Hallucination Reduction in Large Language Models: A Multi-Step Framework for Detection and Correction

Master Thesis by Cristian Longoni
Supervised by Prof. Sivio Gerli
Co-Supervised by Dr. Teresa Cigna

What is a Hallucination?

- LLMs generate fluent but sometimes *unreliable* text
- *Hallucinations*: fabricated or unsupported information
- Originate from Data, Training or Inference
- High-risk in **medicine, law, journalism**
- Need for **factual reliability** and **source traceability**

Objective of the research

- Detect both ***factual*** and ***faithfulness*** hallucinations
- **Automatically** correct or remove false claims
- Ensure every sentence is backed by at least one source
- Evaluate performance through quantitative metrics

Factual
• The construction of the Eiffel Tower led to the extinction of the “Parisian tiger" • Thomas Edison invented the telephone

Faithful
• **Translate:** “What is the capital of France?” Response: “The capital of France is Paris,” • **Source:** "Nile originates from the Great Lakes region" Model:"It originates from mountain ranges" • 2x +3 = 11 x = 3

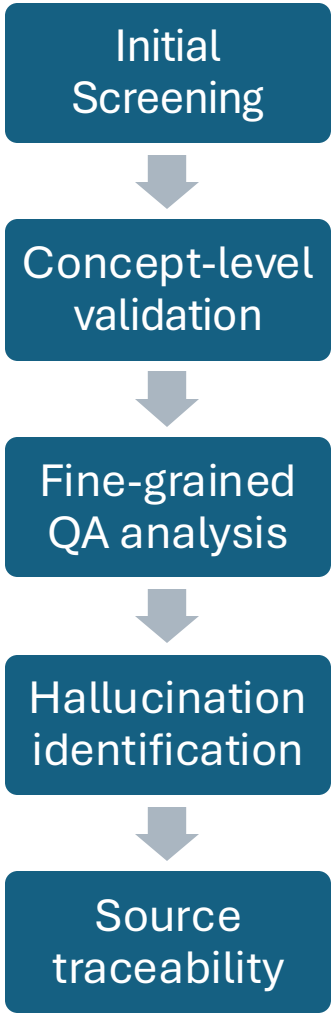
Related Works

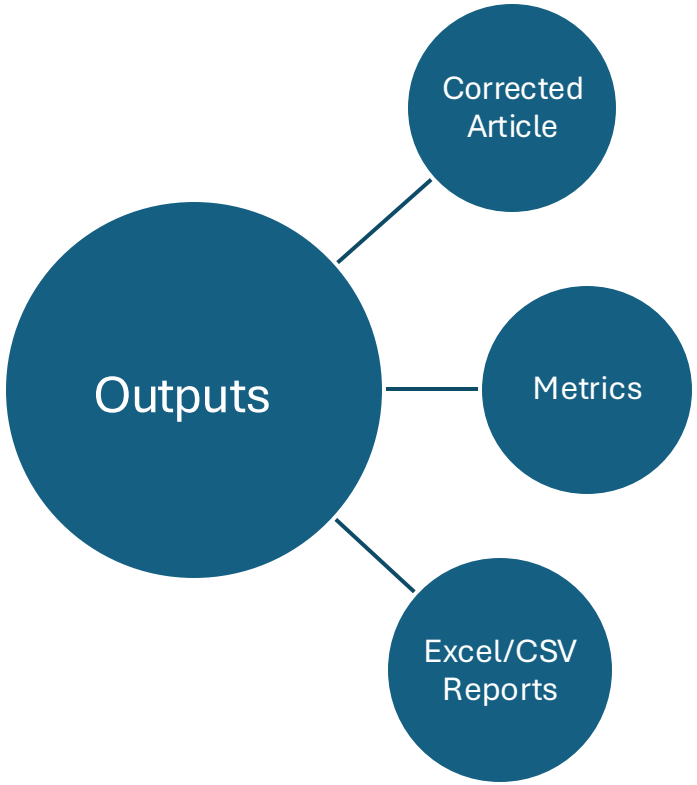
- **Previously:**
  - Hallucination *Detection*

  OR
  - Hallucination *Mitigation*

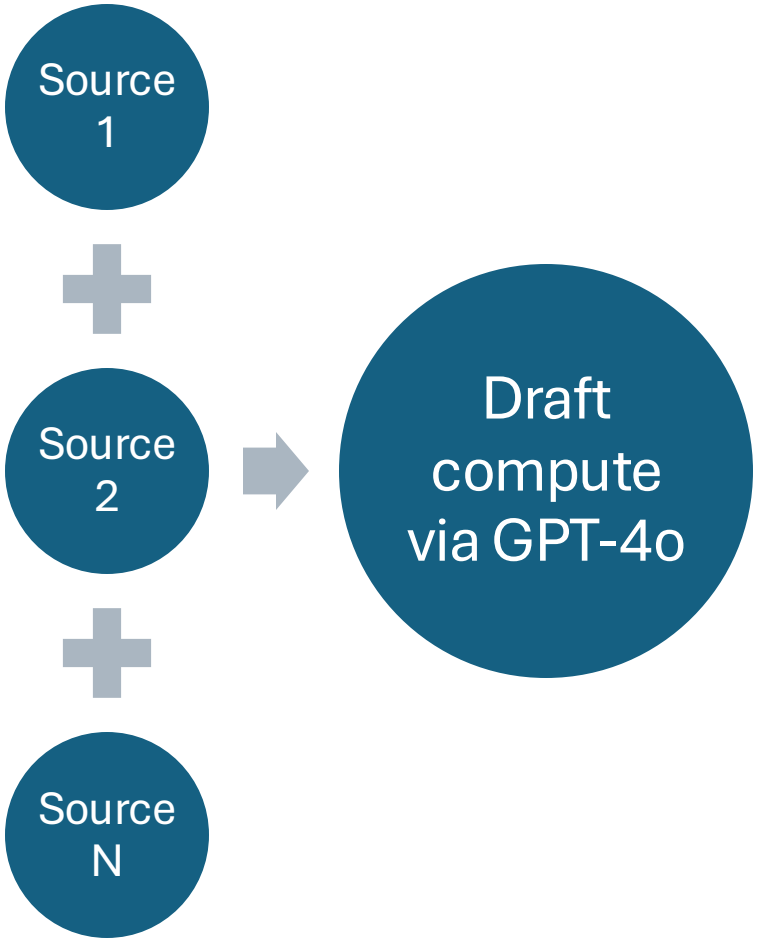
- **Positioning of this Work**
  - Extends COV-style reasoning into a **five-step modular pipeline**.
  - Integrates **automatic correction**, not only detection.
  - Combines **quantitative metrics + qualitative evaluation** for factual reliability.

Framework Overview





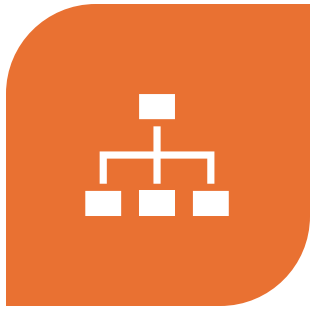
Experimental Setup



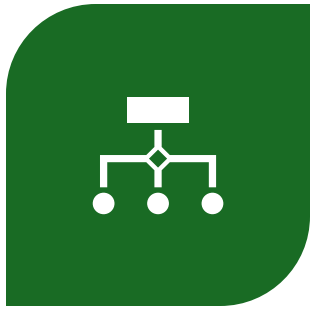
Models via **Ollama** for Framework:

- **gemma2:9b** (Primary)
- **llama3.1:8b** (Secondary)

Prompt Engineering



USE OF *STRUCTURED, ROLE-BASED PROMPTS* (E.G., "YOU ARE A FACT-CHECKER...")



STEP-SPECIFIC TEMPLATES FOR CORRECTION, QA, AND VERIFICATION

COMBINATION OF *INSTRUCTION TUNING* AND *FEW-SHOT EXAMPLES*

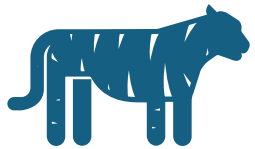
CONTROLLED TEMPERATURE AND DETERMINISTIC RESPONSES

Initial Screening (Zero-Check)

Compare each sentence with all sources

Remove or rewrite unsupported claims

Output: cleaned article



Kept

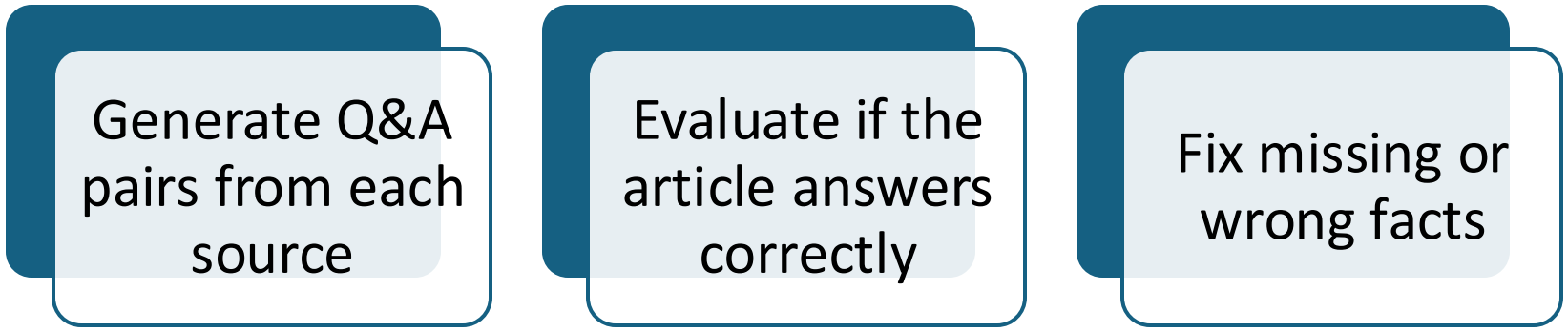
Article says “the concert had 800 attendees”

Source reports “100 attendees”

Removed

A Parisian tiger extinct due to the Eiffel Tower’s construction

Concept-Level QA Validation (First-check)



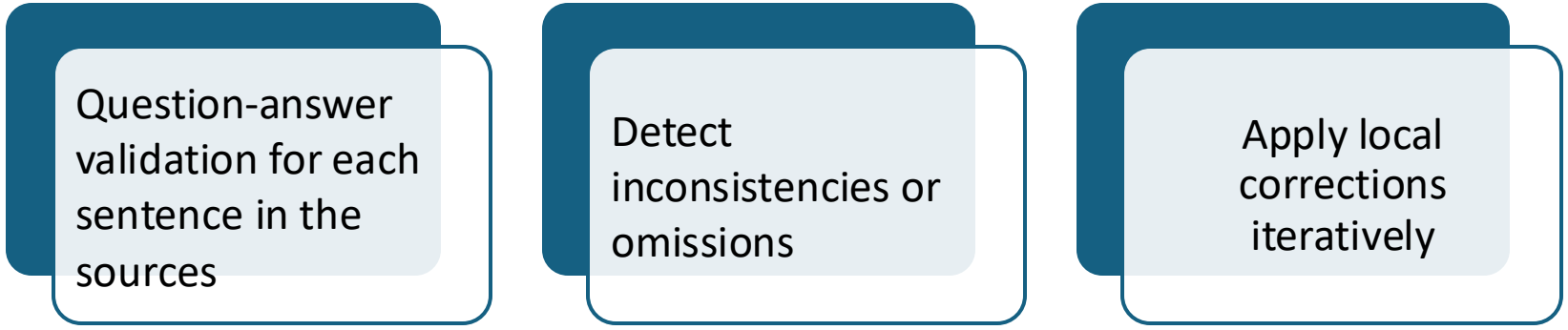
Source text:

"The Nile originates from the Great Lakes region of central Africa"

The QA pair:

- Q: Where does the Nile originate?
- A: The Great Lakes region of central Africa

Fine-Grained QA Analysis (Second-check)



Source text:

"Exposure to microplastics may alter gut microbial composition, showing profiles similar to those observed in depression and colorectal cancer."

The QA pair:

- Q: What effect do microplastics have on the human gut microbiome
- A: They may alter microbial composition, resembling profiles linked to depression and colorectal cancer.

Hallucination Identification (Third-check:)

Generate questions from the article itself

Ask all sources

If *none* can answer → hallucinated claim

Two variants: **classic** (per sentence) | **optimized** (per claim)

Draft sentence:
"Ferrari will introduce an aerodynamic upgrade at Suzuka that will improve lap times by 0.3 seconds."

Verification:
- "Did Ferrari announce an aerodynamic upgrade for Suzuka?"
- "What is the reported time improvement?"

Source Traceability (Fourth-check)

- Verify each final sentence across all sources
- Require explicit short quote (≤ 50 words)
- Remove non-traceable sentences
- Use the secondary model as double check

Draft:

"Sergio Casaro collaborated with director Sergio Leone on several film posters."

Literal quote (≤ 50 words) :

"Casaro designed posters for Leone's Western trilogy during the 1960s."

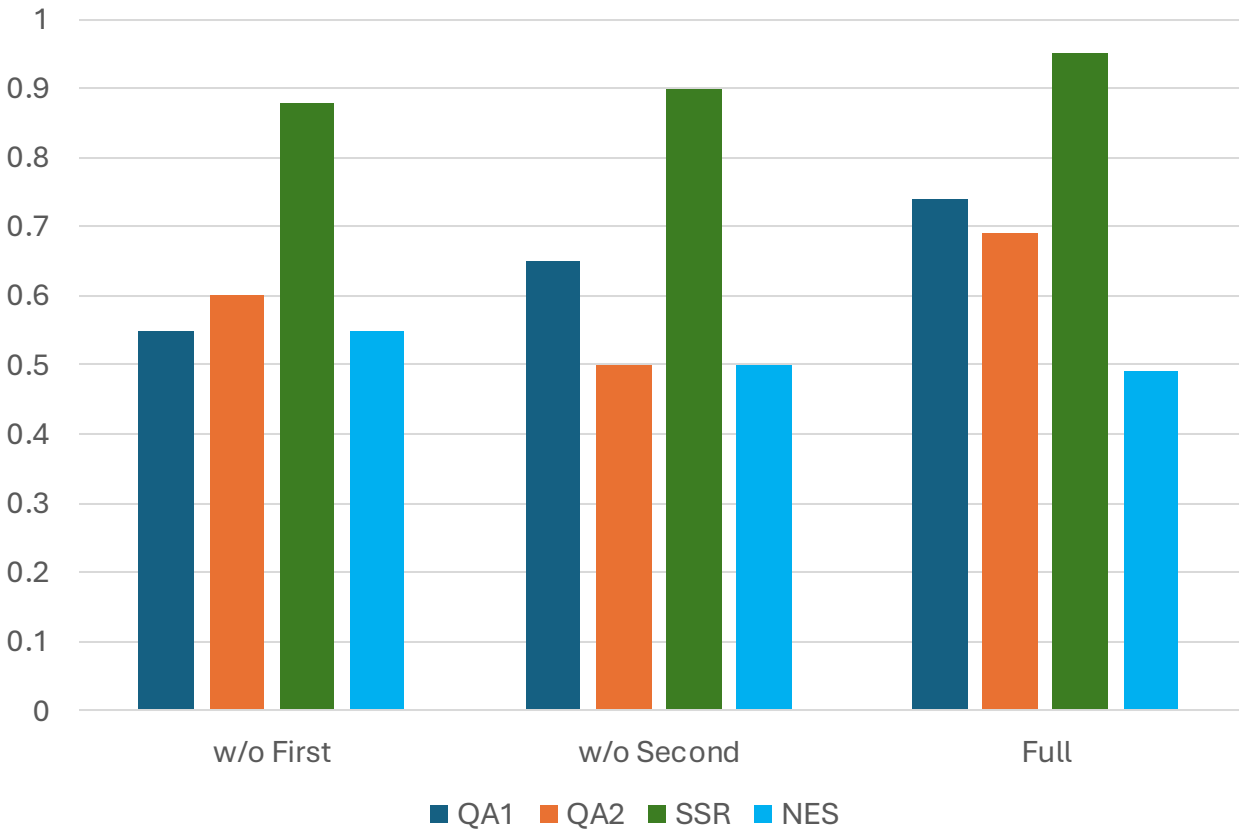
The secondary validator checks all factual elements of the sentence (entities, action, timeframe).

Evaluation Metrics

CATEGORY	METRIC	FORMULA
Traceability	Sentence Support Rate (SSR)	`#supportedsentences / #totalsentences`
Traceability	Attribution Coverage (AC)	`#sentenceswithcitations / #supportedsentences`
Traceability	Fourth Precision (Prec4)	`#non-supportedremoved / #actions(Fourth)`
Traceability	Fourth Recall (Rec4)	`#non-supportedremoved / #non-supportedtotal`
QA Accuracy	QA Accuracy Concept-level (QA1)	`#correctanswers(Concept-level) / Q(Concept-level)`
QA Accuracy	QA Accuracy Sentence-level (QA2)	`#correctanswers(Sentence-level) / Q(Sentence-level)`
Hallucination	Removal Success Rate (RSR)	`#unsupportedclaimscorrected / #unsupportedclaims`
Preservation	Retention Rate (RR)	`Nfinal / Ninitial`
Preservation	Normalized Edit Similarity (NES)	`SequenceMatcher(initial,final)`

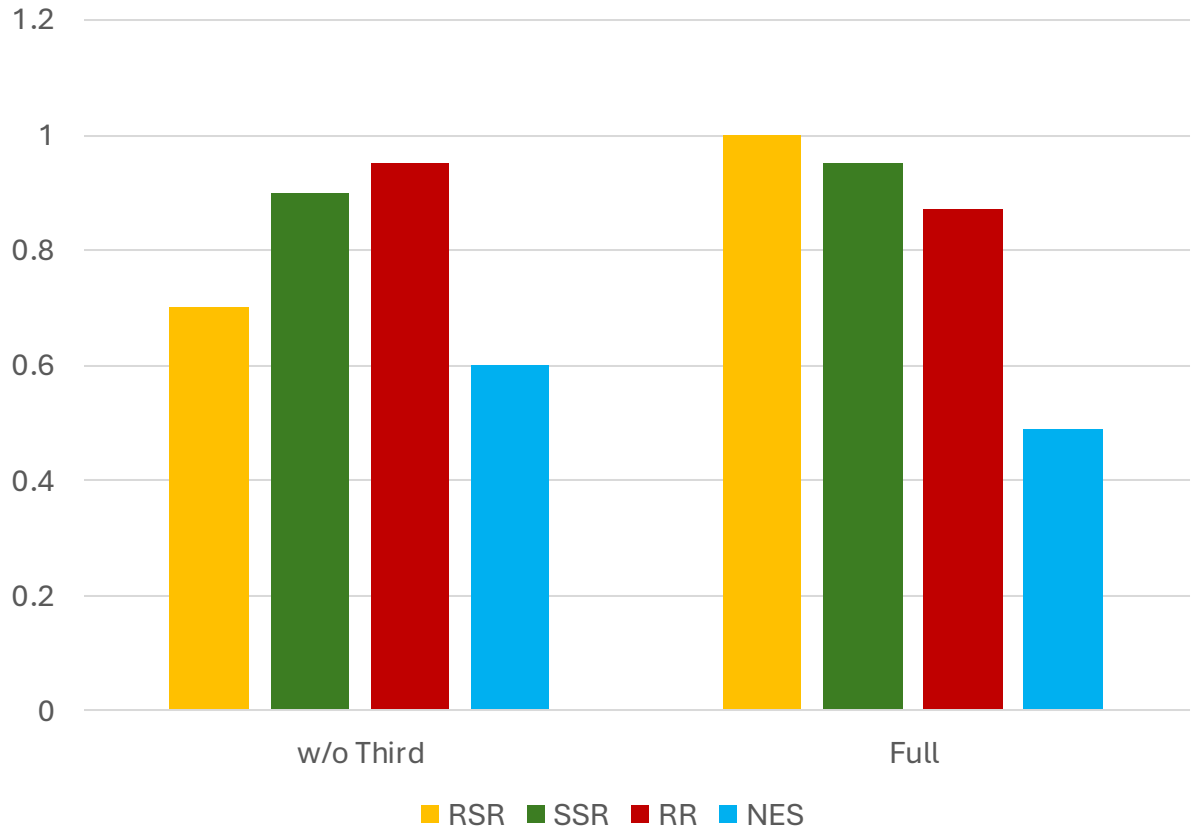
Effect of First and Second Check

Ablation	QA1	QA2	SSR	NES
w/o First	0.55	0.60	0.88	0.55
w/o Second	0.65	0.50	0.90	0.50
Full	**0.74**	**0.69**	**0.95**	**0.49**



Effect of Third Check

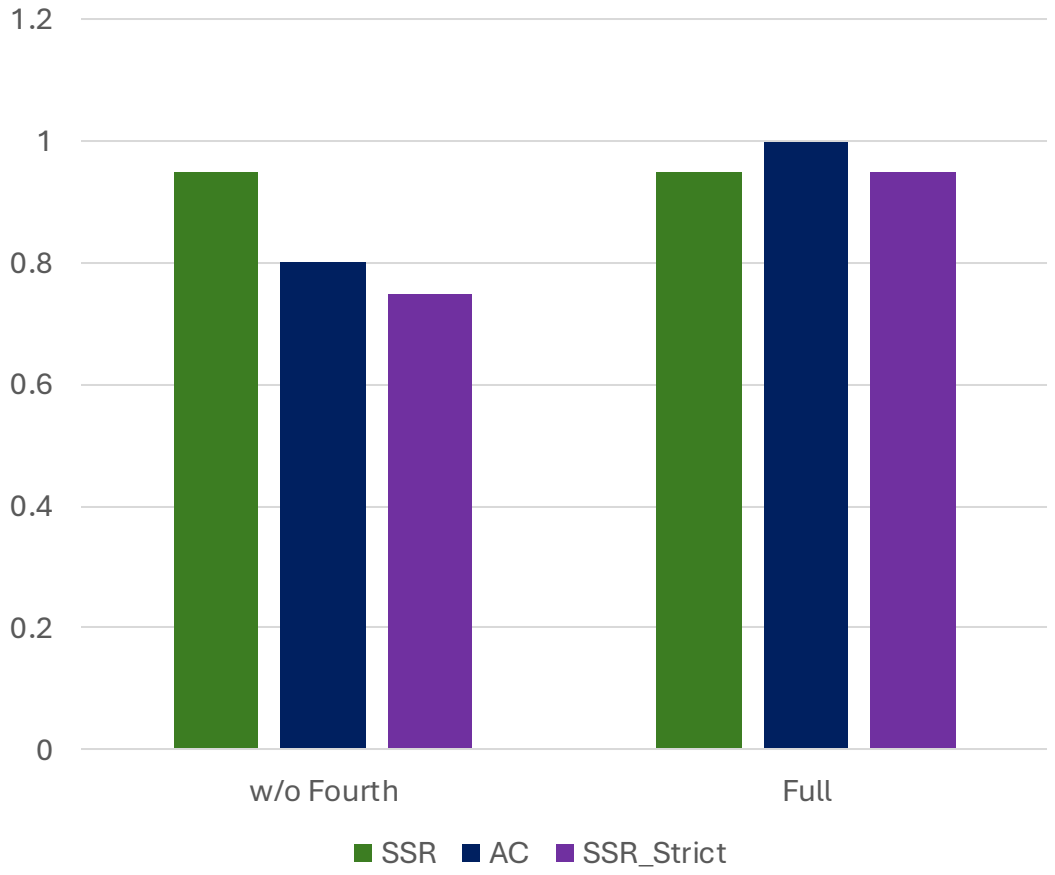
Ablation	RSR	SSR	RR	NES
w/o Third	0.70	0.90	**0.95**	0.60
Full	**1.00**	**0.95**	0.87	**0.49**



Effect of Fourth Check

Ablation	SSR	AC	SSR_strict
w/o Fourth	**0.95**	0.80	0.75
Full	**0.95**	**1.00**	**0.95**

- **Removing** one check **lowers** *SSR* and *RSR*
- Trade-off: factuality ↑, small reduction in length



Qualitative Results (Part I — Factual Corrections)

Scenario	Before	After	Observetion
Medicine	“Emily Chen awarded the 2025 Nobel Prize for immune memory.”	“Mary E. Brunkow, Fred Ramsdell and Shimon Sakaguchi for the discovery of regulatory T cells.”	Invented name replaced with real laureates.
Science	“Microplastics were proven to increase serotonin.”	“Exposure to microplastics may alter gut microbiome profiles similar to those in depression or colorectal cancer.”	Rewritten to match scientific evidence.

Qualitative Results (Part II — Hallucination Removal)

Scenario	Before	After	Observation
Economy	“ECB expected a 1.8 % growth in 2026.”	*(Removed)*	Fabricated numeric forecast deleted.
Sport	“Ferrari confirmed new contract for 2026.”	*(Removed)*	Non-existent contract claim erased.
Art	“Casaro collaborated with Christopher Nolan for *Oppenheimer*.”	*(Removed)*	False collaboration removed.

Strengths and Limitations

Advantages

Modular design

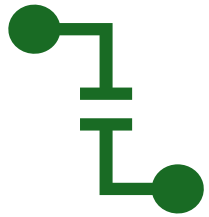
High factual reliability

Automatic correction capability

Traceability assurance

Cross-domain robustness

Transparent evaluation



Challenges

Computational cost

Dependency on prompt quality

Risk of over-correction

Model bias inheritance

Limited multi-source reasoning

Conclusions and Future Perspective

Key Achievements

- **Modular multi-step framework** for detection, correction, and removal.
- **Fact-checking, correction, and source traceability** within a single end-to-end pipeline.
- Improvements in **factual reliability** and **coherence preservation**.
- **Evaluation metrics** to quantify factual robustness.

Future Directions

- **Multi-document aggregation**
- **Cross-media verification**
- **Explainable fact-verification**
- **Efficiency optimization**

Thank you for Your attention