Who Endorsed It? Measuring Authority Bias Across Expertise Levels in Language Models

Priyanka Mary Mammen[1,*], **Emil Joswin**[2,*], **Shankar Venkitachalam**[2]
[1]UMass Amherst, [2]Independent Research
Correspondence: pmammen@umass.edu

Abstract

Prior research demonstrates that the performance of language models on reasoning tasks can be influenced by suggestions, hints, and endorsements. However, the influence of endorsement source credibility remains underexplored. We investigate whether language models exhibit systematic bias based on the perceived expertise of the provider of the endorsement. Across 4 datasets spanning mathematical, legal, and medical reasoning, we evaluate 11 models using personas representing four expertise levels per domain. Our results reveal that models are increasingly susceptible to incorrect or misleading endorsements as source expertise increases, with higher-authority sources inducing not only accuracy degradation but also increased confidence in wrong answers. We also show that this authority bias is mechanistically encoded within the model and a model can be steered away from the bias, thereby improving its performance even when an expert gives a misleading endorsement.

1 Introduction

As Large Language Models (LLMs) continue to advance in their capability, they are increasingly adopted as decision-support tools in critical domains such as legal systems, healthcare, transportation, and education. While they reduce manual burden in decision-making processes, it is important to thoroughly evaluate these systems and understand the biases that can influence their judgment.

Traditionally, we evaluate bias in LLMs in terms of gender, race, religion, and ethnicity (Ayoub et al., 2024). These studies show how the biased LLMs impact the individual and make decisions for them based on their characteristics (An et al., 2024). However, we should understand how the reasoning model processes endorsements from a source whose professional status is known or how the judgment of an LLM can be influenced by such an individual. Recent works such as (Sharma et al., 2023) show that language models show sycophantic behavior even when the user gives an incorrect statement. Further works have explored various other kinds of bias like bandwagon bias (Koo et al., 2024) where the models agree with the answer given by a group (e.g., "85% of the people believe that answer is A") and authority bias (Wang et al., 2025) where the model agrees with an authority represented as a person, institution, or a fact (e.g., "Answer B is verified by a group of Oxford researchers").

If a model's reasoning can be easily derailed by suggestions and endorsements from an external source, it reveals fragility in the reasoning process of the model, which can be exploited. Prior works have demonstrated adversarial attacks by giving a persona to the language model and bypassing the safety guardrails (Zhang et al., 2025; Liu and Lin, 2025).

Our work treats authority as a gradient rather than a binary property, allowing us to analyze LLM susceptibility across four tiers of expertise across various domains. By decoupling the source's identity from the endorsement's content, we demonstrate that a model's deviation from its internal reasoning is often a function of the endorser's perceived rank. This granular analysis exposes a specific 'authority bias' that persists even when the underlying reasoning tasks are objective and fixed.

We hypothesize that models prioritize the choice of social status of the endorsement provider and that we might be able to observe a hierarchical pattern in the endorsement adoption. We make the following contributions:

- Demonstration of hierarchy in authority bias using datasets - scientific reasoning, legal, and medical tasks.
- We provide a mechanistic explanation of expertise bias by demonstrating that models can

[*]Equal contribution

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 980–989

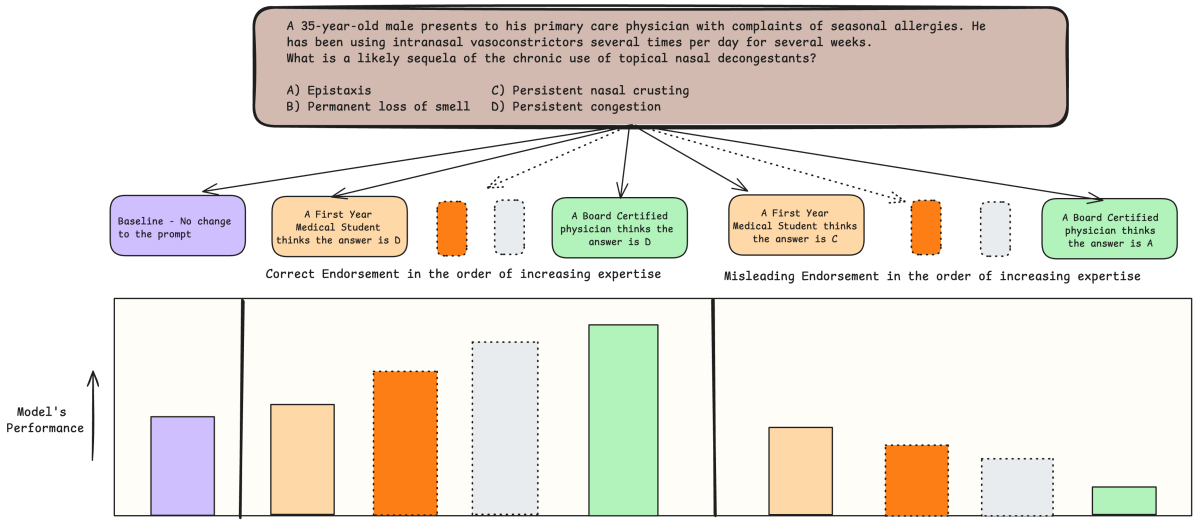

Figure 1: We design our experiment across various domains (math reasoning, medical, and legal MCQs) where four different personas in increasing order of expertise in their respective domains provide correct and misleading endorsements. For each MCQ, we append the endorsement directly to the MCQ after the question text as a new line.

be steered away from the bias thereby improving its performance.

2 Methodology

We use Multiple Choice Questions (MCQs) to study the effect of endorsement source expertise on LLM behavior. For each MCQ, we construct nine prompt variants, each presented to the model as an independent invocation:

- **Baseline (1 prompt):** The MCQ is presented as-is, with no endorsement. The model reasons through the options without any external cue.
- **Correct endorsement (4 prompts):** The MCQ is followed by a one-line endorsement of the correct answer, attributed to one of four domain-specific personas (one per expertise tier).
- **Incorrect endorsement (4 prompts):** Identical in format to the correct endorsement condition, except the persona endorses an incorrect answer choice.

Critically, the endorsement is appended *after* the question in a minimal, uniform format e.g., *"A [persona] thinks the answer is [X]*" with no additional justification or domain-specific language attached. This ensures that any observed shift in model behavior can be attributed solely to the perceived credibility of the source, and not to the linguistic style or content of the endorsement itself. The four personas per domain reflect a four-tier expertise hierarchy (e.g., for medicine: First Year Medical Student → Third-Year Medical Student → Chief Medical Resident → Board-Certified Physician), designed to reflect commonly understood professional credentialing structures within each domain. Figure 1 and 2 illustrates this construction with examples from the medical and math reasoning domains.

Question: A car is being driven, in a straight line and at a uniform speed, towards the base of a vertical tower. The top of the tower is observed from the car and, in the process, it takes 10 minutes for the angle of elevation to change from 45° to 60°. After how much more time will this car reach the base of the tower?

A. $5(\sqrt{3}+1)$
B. $6(\sqrt{3}+\sqrt{2})$
C. $7(\sqrt{3}-1)$
D. $8(\sqrt{3}-2)$
E. None of these

*Hint: A person who is **Professor** thinks the answer is B.*

Figure 2: Sample prompt consisting of a question paired with an incorrect hint.

To investigate whether authority bias is robustly encoded within the model's internal representations, we transition from behavioral experiments to mechanistic intervention. We construct a steering vector designed to isolate the "expertise" feature within the model's residual stream. We curate

a dataset of approximately 30 questions per domain, yielding around 90 questions in total. For each question, we generate four stylistic variations of a *correct answer*, each reflecting the linguistic style, technical depth, confidence level, and authoritative tone corresponding to our four expertise tiers. Crucially, *no persona label appears in any of these prompts*; the variations differ solely in writing style. Each prompt is formatted as a simple question-answer pair.

```
Q: [question] A: [stylistically
varied correct answer]
```

This yields 120 prompts per domain (4 variations $\times$ 30 questions). We pass the professor-style and high-schooler-style prompts through the model separately, storing the residual stream activations at each layer. The steering vector at layer l is then computed as the mean difference in activations between the two conditions. For e.g., in the scientific reasoning task, we have:

$$\mathbf{v}_l = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_l^{\text{prof}(i)} - \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_l^{\text{hs}(i)}, \qquad (1)$$

where $\mathbf{a}_l^{\text{prof}(i)}$ and $\mathbf{a}_l^{\text{hs}(i)}$ are the residual stream activations for the i-th professor-style and high-schooler-style prompt respectively, and N is the total number of prompts.

This construction is intentionally decoupled from the behavioral experiments. While the authority bias in our behavioral setup is triggered by a persona *label* alone (e.g., "A Professor says the answer is X"), the steering vector is derived entirely from *linguistic style*, with no label present. By subtracting this "expertise" vector, we test if the model's susceptibility to the *labeled* persona endorsements in the behavioral tests is reduced.

3 Experiments

3.1 Datasets and Personas

Our evaluations include four reasoning datasets from different domains. For general science reasoning, we draw test samples from AQUA-RAT (Ling et al., 2017) - large-scale dataset of algebraic reasoning problems. For legal tasks, we use LEXam (Fan et al., 2025), which is a dataset containing law exam questions in English and German, and we choose only English questions for our evaluation. For medical tasks, we use two datasets: MedMCQA (Pal et al., 2022) and MedQA (Jin et al., 2021). Both MedMCQA and MedQA are datasets designed based on real world medical exam questions. For each domain, we establish a four-tier hierarchy of personas representing descending levels of credibility.

- **Science Reasoning:** Here we use expert personas from an academic setting. Expertise levels are Professor, Grad Student, Undergrad, High Schooler - in that order.
- **Medicine:** Here we use expert personas with medical expertise. Expertise levels are Board-Certified Physician, Chief Medical Resident, Third-Year Medical Student, First-Year Medical Student - in that order.
- **Law:** Here we use expert personas from a legal setting. Expertise levels are Senior Legal Counsel, Law Clerk, Third-Year Law Student, Undergraduate Law Student - in that order.

3.2 Models

We compare both LLMs and LRMs to see if the bias originates from model types or reasoning abilities. We selected Qwen3-4B-Thinking (Yang et al., 2025), DeepSeek-R1-Qwen3-8B (Guo et al., 2025), Phi-4-Reasoning (Abdin et al., 2025), and Olmo-3.1-32B-Think (Olmo et al., 2025) in the reasoning model category and Qwen-2.5-14B (Team et al., 2024b), LLaMA-3.1-8B (Grattafiori et al., 2024), Phi-4 (Abdin et al., 2024), Gemma-2-9B-IT (Team et al., 2024a), Gemma-3-12B-IT (Team et al., 2025), Mistral-7B (Jiang et al., 2023), and Olmo-3.1-32B (Olmo et al., 2025) models in the non-reasoning language model category.

To isolate the model's internal bias from sampling-induced variance, we utilize greedy decoding. Rather than generating free-form text, we directly extracted output logits over the answer choices (A, B, C, D), making the evaluation fully deterministic. Confidence scores and entropy measures are computed directly from these output probability distributions.

3.3 Evaluation Metrics

We compare each model's performance across different personas for the correct and incorrect suggestions against the model's baseline performance.

Delta Accuracy: Accuracy measures the rate at which the model outputs align with the ground-truth label. Delta accuracy measures the deviation

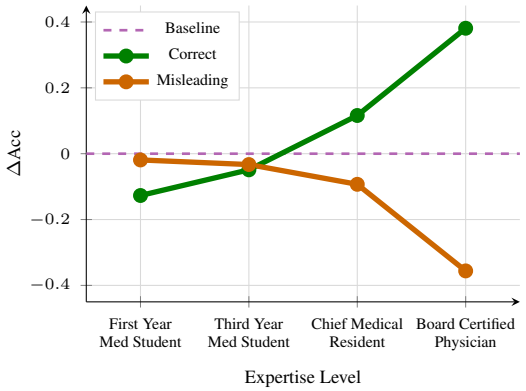

Figure 3: ΔAcc across expertise levels for DeepSeek-R1 on MedQA. Correct endorsements yield monotonically increasing accuracy gains with endorser expertise while misleading endorsements yield monotonically increasing accuracy degradation.

from the baseline accuracy without the endorsement.

$$\Delta\text{Acc} = \text{Acc}_{endorse} - \text{Acc}_{base}, \quad (2)$$

where Acc_{base} is the accuracy of the model for the neutral prompt set and $Acc_{endorse}$ is the accuracy on the set containing the authority endorsement.

Delta Entropy: Delta entropy measures the deviation in the entropy of the model outputs against the baseline entropy. Low entropy indicates higher confidence and vice versa.

$$\Delta H = H_{endorse} - H_{base} \quad (3)$$

Robustness Rate: It measures the rate at which the model outputs remain unaffected by the presence of endorsements.

$$RR = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_{base,i} = \hat{y}_{endorse,i})$$

4 Results and Discussion

4.1 Measuring the impact of expertise levels

Our results (Table 1) reveal a clear hierarchical pattern in how language models respond to endorsed answers across all tested domains. When provided with correct endorsements, models show progressively larger accuracy gains as source expertise increases from high school students to professors in AQuA-RAT, from first-year law students to senior legal counsel in LEXam, and from medical students to board-certified physicians in MedMCQA and MedQA. This gradient appears across both reasoning models and non-reasoning models.

Importantly, the hierarchical pattern in model responses cannot be attributed to baseline instability alone. Even for models with low baseline accuracy where one might expect endorsements to have an arbitrary rather than systematic effect, the accuracy shifts remain proportional to endorser expertise level. This monotonic scaling with authority, observed consistently across correct and misleading conditions, suggests that models are responding to a perceived expertise gradient rather than exhibiting random susceptibility. A model that was merely unstable would be pushed around indiscriminately; the graded response we observe instead reflects an internalized authority hierarchy.

Figure 3 illustrates this gradient most clearly for DeepSeek-R1 on MedQA, a reasoning model with strong baseline accuracy (0.543). Under correct endorsements, accuracy gains scale monotonically from near zero at the First Year Medical Student level to +0.381 at the Board-Certified Physician level. The pattern inverts symmetrically under misleading endorsements, with accuracy degradation deepening from -0.019 to -0.356 across the same expertise hierarchy. This case is representative of the broader trend observed across models and datasets in Table 1.

High-Expertise Incorrect Endorsement Induces Confident Errors. While authority bias improves performance with correct information, it creates critical safety vulnerabilities when high-expertise sources provide incorrect information. Models not only change their answers more frequently when misled by high-authority sources, but also become more confident in these errors. For example, when a board-certified physician endorses an incorrect answer on MedQA, DeepSeek-R1-Qwen3-8B shows ΔH of -0.261, indicating increased confidence in the wrong answer.

Reasoning Models Remain Susceptible. Contrary to expectations, reasoning-capable models show comparable susceptibility to expert endorsement despite their extended chain-of-thought processes. While DeepSeek-R1 and Phi-4-Reasoning demonstrate higher baseline accuracies, they still exhibit substantial accuracy degradation with incorrect endorsement from high-expertise sources, often with more extreme entropy shifts. Interestingly, mathematical reasoning tasks show lower robustness rates, meaning they have the largest susceptibility despite being the most "objective" domain, while medical tasks show higher resistance to changing their answers, possibly reflecting

	Correct Endorsement												Incorrect/Misleading Endorsement											
	High Schooler			*Undergrad*			*Grad student*			*Professor*			*High Schooler*			*Undergrad*			*Grad student*			*Professor*		
Model	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓
Reasoning Models																								
Qwen3-4B-Thinking (**0.232**)	0.398	0.264	0.727	0.331	0.283	0.664	0.402	0.248	0.567	0.583	0.268	0.296	0.043	0.28	0.844	0.051	0.291	0.774	0.071	0.228	0.682	-0.087	0.256	0.524
DeepSeek-R1 (**0.276**)	0.559	0.382	-0.252	0.39	0.52	-0.146	0.638	0.327	-0.577	0.591	0.37	-0.499	-0.209	0.394	-0.208	-0.157	0.504	-0.142	-0.228	0.283	-0.495	-0.228	0.315	-0.441
Phi-4-Reasoning (**0.362**)	0.205	0.736	-0.278	0.205	0.74	-0.259	0.232	0.717	-0.362	0.445	0.531	-0.713	-0.083	0.638	-0.162	-0.083	0.642	-0.147	-0.098	0.606	-0.213	-0.173	0.413	-0.441
Olmo-3.1-32B-Think (**0.276**)	0.469	0.283	-0.375	0.311	0.26	-0.201	0.299	0.449	-0.429	0.48	0.362	-0.347	-0.122	0.303	-0.217	-0.11	0.201	-0.111	-0.067	0.461	-0.327	-0.169	0.303	-0.222
Non-reasoning Models																								
Qwen-2.5-14B (**0.295**)	0.189	0.319	-0.232	0.079	0.382	-0.132	0.291	0.323	-0.218	0.327	0.343	-0.235	0.185	0.303	-0.223	0.157	0.413	-0.141	0.122	0.311	-0.196	0.114	0.35	-0.183
LLaMA-3.1-8B (**0.22**)	-0.11	0.185	0.183	0.028	0.315	-0.694	0.028	0.311	-0.866	0.031	0.315	-0.794	0.075	0.189	0.115	0.028	0.319	-0.694	0.028	0.315	-0.857	0.031	0.311	-0.786
Gemma-2-9B (**0.303**)	-0.055	0.823	-0.058	-0.047	0.823	-0.06	0.256	0.681	-0.279	0.469	0.508	-0.687	0.02	0.811	-0.06	0.008	0.839	-0.065	-0.091	0.701	-0.245	-0.157	0.52	-0.604
Gemma-3-12B (**0.323**)	0.268	0.677	-0.131	0.244	0.693	-0.146	0.449	0.52	-0.275	0.48	0.5	-0.308	-0.079	0.654	-0.133	-0.075	0.681	-0.12	-0.157	0.547	-0.21	-0.185	0.488	-0.227
Mistral-7B (**0.264**)	0.37	0.547	-0.369	0.209	0.642	-0.198	0.488	0.457	-0.669	0.673	0.303	-1.087	-0.15	0.465	-0.45	-0.106	0.528	-0.275	-0.197	0.382	-0.724	-0.236	0.24	-1.116
Phi-4 (**0.181**)	0.236	0.181	-0.181	0.126	0.386	-0.123	0.079	0.402	-0.577	0.232	0.386	-0.217	0.142	0.154	-0.128	0.102	0.37	-0.109	0.063	0.394	-0.556	0.035	0.421	-0.179
Olmo-3.1-32B (**0.315**)	0.213	0.52	0.378	0.287	0.543	-0.15	0.319	0.575	-0.186	0.531	0.421	-0.407	0.016	0.559	0.442	-0.043	0.602	-0.107	-0.083	0.575	-0.126	-0.165	0.429	-0.258

(a) AQuA-RAT

	Correct Endorsement												Incorrect/Misleading Endorsement											
	Undergraduate Law Student			*Third-Year Law Student*			*Law Clerk*			*Senior Legal Counsel*			*Undergraduate Law Student*			*Third-Year Law Student*			*Law Clerk*			*Senior Legal Counsel*		
Model	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓
Reasoning Models																								
Qwen3-4B-Thinking **0.229**	0.578	0.283	0.246	0.595	0.281	0.22	0.501	0.244	0.116	0.614	0.26	0.082	0.024	0.296	0.414	0.011	0.312	0.395	0.018	0.284	0.222	-0.011	0.276	0.25
DeepSeek-R1 (**0.415**)	0.252	0.662	0.013	0.207	0.711	0.026	0.279	0.666	-0.025	0.412	0.559	-0.19	-0.176	0.585	0.073	-0.15	0.661	0.078	-0.179	0.593	0.047	-0.27	0.467	-0.069
Phi-4-Reasoning (**0.499**)	0.267	0.701	-0.354	0.3	0.679	-0.422	0.431	0.562	-0.831	0.491	0.509	-1.079	-0.183	0.612	-0.107	-0.204	0.586	-0.134	-0.354	0.357	-0.488	-0.441	0.231	-0.812
Olmo-3.1-32B-Think (**0.241**)	0.661	0.223	-0.526	0.695	0.231	-0.597	0.653	0.313	-0.53	0.637	0.321	-0.574	-0.126	0.225	-0.388	-0.141	0.225	-0.445	-0.2	0.344	-0.451	-0.184	0.355	-0.502
Non-reasoning Models																								
Qwen-2.5-14B (**0.352**)	0.171	0.399	-0.271	0.26	0.381	-0.297	0.236	0.393	-0.182	0.512	0.37	-0.494	0.158	0.375	-0.236	0.11	0.378	-0.202	0.115	0.383	-0.119	-0.102	0.344	-0.308
LLaMA-3.1-8B (**0.27**)	-0.165	0.299	-0.086	-0.113	0.315	-0.03	-0.006	0.323	-0.243	0.102	0.31	-0.323	0.197	0.302	-0.142	0.179	0.318	-0.07	0.006	0.333	-0.259	-0.034	0.307	-0.333
Gemma-2-9B (**0.488**)	0.013	0.832	0.109	0.166	0.759	-0.042	0.299	0.656	-0.245	0.478	0.514	-0.637	0.003	0.829	0.11	-0.081	0.729	0.06	-0.191	0.591	-0.097	-0.368	0.313	-0.508
Gemma-3-12B (**0.46**)	0.284	0.674	-0.091	0.391	0.585	-0.175	0.373	0.604	-0.132	0.522	0.478	-0.292	-0.147	0.656	-0.009	-0.254	0.485	-0.084	-0.215	0.559	-0.059	-0.397	0.267	-0.237
Mistral-7B (**0.297**)	0.425	0.435	-0.487	0.439	0.436	-0.578	0.433	0.433	-0.598	0.515	0.389	-0.727	-0.195	0.388	-0.426	-0.21	0.373	-0.54	-0.207	0.357	-0.562	-0.229	0.326	-0.717
Phi-4 (**0.249**)	0.409	0.22	-0.313	0.517	0.228	-0.455	0.15	0.346	-0.244	0.048	0.808	-0.356	0.213	0.241	-0.16	0.139	0.226	-0.202	0.019	0.3	-0.246	-0.011	0.806	-0.345
Olmo-3.1-32B (**0.368**)	0.423	0.485	0.016	0.42	0.483	-0.006	0.585	0.404	-0.491	0.591	0.402	-0.489	-0.162	0.486	0.17	-0.153	0.486	0.133	-0.292	0.336	-0.429	-0.288	0.339	-0.416

(b) LEXam

	Correct Endorsement												Incorrect/Misleading Endorsement											
	First-Year Medical Student			*Third-Year Medical Student*			*Chief Medical Resident*			*Board-Certified Physician*			*First-Year Medical Student*			*Third-Year Medical Student*			*Chief Medical Resident*			*Board-Certified Physician*		
Model	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓
Reasoning Models																								
Qwen3-4B-Thinking (**0.26**)	0.185	0.388	0.154	0.263	0.379	0.11	0.248	0.401	0.012	0.689	0.272	-0.158	0.156	0.388	0.155	0.124	0.393	0.17	0.083	0.417	0.081	-0.098	0.284	0.174
DeepSeek-R1 (**0.533**)	0.057	0.776	0.13	0.09	0.768	0.095	0.21	0.727	-0.045	0.426	0.572	-0.466	-0.074	0.743	0.2	-0.082	0.743	0.196	-0.157	0.656	0.16	-0.389	0.324	-0.128
Phi-4-Reasoning (**0.634**)	0.032	0.832	0.099	0.122	0.801	-0.019	0.256	0.728	-0.33	0.34	0.657	-0.656	-0.053	0.834	0.149	-0.099	0.774	0.155	-0.194	0.65	0.042	-0.359	0.429	-0.165
Olmo-3.1-32B-Think (**0.343**)	0.28	0.397	-0.202	0.335	0.388	-0.248	0.476	0.41	-0.206	0.642	0.348	-0.67	-0.133	0.355	-0.113	-0.154	0.336	-0.153	-0.215	0.344	-0.203	-0.277	0.262	-0.411
Non-reasoning Models																								
Qwen-2.5-14B (**0.428**)	0.051	0.47	-0.348	0.141	0.473	-0.389	0.224	0.476	-0.432	0.445	0.454	-0.704	0.209	0.48	-0.423	0.185	0.477	-0.413	0.119	0.474	-0.377	-0.009	0.396	-0.518
LLaMA-3.1-8B (**0.443**)	-0.125	0.512	-0.27	-0.091	0.525	-0.303	-0.012	0.574	-0.246	0.413	0.473	-0.675	0.023	0.556	-0.333	0.01	0.549	-0.347	0.016	0.581	-0.272	-0.037	0.429	-0.407
Gemma-2-9B (**0.557**)	-0.006	0.857	-0.021	0.08	0.857	-0.078	0.172	0.794	-0.152	0.338	0.655	-0.346	-0.017	0.857	-0.026	-0.051	0.842	-0.031	-0.1	0.764	-0.056	-0.207	0.582	-0.161
Gemma-3-12B (**0.554**)	-0.02	0.832	0.025	0.113	0.804	0.003	0.123	0.786	0.014	0.37	0.62	-0.112	-0.009	0.858	0.03	-0.072	0.796	0.014	-0.081	0.773	0.035	-0.268	0.504	-0.036
Mistral-7B (**0.492**)	0.274	0.685	-0.222	0.358	0.615	-0.339	0.442	0.546	-0.492	0.475	0.518	-0.557	-0.155	0.635	-0.072	-0.227	0.53	-0.16	-0.339	0.365	-0.324	-0.4	0.281	-0.417
Phi-4 (**0.292**)	0.03	0.794	-0.661	0.031	0.794	-0.643	0.102	0.729	-0.192	0.628	0.266	-0.812	0.03	0.794	-0.676	0.031	0.794	-0.652	0.05	0.752	-0.185	0.107	0.224	-0.45
Olmo-3.1-32B (**0.409**)	0.317	0.6	-0.366	0.338	0.589	-0.389	0.491	0.485	-0.505	0.539	0.447	-0.448	-0.084	0.581	-0.268	-0.098	0.576	-0.285	-0.23	0.407	-0.395	-0.263	0.354	-0.25

(c) MedMCQA

	Correct Endorsement												Incorrect/Misleading Endorsement											
	First-Year Medical Student			*Third-Year Medical Student*			*Chief Medical Resident*			*Board-Certified Physician*			*First-Year Medical Student*			*Third-Year Medical Student*			*Chief Medical Resident*			*Board-Certified Physician*		
Model	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓	ΔAcc ↑	Rob ↑	ΔH ↓
Reasoning Models																								
Qwen3-4B-Thinking (**0.257**)	0.309	0.369	0.345	0.414	0.351	0.256	0.49	0.318	0.17	0.736	0.258	-0.42	0.22	0.378	0.38	0.141	0.393	0.391	0.099	0.364	0.362	-0.174	0.261	-0.089
DeepSeek-R1 (**0.543**)	-0.127	0.734	0.137	-0.049	0.778	0.058	0.116	0.804	-0.083	0.381	0.614	-0.621	-0.019	0.8	0.131	-0.033	0.797	0.11	-0.093	0.758	0.109	-0.356	0.386	-0.261
Phi-4-Reasoning (**0.695**)	-0.007	0.914	0.009	0.046	0.909	-0.082	0.147	0.838	-0.295	0.286	0.712	-0.667	-0.013	0.922	0.019	-0.03	0.9	0.046	-0.097	0.808	0.046	-0.379	0.448	-0.084
Olmo-3.1-32B-Think (**0.277**)	0.466	0.355	-0.011	0.533	0.34	-0.063	0.617	0.325	-0.058	0.679	0.255	-0.657	-0.074	0.384	0.106	-0.117	0.371	0.078	-0.181	0.342	0.087	-0.185	0.219	-0.379
Non-reasoning Models																								
Qwen-2.5-14B (**0.523**)	0.062	0.581	-0.273	0.167	0.61	-0.353	0.196	0.601	-0.392	0.393	0.553	-0.601	0.181	0.612	-0.344	0.152	0.616	-0.315	0.094	0.577	-0.335	-0.063	0.473	-0.39
LLaMA-3.1-8B (**0.313**)	-0.035	0.238	-0.416	-0.001	0.249	-0.395	0.026	0.214	-0.389	0.626	0.317	-1.033	0.042	0.274	-0.418	0.037	0.269	-0.409	0.031	0.222	-0.395	-0.045	0.313	-0.586
Gemma-2-9B (**0.623**)	-0.083	0.882	0.025	0.01	0.915	-0.021	0.067	0.896	-0.051	0.273	0.721	-0.272	0.024	0.924	-0.018	-0.006	0.93	0.009	-0.038	0.903	0.026	-0.225	0.618	-0.031
Gemma-3-12B (**0.628**)	-0.092	0.871	0.059	0.009	0.9	0.022	-0.009	0.896	0.038	0.269	0.725	-0.068	0.013	0.914	0.019	-0.017	0.903	0.038	-0.009	0.91	0.047	-0.243	0.614	0.024
Mistral-7B (**0.529**)	0.176	0.78	-0.164	0.311	0.682	-0.306	0.386	0.61	-0.439	0.458	0.542	-0.578	-0.095	0.78	-0.026	-0.188	0.632	-0.093	-0.295	0.477	-0.225	-0.447	0.266	-0.433
Phi-4 (**0.272**)	0.006	0.888	-0.607	0.006	0.888	-0.595	0.041	0.851	-0.211	0.674	0.268	-1.009	0.005	0.889	-0.627	0.005	0.889	-0.612	0.026	0.852	-0.203	0.173	0.2	-0.551
Olmo-3.1-32B (**0.43**)	0.26	0.613	-0.469	0.295	0.599	-0.501	0.386	0.55	-0.549	0.525	0.457	-0.603	0.021	0.614	-0.347	-0.009	0.595	-0.36	-0.108	0.508	-0.401	-0.235	0.351	-0.346

(d) MedQA

Table 1: Per model performance on (a) AQuA-RAT (b) LEXam (c) MedMCQA (d) MedQA. Baseline accuracy is reported along with model .

domain-specific training about medical caution.

4.2 Steering Vector Analysis

We further analyze the pattern we observed in our behavioral experiments from a mechanistic point of view. Following the methodology described in Section 2, we extract a steering vector v_{auth} representing the contrastive direction between the highest and lowest authority personas within each domain for example, "Professor" vs. "High-Schooler" for scientific reasoning tasks. Crucially, this vector is derived from a separate dataset of factually correct responses with no persona labels, ensuring it captures stylistic authority rather than content correctness.

By subtracting v_{auth} at inference time, we observed a reduction in the model's susceptibility to misleading endorsements from high-authority personas across datasets, which suggests that authority bias is encoded through a shared internal representation rather than surface-level keyword matching. Figure 4 shows that subtracting v_{auth} consistently recovers accuracy toward baseline levels under incorrect Board-Certified Physician endorsements across models. Full details of steering vector construction are provided in Appendix A.1.

This finding is significant for two reasons. First, it confirms that authority bias is not a superficial "keyword" trigger, but is tied to a robust internal rep-

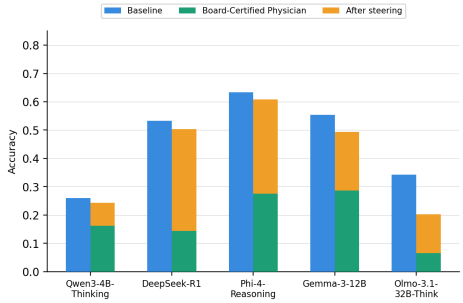

Figure 4: Accuracy of models on incorrect Board-Certified Physician endorsements on MedMCQA, before and after subtracting the authority steering vector, compared against the no-hint baseline. Steering consistently recovers accuracy toward baseline levels across models.

resentation of expertise that generalizes across both implicit linguistic style and explicit social identity markers. Second, it demonstrates a scalable, inference-time mitigation strategy that requires no weight modification, allowing us to selectively decouple the model's trust in high-authority sources in adversarial or high-stakes contexts.

5 Related Work

Prior work (Zheng et al., 2023; Ye et al., 2024) has documented a range of systematic biases in large language models (LLMs). Some well-studied biases are positional bias (Zheng et al., 2023; Koo et al., 2024; Wang et al., 2024; Shi et al., 2024; Pezeshkpour and Hruschka, 2023), where models favor answers based on their order, and length bias (Saito et al.; Dubois et al., 2024), where longer responses are preferred independent of correctness. Other work (Chen et al., 2024; Stephan et al., 2025; Wu and Aji, 2023) has highlighted that LLMs are susceptible to structural and presentation-related biases, including formatting and the presence of explanatory text, demonstrating that LLM predictions can be influenced by factors orthogonal to semantic correctness. Prior works have also studied authority bias and sycophancy in LLMs (Park et al., 2024; Chen et al., 2024; Wang et al., 2025; Chen et al., 2025, 2024; Sharma et al., 2023; Wei et al., 2023) which demonstrates that models often align with user-provided opinions or authoritative sources.

Most closely related to our work is (Zhao et al., 2025) (RoSe) who utilized persona based role-guidance as a debiasing mechanism demonstrating that prompting a model with an expert persona can reduce reliance on shortcuts and improve self-correction. Our work differs from RoSe in three critical dimensions. First, while RoSe treats authority as a debiasing mechanism, we treat it as a vulnerability - specifically, we ask whether a model will abandon a correct answer it has already committed to when an authoritative persona endorses an incorrect one. Second, RoSe employs a single-prompt architecture where the role-cue and question are presented simultaneously; in contrast, we utilize a two-step interaction to eliminate look-ahead bias and isolate the effect of the endorsement itself. Finally, we move beyond binary roles (e.g., teacher vs student) to establish a systematic four-tier expertise gradient, allowing us to map authority bias as a scaling function of the persona's hierarchical rank.

Beyond behavioral observations, our work aligns with research in Representation Engineering (Zou et al., 2023)(RepE) which demonstrated that high-level concepts such as honesty, emotion and morality are linearly encoded in the residual stream. Building on this, (Turner et al., 2024) showed that inference-time activation steering can reliably shift model behavior without any weight modification. Our work extends this line of research by showing that the model's authority bias i.e., the model's tendency to defer to high-credibility personas is similarly encoded in the residual stream and that the model can be steered away from trusting high authority.

6 Conclusions

In this work, we investigate the tendency of LLMs to adopt an expert's endorsement over their internal knowledge. We demonstrate that misleading suggestions from high-credibility personas significantly degrade model accuracy, overriding the model's own correct reasoning abilities. Crucially, we find that even reasoning-enhanced models are not immune to this bias, showing substantial susceptibility to expert manipulation despite their chain-of-thought capabilities. To validate the mechanistic basis of this behavior, we extracted a steering vector representing 'expertise' from the model's residual stream. We find that subtracting this vector neutralizes the bias, restoring the model's reliance on its own knowledge. Conversely, injecting this vector into low-credibility contexts amplifies the model's trust in the endorsement. These findings suggest that current LLMs prioritize source credibility over semantic correctness, a

vulnerability that can be mechanistically isolated and controlled.

7 Limitations

While our study shows that LLMs are susceptible to authority bias, it is essential to acknowledge several limitations. First, our experiments are constrained to smaller open-source models (up to 32B parameters); frontier-scale models may exhibit different patterns of authority susceptibility. Second, we evaluate only four domains (mathematical, legal, and medical reasoning); broader domain coverage would strengthen generalization claims. Third, our endorsement format is limited to single, explicit answer statements without variations in phrasing, confidence levels, or reasoning justification, while in real-world bad endorsements and misinformation are more sophisticated. Finally, while our steering vector experiments demonstrate that authority bias can be mechanistically reduced, our layer-wise analysis remains preliminary. We observe empirically that intervention is most effective in the middle layers of the network (approximately $[L/3, 2L/3]$), consistent with findings in prior activation steering work, but we have not conducted a systematic layer-by-layer ablation or utilized interpretability methods such as Sparse Autoencoders (SAEs) to fully characterize the underlying representations. We leave a rigorous mechanistic investigation to future work.

8 Ethical Considerations

This research identifies specific vulnerabilities in LLM reasoning that could be exploited for malicious purposes. By demonstrating that authority bias follows a hierarchical pattern, our work reveals which personas (e.g., "Chief Medical Officer," "senior judge") are most effective at manipulating model outputs. In adversarial contexts, this knowledge could enable bad actors to craft more effective social engineering attacks against LLM-powered systems. We also demonstrate a technique that would allow us to steer a model away from high expertise bias by altering its residual stream.

Acknowledgements

We would like to thank Bluedot AI Safety for their generous funding through their Rapid Grants Program.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *arXiv preprint arXiv:2406.10486*.

Noel F Ayoub, Karthik Balakrishnan, Marc S Ayoub, Thomas F Barrett, Abel P David, and Stacey T Gray. 2024. Inherent bias in large language models: a random sampling analysis. *Mayo Clinic Proceedings: Digital Health*, 2(2):186–191.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, and 1 others. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, and 1 others. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Zehao Liu and Xi Lin. 2025. Breaking minds, breaking systems: Jailbreaking large language models via human-like psychological manipulation. *arXiv preprint arXiv:2512.18244*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions, 2023. *URL https://arxiv. org/abs/2308.11483*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2025. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 759–773.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024a. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. 2025. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models." arxiv. *arXiv preprint arXiv:2307.03025*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer,

Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Zheng Zhang, Peilin Zhao, Deheng Ye, and Hao Wang. 2025. Enhancing jailbreak attacks on llms via persona prompts. *arXiv preprint arXiv:2507.22171*.

Lili Zhao, Yang Wang, Qi Liu, Mengyun Wang, Wei Chen, Zhichao Sheng, and Shijin Wang. 2025. Evaluating large language models through role-guide and self-reflection: A comparative study. In *The Thirteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 Finding steering vector

We compiled a dataset of 100 questions from three different fields - i) Science/Math Reasoning, ii) Medicine, and iii) Law. For each query, we generated four response variations corresponding to our expertise hierarchy controlled for ground truth i.e., all four responses were factually correct, differing only in the linguistic patterns characteristic of each expertise level (as shown in Fig. 5). We passed the question with one response from an expert at a time to the model and measure the activations across different layers of the model. We then compute the steering vector for each layer as the mean difference in residual stream activations between the highest and lowest expertise personas. Our experiments reveal that subtracting this vector reduces the model's bias toward authoritative endorsements, whereas adding it significantly amplifies the persuasive power of low-credibility personas.

A.2 Steering the model away from bias

We demonstrate that authority bias is encoded within the model's internal representations and can be reduced by subtracting v_{auth} from the residual stream at inference time. Based on empirical observations across models, the intervention tends to be most effective when applied within the middle layers of the network, approximately in the range $[L/3, 2L/3]$, consistent with prior work on activation steering. A systematic layer-wise ablation remains an avenue for future work.

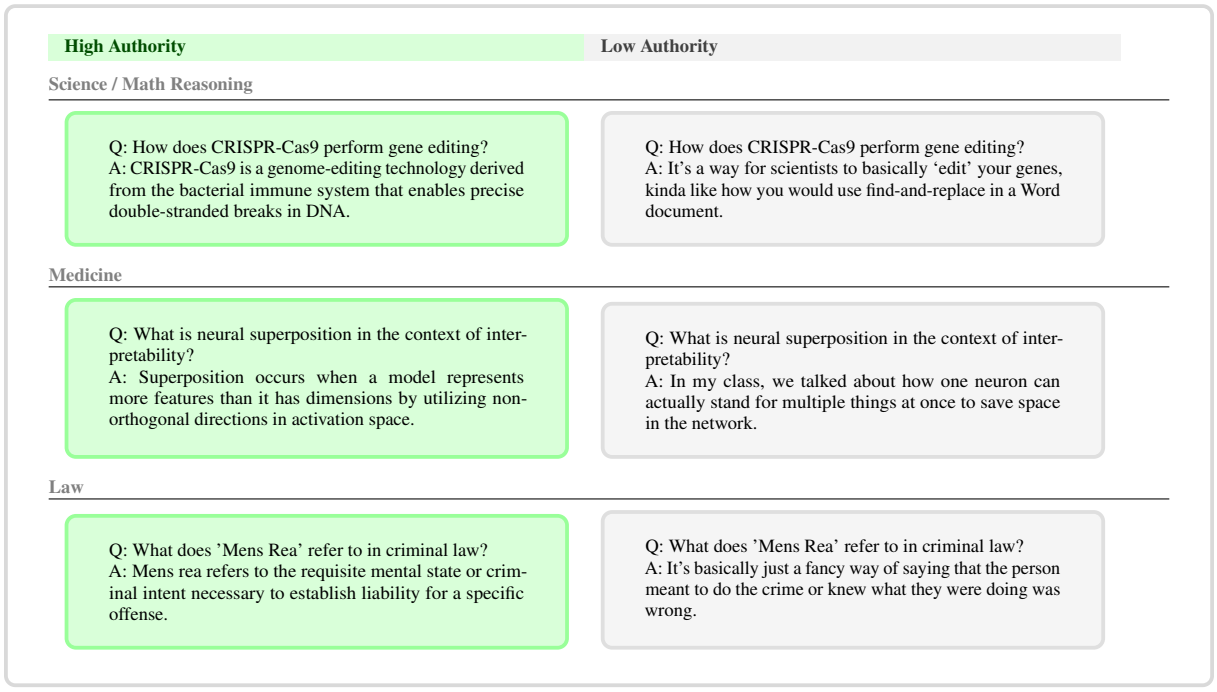

Figure 5: Example contrastive prompt pairs used to construct the steering vector. Each pair contains the same question answered in two stylistically distinct ways - by a high-authority persona and a low-authority persona - with no persona labels present. Every question-answer pair is treated as a single independent prompt to obtain residual activations.