ACL 2026 · SAN DIEGO

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

Presenter: Shiping Yang

THE PHENOMENON

Same content, different format — inconsistent response

Q “What is the atomic number of indium?” *+ the Wikipedia paragraph that contains the correct answer.*

Document formatted as JSON

Identical content. Identical meaning.

Model answers **49**

CORRECT

Document formatted as YAML

Identical content. Identical meaning.

Model answers **59**

WRONG

Correctness flips ! But the semantics never changed.

Research Gap

Prior work studied the RAG Robustness on explicit noise

STUDIED BEFORE

Explicit noise

Semantics is altered

Irrelevant or counterfactual documents — the well-known RAG robustness issue.

THIS PAPER

Spurious features (implicit noise)

Semantics is preserved

differ in Style, format, source, ordering, metadata.

Do these semantic-agnostic differences change the model's answer?

How to Quantify Robustness

Dataset-level metrics hides the instance-level variation

Per-instance, predictions change — but the flips cancel out in the average:

Accuracy(raw) ≈ Accuracy(perturbed) → **Looks robust. The Sensitivity is invisible.**

Robustness Rate

How often the prediction stays consistent.

Win Rate

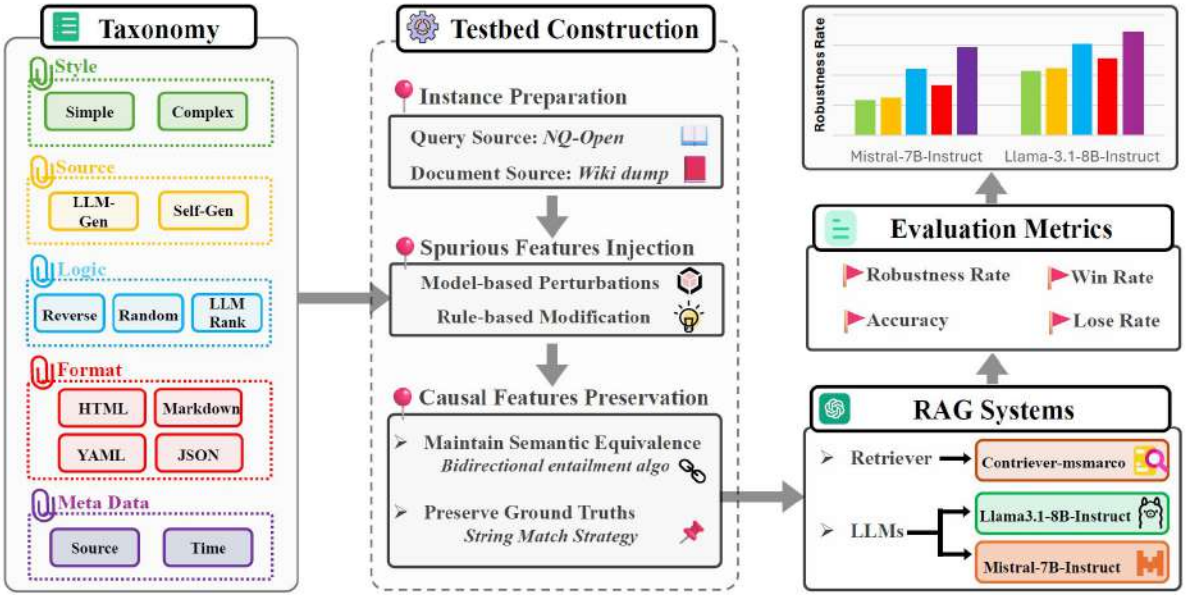
Perturbation flips a wrong answer to right.

Lose Rate

Perturbation flips a correct answer to wrong.

How to Evaluate Robustness

SURE Framework: Spurious featUres Robustness Evaluation



1 Inject

model-based & rule-based perturbations.

2 Preserve

Bidirectional entailment confirms equivalence; string-match confirms the answer survives.

3 Evaluate

Feed to RAG systems and measure RR / WR / LR at the instance level.

Takeaway

It's everywhere — but not all of it is bad

Finding 1

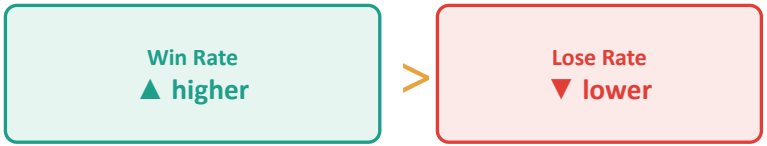
A widespread robustness gap

Sensitivity to spurious features holds **regardless of model scale, architecture, or whether the model already knows the answer internally.**

It is a fundamental gap — not a flaw of one model.

Finding 2

Not every feature is harmful



For some perturbations, Win Rate exceeds Lose Rate.

Further Analysis

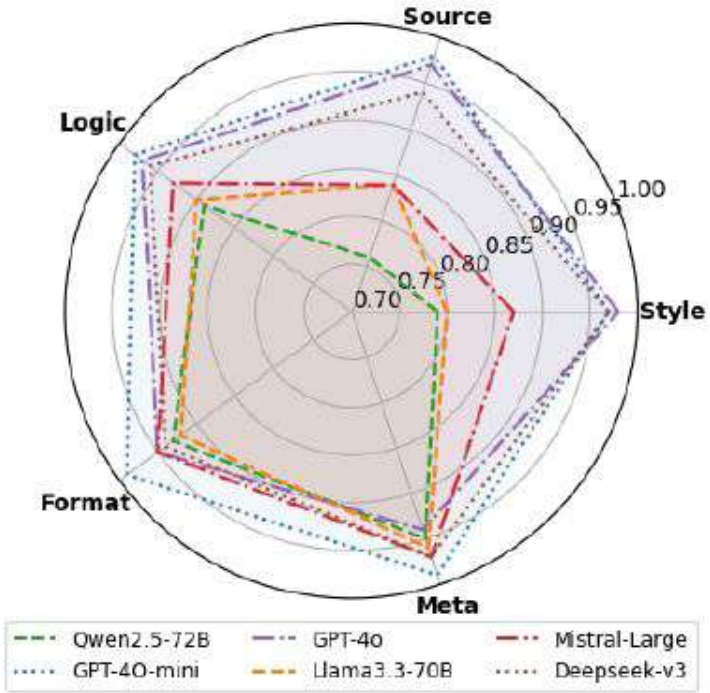


Figure 4: Robustness comparison of six SOTA LLMs.

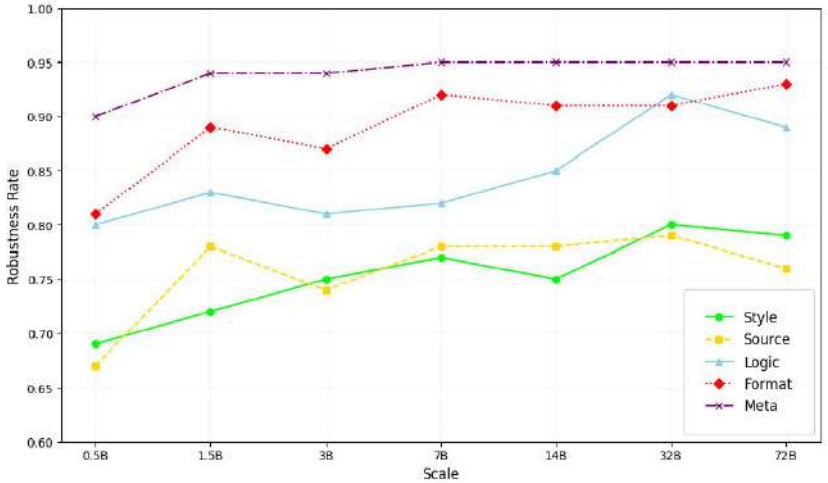


Figure 5: Scaling analysis on Qwen2.5 series.

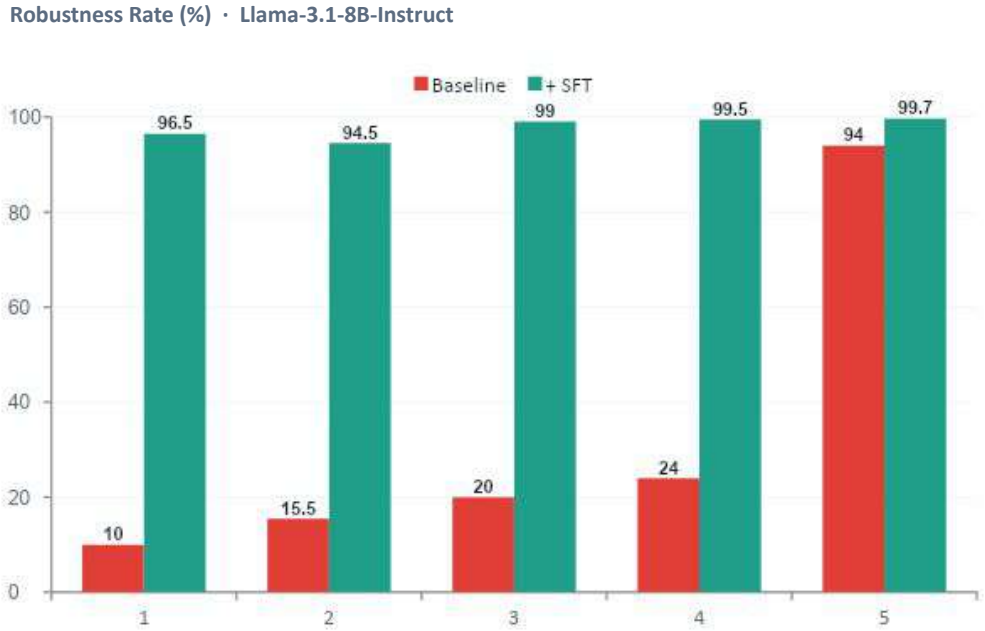
How to Improve Robustness

Train on synthetic data — robustness jumps to ~95%+

The recipe

For each unrobust instance, pair the original doc with its perturbation, and train — via **SFT or DPO** — to answer correctly under both.

	Style	Source	Logic	Format	Meta
Llama3.1-8B (Wiki)	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
Llama3.1-8B (Trivial)	87.5	93.5	93.0	90.8	97.0
+ SFT	88.5	91.5	95.0	**96.3**	**99.0**
+ DPO	**94.5**	**94.5**	**97.3**	95.8	98.0



TAKEAWAY

RAG models are sensitive not just to what a document says, but to how it's written.

- **Standard metrics hide it** — instance-level RR / WR / LR reveal it.
- **SURE quantifies it** — a perturb-then-evaluate framework.
- **Training on SURE generated samples fixes it** — SFT / DPO improve robustness to ~95%+.

yangshipingnlp@gmail.com

Thank you!