBUST
Nemotron-4B	General	.919 [.880, .955]	.888 ± .026	.000	ROBUST
	Legal	.843 [.758, .919]	.853 ± .026	.000	ROBUST
	Financial	.754 [.647, .851]	.752 ± .062	.000	ROBUST
	Science	.811 [.730, .889]	.779 ± .071	.000	ROBUST
	Moral	.884 [.786, .965]	.872 ± .031	.000	ROBUST
	Code	.628 [.559, .697]	.603 ± .041	.000	ROBUST
Phi-3.5-mini	General	.904 [.856, .944]	.912 ± .020	.000	ROBUST
	Legal	.769 [.692, .838]	.769 ± .031	.000	ROBUST
	Financial	.773 [.682, .857]	.780 ± .033	.000	ROBUST
	Science	.957 [.904, .991]	.916 ± .029	.000	ROBUST
	Moral	.898 [.845, .941]	.901 ± .035	.000	ROBUST
	Code	.570 [.497, .641]	.566 ± .030	.000	WEAK
Mistral-7B	General	.850 [.783, .911]	.808 ± .028	.000	ROBUST
	Legal	.857 [.757, .936]	.875 ± .055	.000	ROBUST
	Financial	.745 [.648, .834]	.791 ± .031	.000	ROBUST
	Science	.865 [.775, .944]	.921 ± .037	.000	ROBUST
	Moral	.872 [.792, .934]	.839 ± .052	.000	ROBUST
	Code	.726 [.630, .819]	.643 ± .054	.000	ROBUST
Llama-3.1-8B	General	.839 [.738, .930]	.865 ± .032	.000	ROBUST
	Legal	.862 [.783, .942]	.874 ± .047	.000	ROBUST
	Financial	.662 [.529, .777]	.769 ± .071	.000	ROBUST
	Science	.850 [.747, .931]	.837 ± .066	.000	ROBUST
	Moral	.753 [.544, .935]	.869 ± .079	.000	ROBUST
	Code	.537 [.456, .620]	.558 ± .076	.600	WEAK

Table 12: Robustness analysis for all experiments under direct prompting. AUROC is the bootstrap point estimate with 95% CI (1,000 resamples). CV reports 5-fold cross-validated AUROC.

Model	Domain	AUROC [95% CI]	CV (mean ± std)	Perm. p	Verdict
Qwen2.5-3B	General	.924 [.880, .961]	.942 ± .016	.000	ROBUST
	Legal	.793 [.738, .850]	.788 ± .017	.000	ROBUST
	Financial	.830 [.750, .901]	.809 ± .009	.000	ROBUST
	Science	.607 [.499, .724]	.692 ± .055	.000	WEAK
	Moral	.850 [.779, .908]	.836 ± .021	.000	ROBUST
	Code	.646 [.574, .713]	.628 ± .041	.050	WEAK
Nemotron-4B	General	.973 [.947, .992]	.973 ± .020	.000	ROBUST
	Legal	.849 [.766, .923]	.854 ± .025	.000	ROBUST
	Financial†	.754 [.647, .851]	.752 ± .062	.000	ROBUST
	Science	.793 [.712, .871]	.773 ± .067	.000	ROBUST
	Moral†	.884 [.786, .965]	.872 ± .031	.000	ROBUST
	Code†	.628 [.559, .697]	.603 ± .041	.000	ROBUST
Phi-3.5-mini	General	.980 [.961, .993]	.988 ± .006	.000	ROBUST
	Legal	.810 [.746, .875]	.809 ± .029	.000	ROBUST
	Financial†	.773 [.682, .857]	.780 ± .033	.000	ROBUST
	Science	.937 [.872, .986]	.856 ± .041	.000	ROBUST
	Moral	.835 [.770, .892]	.900 ± .038	.000	ROBUST
	Code	.555 [.480, .625]	.583 ± .033	.000	WEAK
Mistral-7B	General	.958 [.908, .996]	.963 ± .012	.000	ROBUST
	Legal	.906 [.838, .963]	.890 ± .034	.000	ROBUST
	Financial	.787 [.695, .867]	.806 ± .025	.000	ROBUST
	Science	.871 [.783, .946]	.917 ± .018	.000	ROBUST
	Moral	.768 [.672, .858]	.829 ± .036	.000	ROBUST
	Code	.776 [.688, .853]	.716 ± .054	.000	ROBUST
Llama-3.1-8B	General	.756 [.601, .895]	.756 ± .032	.000	ROBUST
	Legal	.861 [.779, .938]	.872 ± .048	.000	ROBUST
	Financial†	.662 [.529, .777]	.769 ± .071	.000	ROBUST
	Science	.821 [.710, .914]	.832 ± .080	.000	ROBUST
	Moral	.898 [.744, .994]	.877 ± .088	.000	ROBUST
	Code†	.537 [.456, .620]	.558 ± .076	.600	WEAK

Table 13: Robustness analysis for all experiments under CoT prompting. † marks experiments identical to direct prompting (response caching). See Table 12 caption for details.