Quantifying and Improving the Robustness of Retrieval-Augmented Language Models Against Spurious Features in Grounding Data

Work done during an internship at Microsoft

Shiping Yang[1,2], Jie Wu[3], Wenbiao Ding[2], Ning Wu[2], Shining Liang[2], Ming Gong[3], Hongzhi Li[4], Hengyuan Zhang[5]
Angel X. Chang[1,6], Dongmei Zhang[2]

[1] Simon Fraser University · [2] Microsoft · [3] Atlassian · [4] Tongji University · [5] The University of Hong Kong · [6] Canada-CIFAR AI Chair, Amii

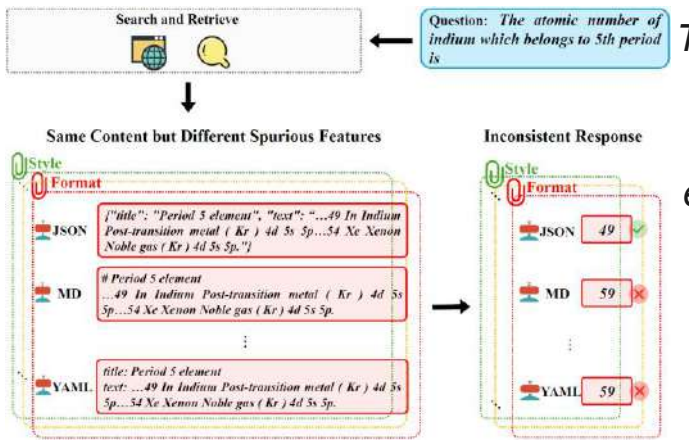
ACL 2026 SAN DIEGO JULY 2-7

Background and Motivation

RAG mitigates hallucinations by grounding answers in retrieved documents. Prior work has shown that RALMs are sensitive to irrelevant or counterfactual grounding data — a well-known RAG robustness issue. However, these studies all examine explicit noise that alters semantics.

In practice, the Internet is saturated with documents that share the same meaning while differing in style, format, source, ordering, or metadata. We instead study implicit noise: spurious features, i.e., semantic-agnostic perturbations, and ask whether they likewise affect RALM responses.

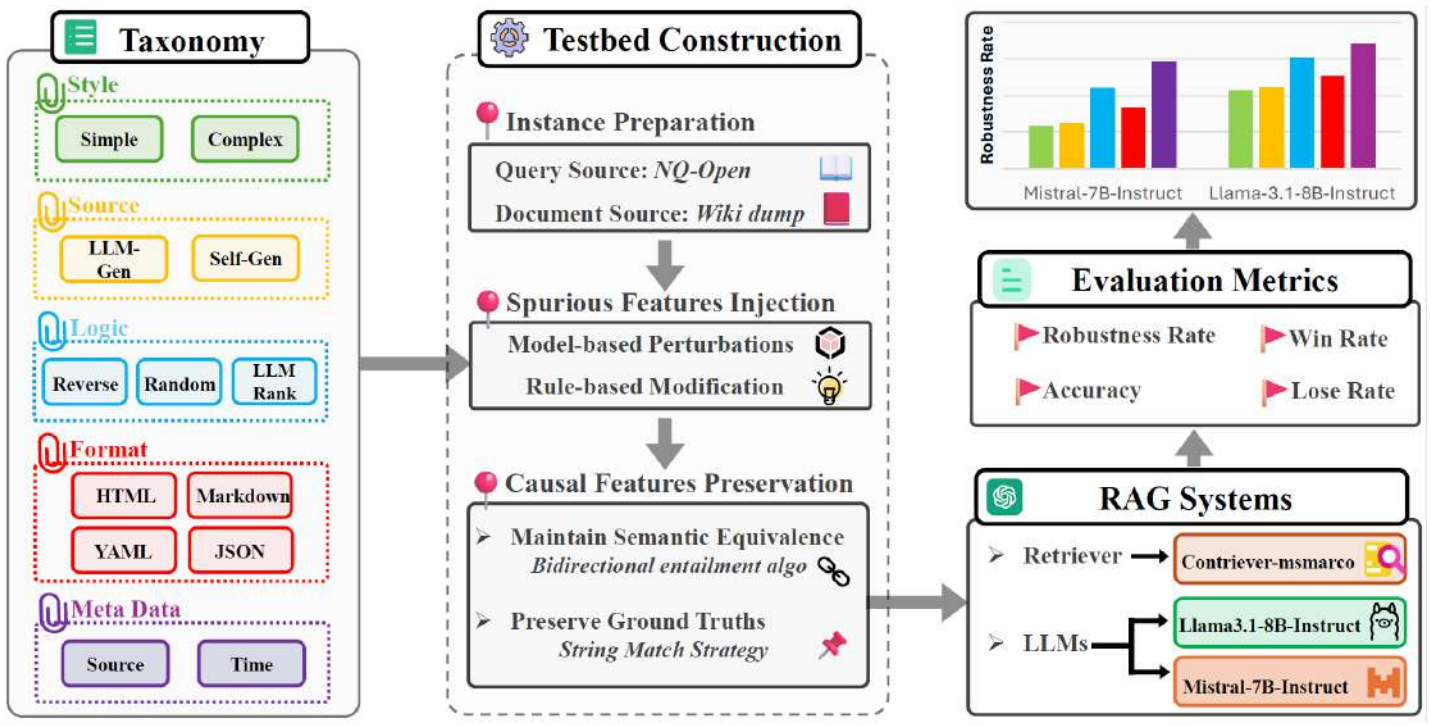
The Phenomenon: Sensitive to Spurious Features



The same retrieved document, when presented to the LLM in different formats, elicits inconsistent responses.

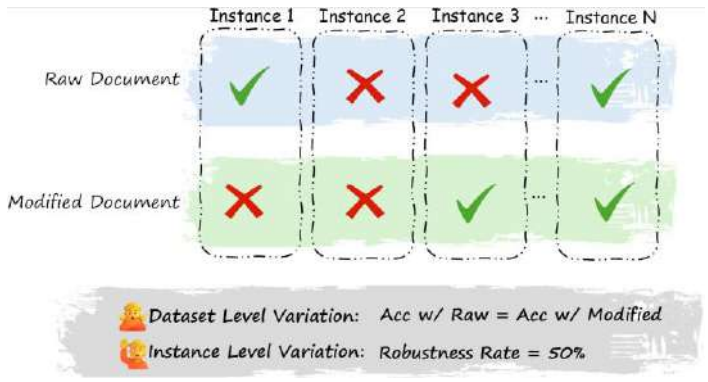
Correctness Flips!!

SURE Framework: Perturb then Evaluate



How to Quantify Robustness

Dataset-level metrics hides the Sensitivity



Robustness Rate	Win Rate (WR)	Lose Rate (LR)
Prediction unchanged	Incorrect to correct	Correct to incorrect

Main Results

Llama-3.1-8B-Instruct

Taxonomy	Perturbations	Known-Golden LR	Known-Golden RR	Known-Golden WR	Known-Golden Org	Known-Golden Acc	Known-Noise LR	Known-Noise RR	Known-Noise WR	Known-Noise Org	Known-Noise Acc	Unknown-Golden LR	Unknown-Golden RR	Unknown-Golden WR	Unknown-Golden Org	Unknown-Golden Acc	U-N RR
Style	Simple	7.79	83.04	**9.18**	66.03	67.42	1.70	95.80	**2.50**	4.12	4.92	8.43	82.88	**8.69**	51.42	51.68	99.45
	Complex	6.00	85.60	**8.40**		68.43	1.91	96.59	1.50		3.71	6.71	84.86	**8.43**		53.13	99.57
Source	LLM-Generated	5.89	86.43	**7.69**	65.62	67.43	1.43	96.83	**1.74**	4.13	4.45	6.20	85.71	**8.09**	49.15	51.04	99.56
	Self-Generated	6.55	85.01	**8.44**		67.52	1.55	96.37	**2.09**		4.67	6.52	86.36	**7.12**		49.74	99.57
Logic	Reverse	5.06	90.82	4.12	62.95	62.01	1.13	97.82	1.06	4.43	4.36	5.73	89.71	4.56	45.84	44.66	99.67
	Random	3.91	93.16	2.93		61.97	0.86	98.31	0.83		4.40	4.21	91.67	4.12		45.74	99.72
	LLM-Ranked	3.24	93.93	2.83		62.54	0.82	98.43	0.74		4.36	3.58	93.36	3.06		45.32	99.76
Format	JSON	7.01	88.25	4.74	63.91	61.64	1.70	97.25	1.05	3.87	3.21	5.92	89.63	4.45	49.35	47.88	99.61
	HTML	11.85	84.46	3.69		55.75	2.70	96.90	0.40		1.56	9.33	86.78	3.90		43.92	99.61
	YAML	5.26	89.94	4.80		63.45	1.26	97.41	1.33		3.94	4.79	90.80	4.41		48.97	99.67
	Markdown	2.32	92.23	**5.45**		67.04	0.60	96.89	**2.51**		5.77	2.34	93.46	**4.19**		51.20	99.61
Metadata	Timestamp (pre)	2.08	95.81	**2.11**	55.77	55.80	0.28	99.42	**0.29**	1.58	1.59	2.54	95.56	1.90	43.31	42.66	99.95
	Timestamp (post)	2.04	95.86	**2.10**		55.84	0.25	99.43	**0.32**		1.64	2.81	95.56	1.63		42.12	99.95
	Datasource (wiki)	2.11	93.45	**4.44**		58.10	0.23	98.96	**0.81**		2.17	3.25	92.47	**4.27**		44.33	99.86
	Datasource (twitter)	2.27	94.11	**3.62**		57.11	0.31	99.25	**0.43**		1.70	2.77	93.97	**3.25**		43.79	99.91

How to Improve Robustness

PER UNROBUST INSTANCE

Q query · ✓ correct ans · ✗ incorrect ans · G Golden Doc · g(G) Perturbed Doc

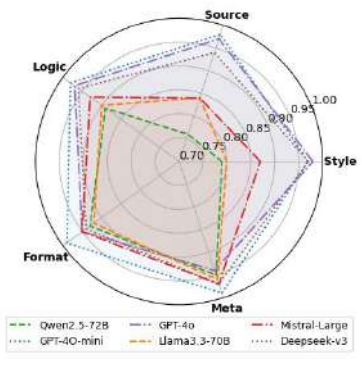
▼ Pair Q with G and g(G) ▼

SFT
Train model to produce ✓ on both G and g(G) .
2 training samples per instance

DPO
Prefer ✓ > ✗ under both G and g(G) .
2 preference pairs per instance

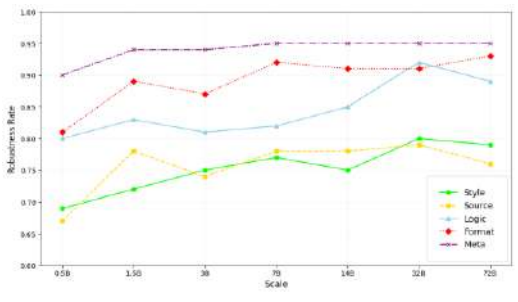
METHOD	STYLE	SOURCE	LOGIC	FORMAT	META
Llama-3.1-8B-It	10.0	15.5	20.0	24.0	94.0
+ SFT	**96.5**	**94.5**	**99.0**	**99.5**	**99.7**
+ DPO	**96.5**	**96.0**	**96.0**	**98.0**	**98.0**

Further Analysis



TAKEAWAY 1

Spurious features are a widespread robustness gap in RAG — independent of internal knowledge, architecture, and scale.



TAKEAWAY 2

Not every spurious feature is harmful — **WR > LR** on some perturbations shows certain features can even help.

TAKEAWAY 3

Traditional strategies for improving robustness against explicit noise do not work for spurious features; training on SURE-synthesized pairs (SFT / DPO) effectively mitigates this sensitivity.

Contact Information

Shiping Yang
Email: yangshipingnlp@gmail.com

PAPER · CODE · Personal Webpage