Analyzing and Improving Post-hoc Approaches for the Detection and Correction of Hallucinations in Long-form Text Generation

Ihsan Soydemir

28.10.2024, Master Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

- Motivation
- Research Questions

- Methodology
  - Fact-Check Systems
  - Datasets
  - Experiment Setup
  - Evaluation

- Results
  - Analysis
  - Improvements
  - Human Evaluation
  - Taxonomy

- Conclusion
- Future Work

Motivation

Motivation

- Inevitability of hallucinations
- Fictional facts
- Importance for critical applications
- Building trust for LLMs
- Improving detection and correction mechanisms

This image was generated with the assistance of AI

Research Questions

Research Questions

RQ1

What is an appropriate taxonomy for categorizing hallucinations in LLMs?

RQ2

RQ3

Research Questions

RQ1

What is an appropriate taxonomy for categorizing hallucinations in LLMs?

RQ2

How can RAG model be effectively applied to handle long contexts without compromising performance or efficiency?

RQ3

Research Questions

RQ1

What is an appropriate taxonomy for categorizing hallucinations in LLMs?

RQ2

How can RAG model be effectively applied to handle long contexts without compromising performance or efficiency?

RQ3

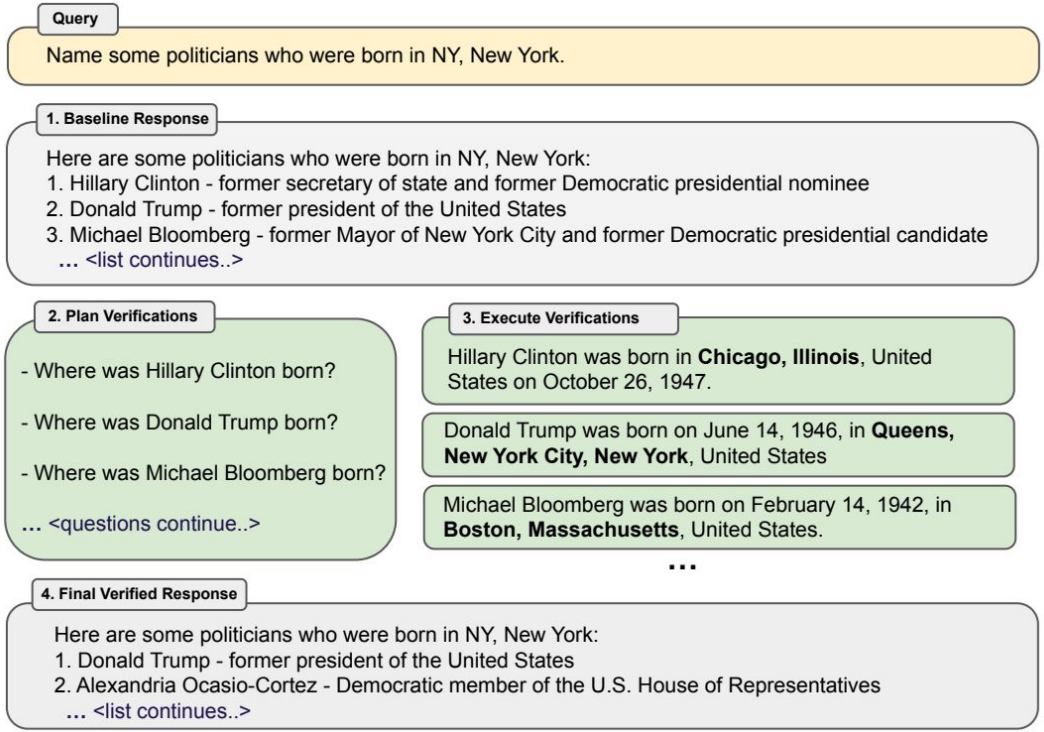
How can we optimally balance editing and faithfulness when refining generated text?

Fact-Check Systems

Chain-of-Verification [Meta AI, 2023]

- Steps
  - Initial Claim
  - Question Generation
  - Verification of Questions
    - Internal Knowledge
    - External Knowledge
      - Online Search
      - DBs
  - Refinement of final response

- Gaps
  - Minimal changes with executed verifications
  - Sometimes reasoning fine but lacks refinement

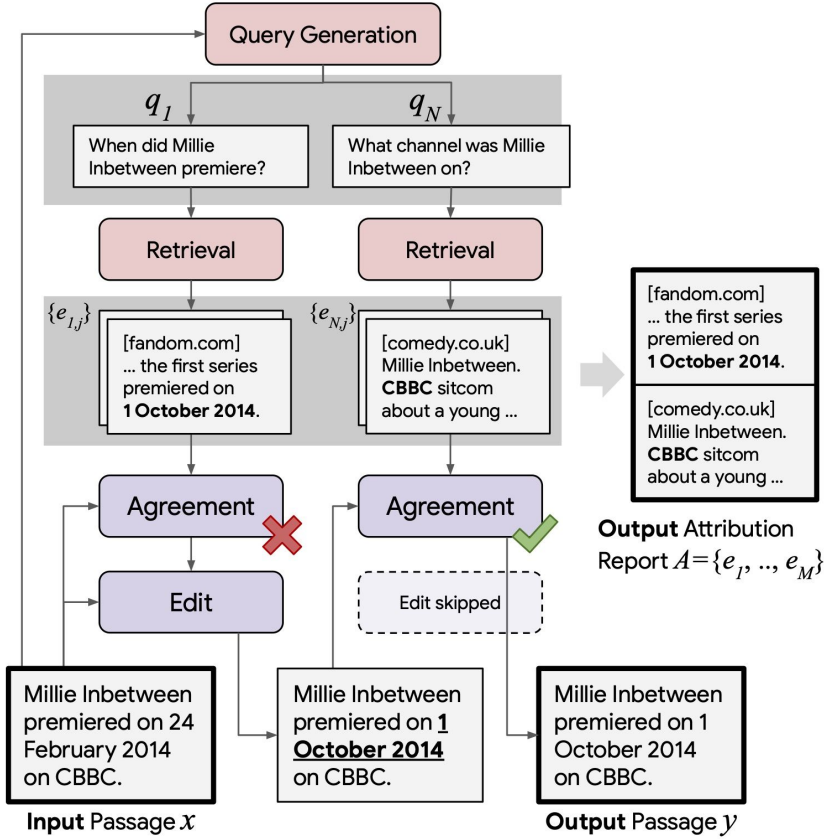


Source: https://arxiv.org/abs/2309.11495

RARR [Google, 2023]

- 4-Step Process
  - Question Generation
  - Evidence Collection
  - Agreement Test
  - Editing

- Gaps
  - Excessive refinement
    - Changes original text significantly



Source: https://arxiv.org/abs/2210.08726

Datasets

SummEdits

- “LLMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond”
- 10 Domains
  - News
  - Podcast Transcripts
  - Scientific TL; DRs
- Edit types
  - Entity modification
  - Hallucinated fact insertion
  - Antonym Swap
- Creates a hallucinated summary

Summary: A report by Forbes found that Binance secretly transferred ~~$1.8~~ $1.5 billion worth of funds to several hedge funds in August 2022...

~~original~~
edited

Original Document: A report by Forbes found that Binance transferred nearly $1.8 billion worth of stablecoins collateral to hedge funds last year. The beneficiaries included Justin Sun, Cumberland, and Sam Bankman-Fried's Alameda Research. Forbes has claimed that the transfer of assets was similar to maneuvers made by FTX before art of its daily business activities. He also revealed that other than the wallets, the exchange had ledgers that kept track of all funds owed to customers and tokens that were moved around between wallets.

ExpertQA

- Q&A Dataset
- Key Fields
  - Question
  - Answer
    - LLM generated
    - Expert revised

Question: How can AI help to combat the housing crisis in the UK?

LLM Answer: AI can optimize construction, improve building designs, and analyze housing data. [...] also predicts housing needs, guiding policy to ensure affordable housing where needed [...]

Expert Answer (Revised): AI can address the UK housing crisis by improving construction processes, analyzing housing data to inform decisions [...] also helps analyze housing market data—property values, rent, demographics—informing policy and planning by identifying areas needing affordable [...]

Long Form Factuality (from DeepMind)

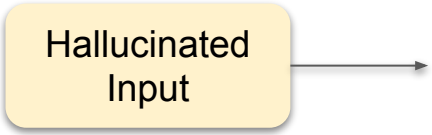
- Concepts
  - Require deeper exploration
  - Multiple Entities
- Objects
  - Easier to address
  - Specific, tangible entities
- 38 Topics
  - Biology
  - Economics
  - History
  - Physics

Generated Concept Question: Could you explain the fundamentals of Natural Language Processing (NLP) in Computer Science, its key techniques [...], and discuss [...] in advancing fields like machine translation, sentiment analysis, and information retrieval?

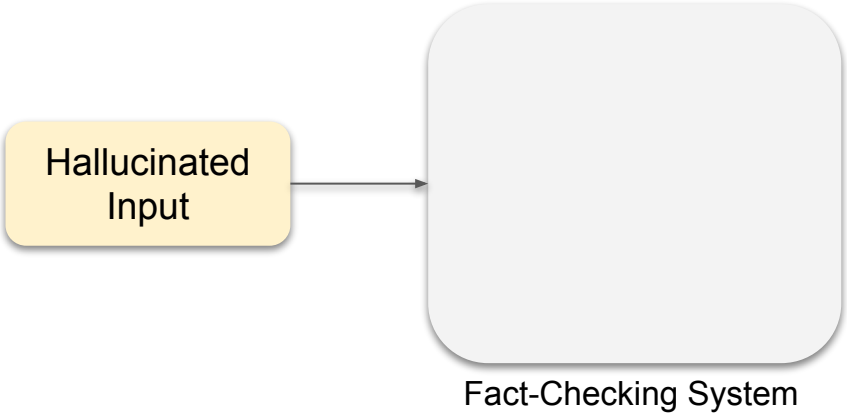
Generated Object Question: Can you tell me about TRIUMF, Canada's national particle accelerator facility?

Experiment Setup

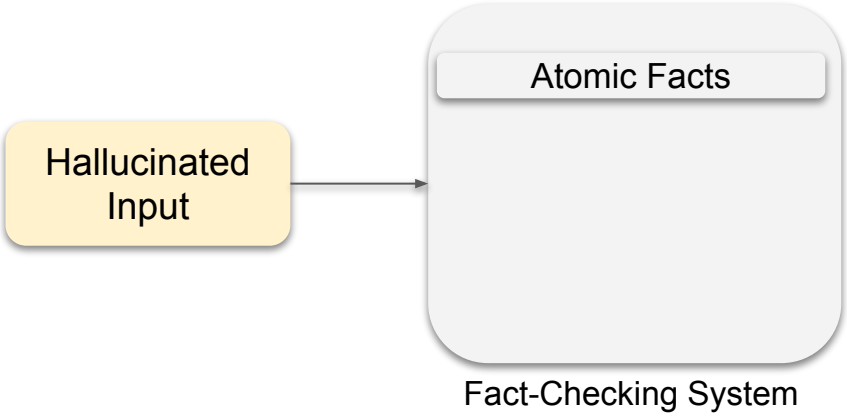
Experiment Setup



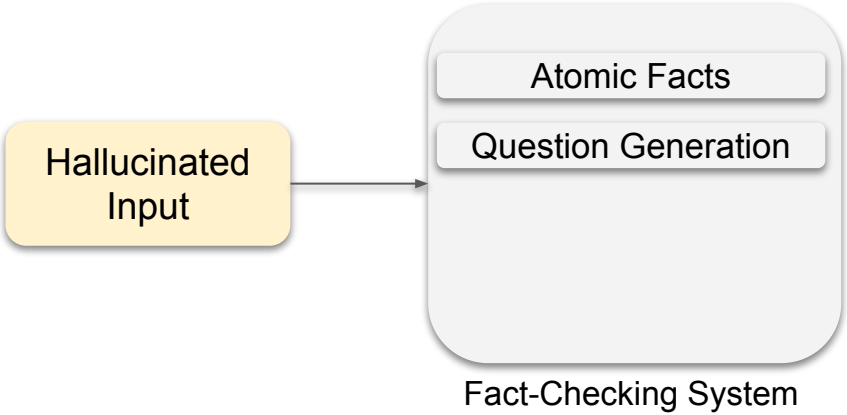
Experiment Setup



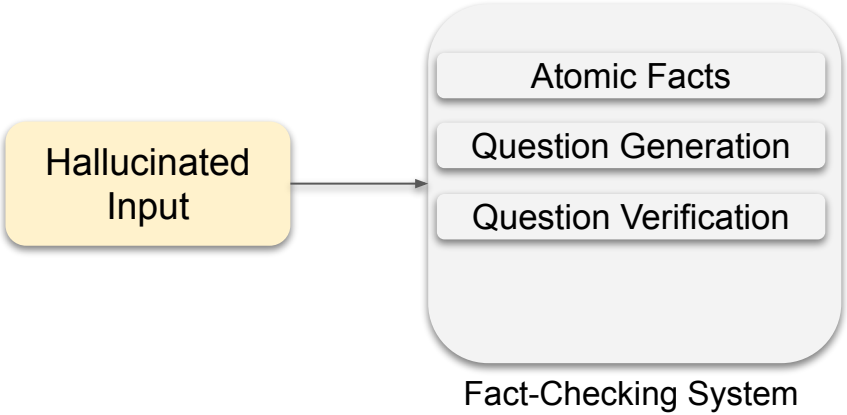
Experiment Setup



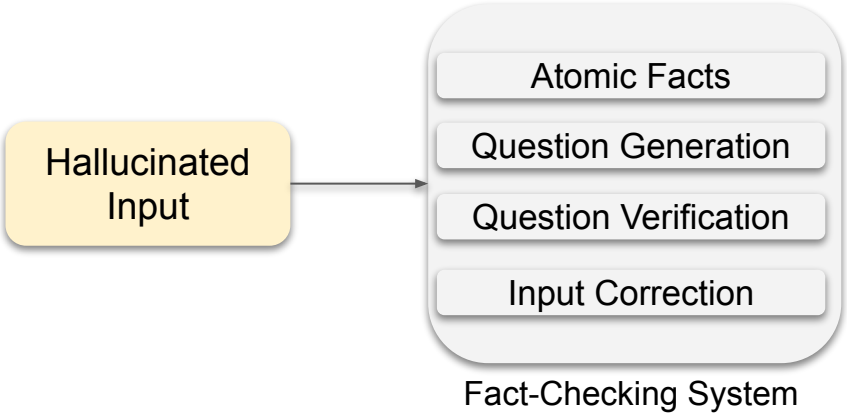
Experiment Setup



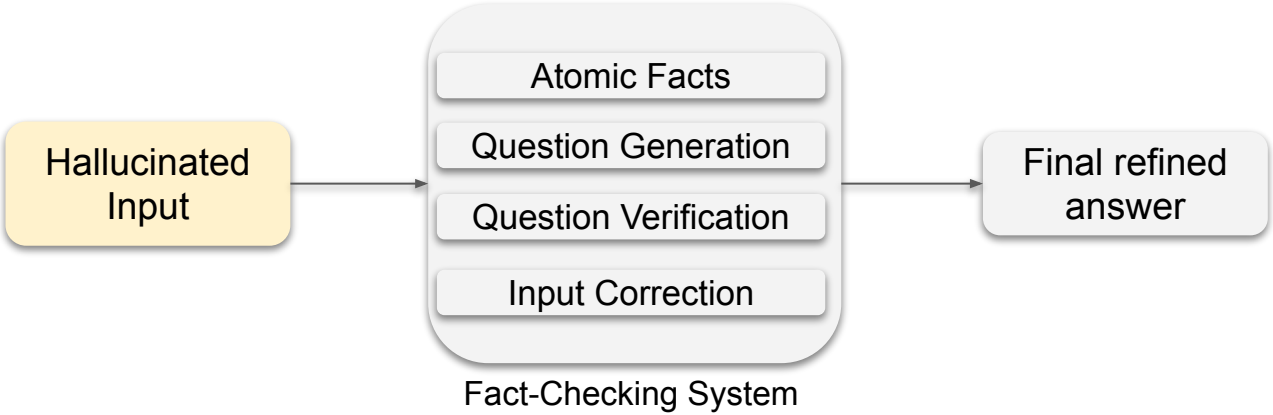
Experiment Setup



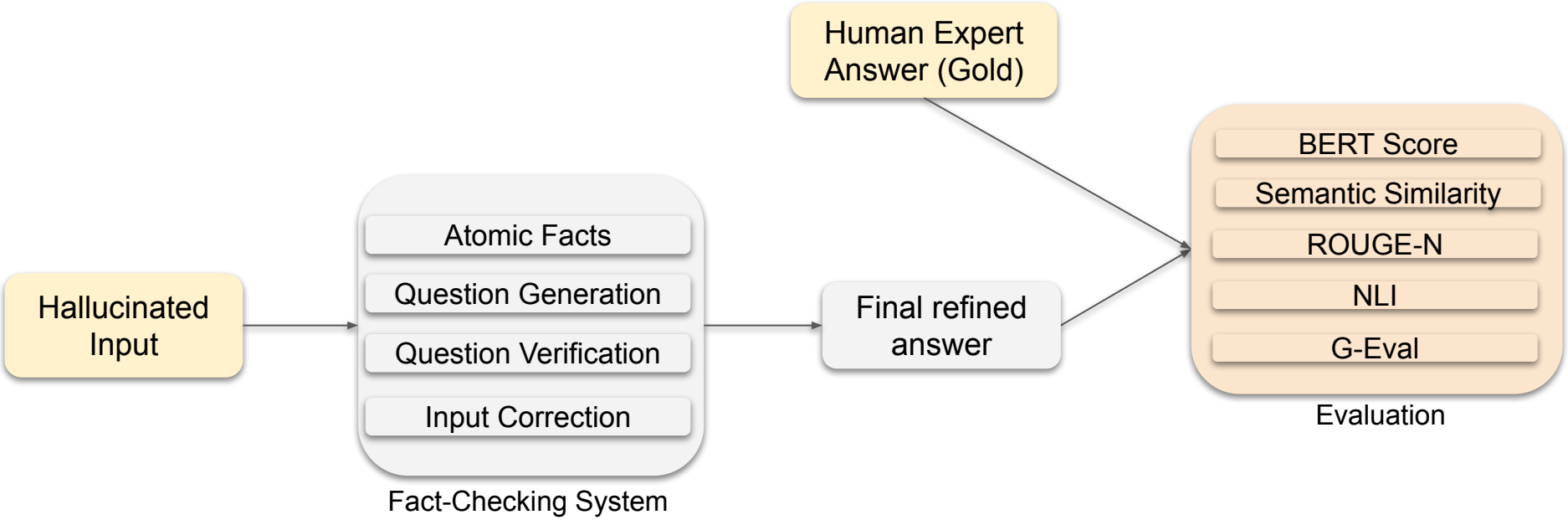
Experiment Setup



Experiment Setup



Experiment Setup



Evaluation

Evaluation

- BERT Score
  - Compares token embeddings to measure text similarity (token-level)
  - $$\frac{1}{n}\sum \cos(\text{Embedding}(S_i), \text{Embedding}(T_i))$$

- Semantic Similarity
  - Cosine similarity between two text embeddings (text-level)
  - $$\frac{\text{Embedding}(A) \cdot \text{Embedding}(B)}{\|\text{Embedding}(A)\| \|\text{Embedding}(B)\|}$$

- Normalized Edit Distance
  - Counts minimum edits needed to match texts, normalized by length
  - $$\frac{\text{Edit Operations}(S_1, S_2)}{\max(\text{Length}(S_1), \text{Length}(S_2))}$$

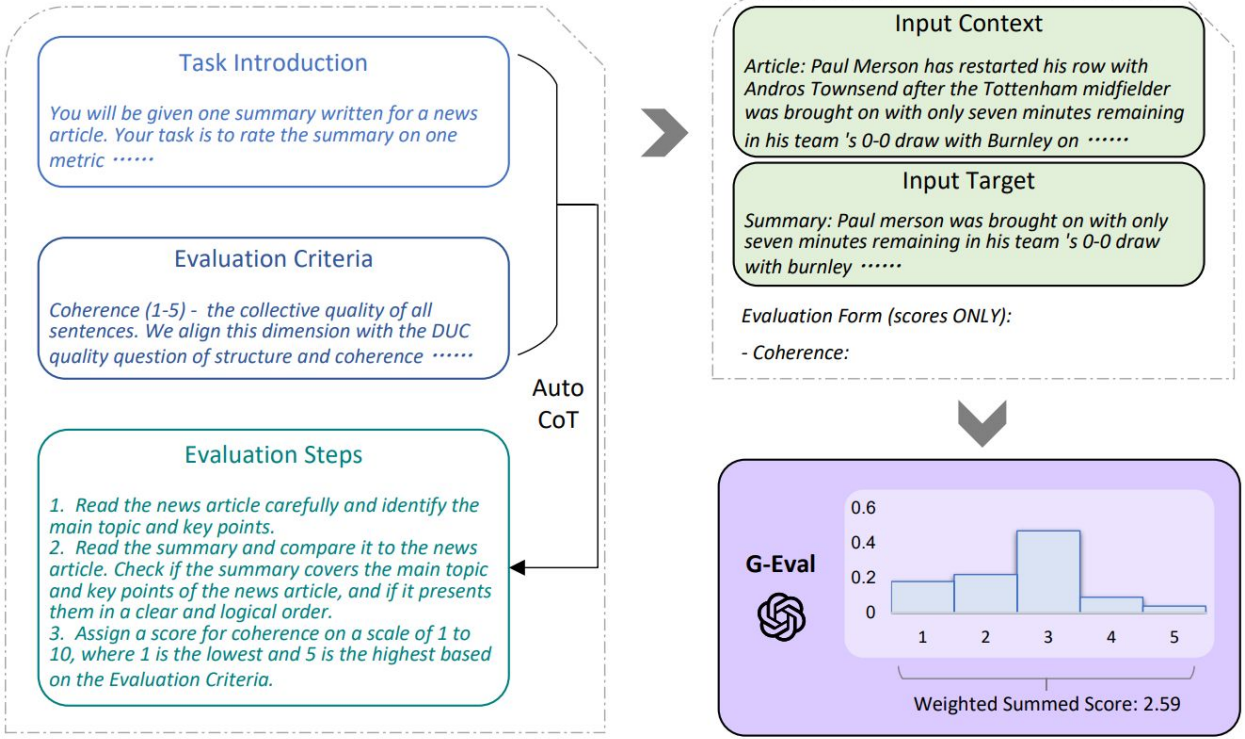
Evaluation

- ROUGE-N
  - Recall of word sequences (n-grams) in generated vs. reference text
  - $$\frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- Natural Language Inference (NLI)
  - Fine-tuned DeBERTa-v3 model
  - Labels
    - Entailment
    - Neutral
    - Contradiction

Evaluation

- G-Eval [Microsoft Research, 2023]
  - “NLG Evaluation with Better Human Alignment”
  - LLM-as-a-Judge
  - CoT prompting



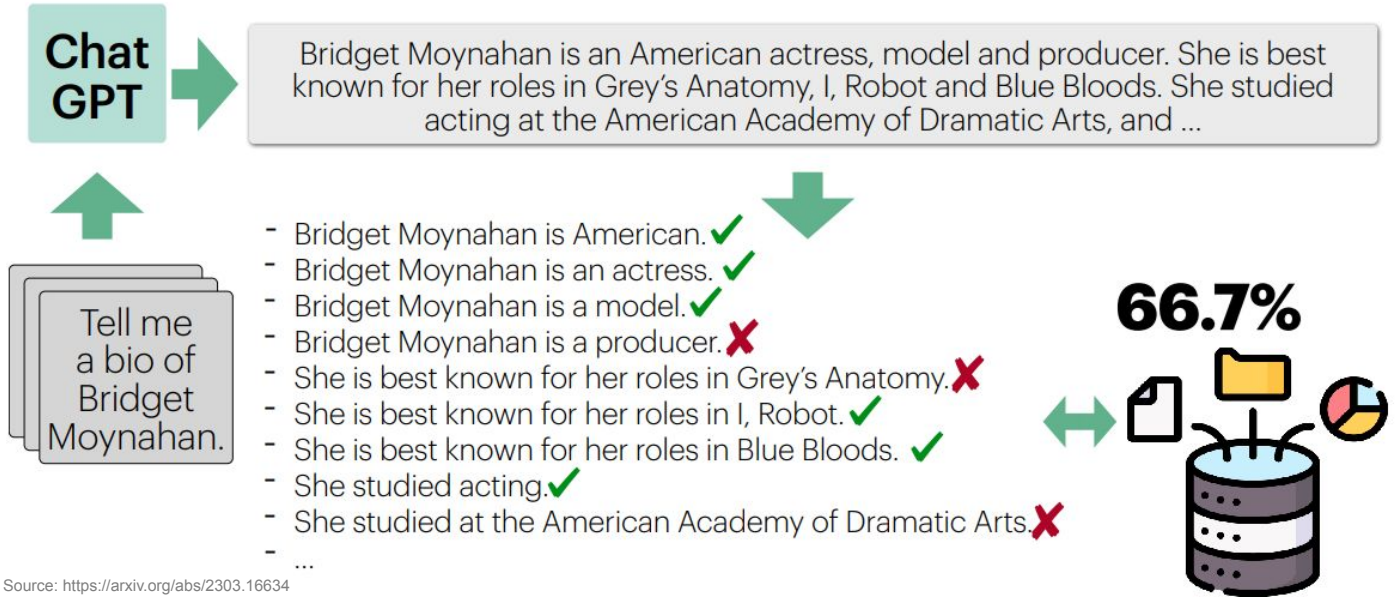
Source: https://arxiv.org/abs/2303.16634

Evaluation

- FActScore
  - “Evaluating Factual Precision of Long-form Text Generation”
  - Register knowledge source

$$f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} \mathbb{I}[a \text{ is supported by } \mathcal{C}],$$

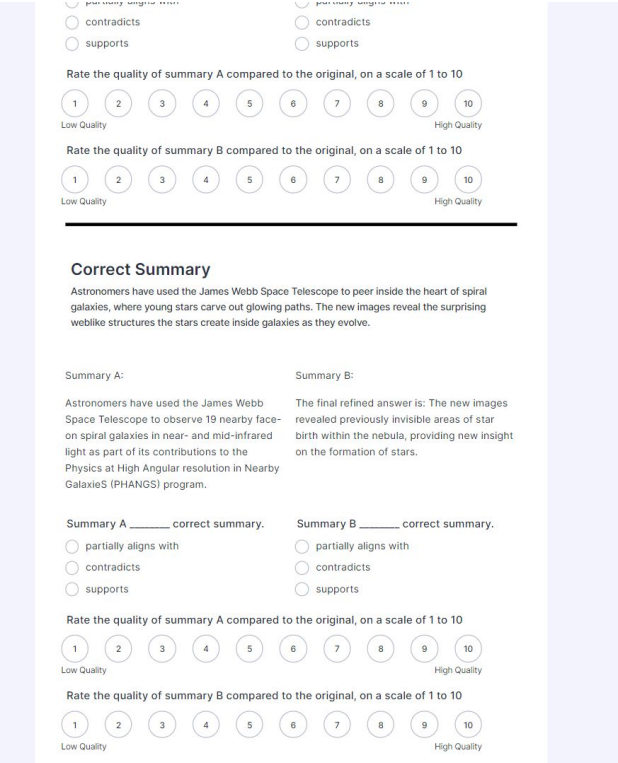
$$\text{FACTSCORE}(\mathcal{M}) = \mathbb{E}_{x \in \mathcal{X}}[f(\mathcal{M}_x) | \mathcal{M}_x \text{ responds}]$$



Source: https://arxiv.org/abs/2303.16634

Evaluation

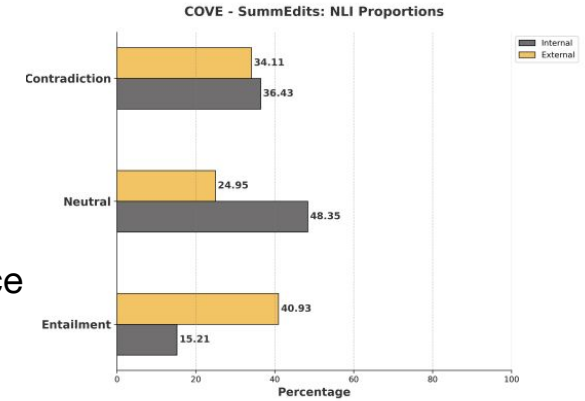
- Human Evaluation
- Goal
  - Check if automatic evaluations for NLI and G-Eval align with human preferences
- Quality (Factual accuracy + Relevance)
  - Factual accuracy
    - Is it based on facts?
    - Avoids misinformation?
  - Relevance
    - Does the summary cover the main points?
    - Off-topic?
- NLI
  - Contradiction
  - Entailment (Supports)
  - Neutral (Partially aligns with)



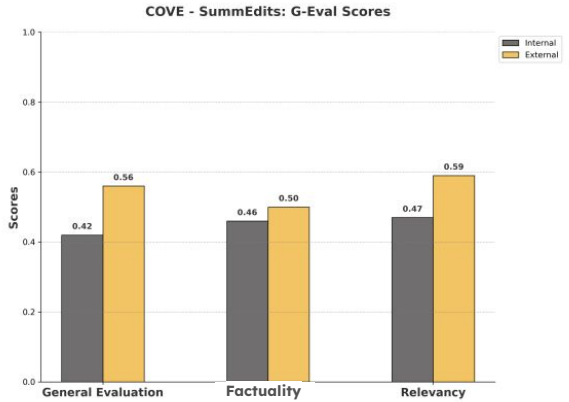
Results

Analysis - Internal and External Evidence Sources

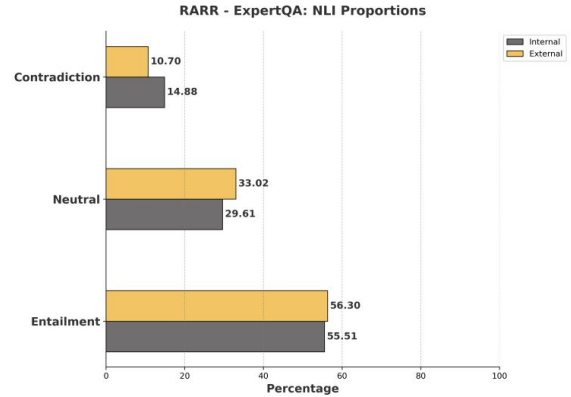
- Internal
  - Self-knowledge
  - Limited by cutoff
- External
  - Outside access
    - Online source
    - Databases



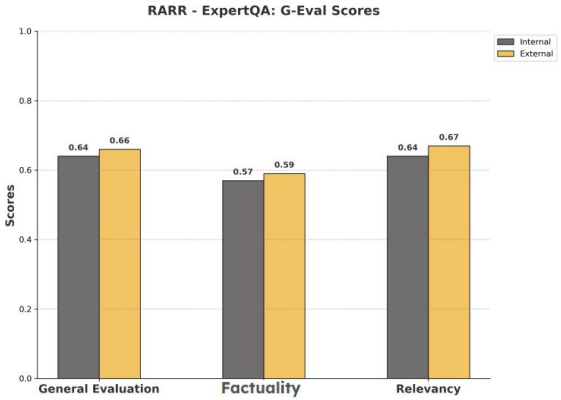
(a) NLI Proportions for SummEdits



(b) G-Eval Scores for SummEdits



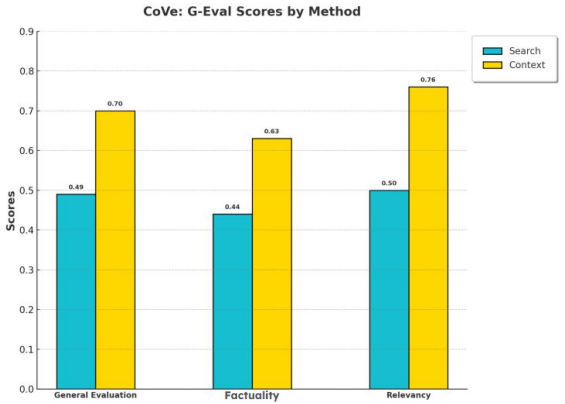
(a) NLI Proportions for ExpertQA



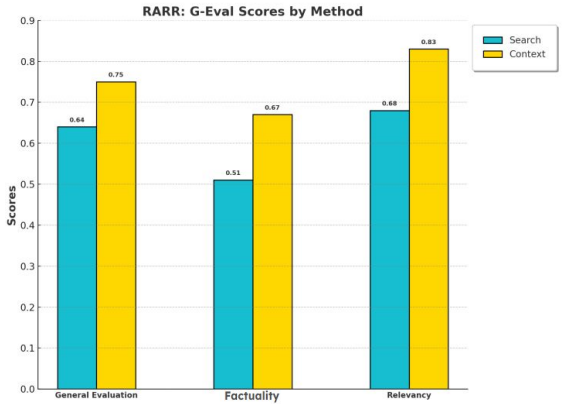
(b) G-Eval Scores for ExpertQA

Analysis - Document as Context

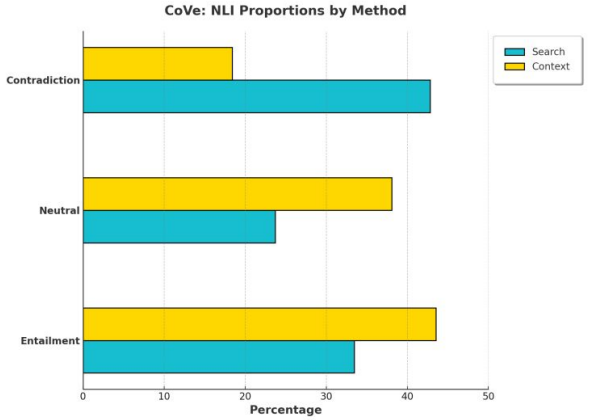
- Search
  - Retrieving relevant information online
  - e.g., Google, Bing
- Context
  - Original Document
  - Primary Source



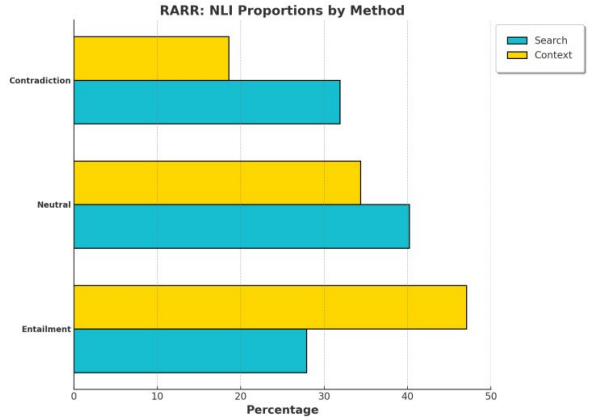
(a) CoVE - G-Eval Comparison



(b) RARR - G-Eval Comparison



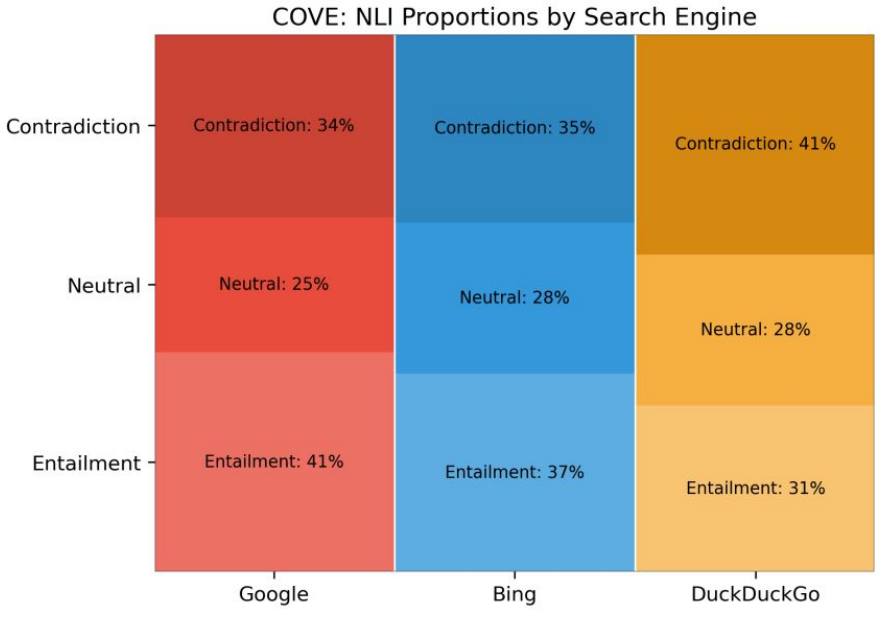
(a) CoVE - NLI Results

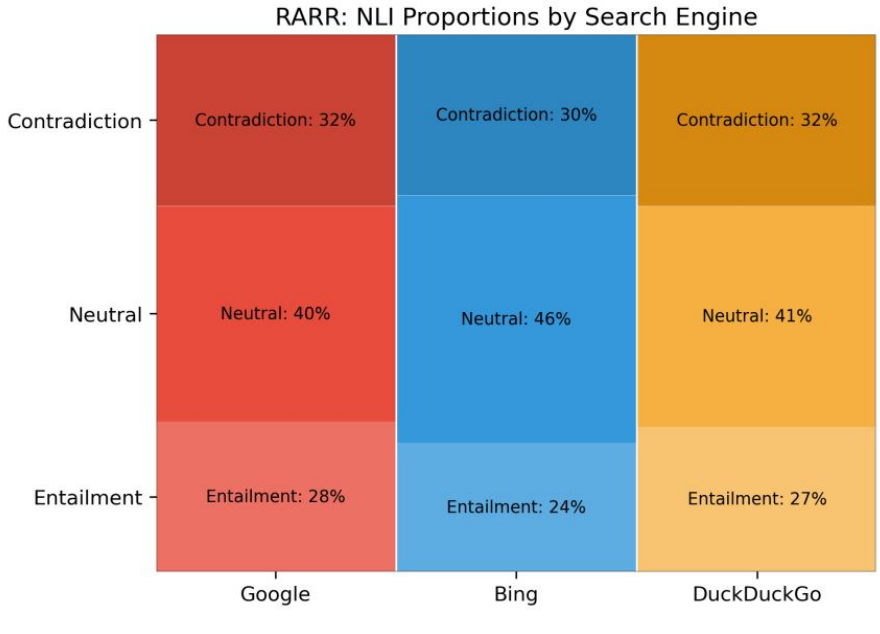


(b) RARR - NLI Results

Analysis - Different Search Engines

- Google
- Bing
- DuckDuckGo





Analysis - LongFact Responses

- LongFact

Metric	Character Count	Word Count
Mean	1758	244
Standard Deviation	517	70
Median	1838	260

Table 5.4: LLM Response Lengths for LongFact questions (`max_tokens = 500`)

Question: Can you tell me about TRIUMF, Canada's national particle accelerator facility?

triumf is canada's national particle accelerator facility, located near vancouver, british columbia established in 1968, it is one of the world's leading centers for research in nuclear and particle physics the facility is operated by a consortium of canadian universities and is funded by the canadian government, as well as various international partnerships

key features of triumf:

1 **particle accelerator**: the centerpiece of triumf is its cyclotron, which accelerates protons to high energies the facility can produce intense beams of protons and other particles, enabling a wide range of experiments in nuclear and particle physics

2 **research areas**: triumf conducts research in several key areas:
- **nuclear physics**: studies the structure and behavior of atomic nuclei
- **particle physics**: investigates fundamental particles and their interactions
- **medical isotope production**: produces isotopes for medical diagnostics and treatment, such as those used in pet scans
- **material science**: uses particle beams to study and develop new materials

3 **collaborations**: triumf collaborates with various international laboratories and universities, contributing to global research efforts in fundamental physics, including projects at cern and other major facilities

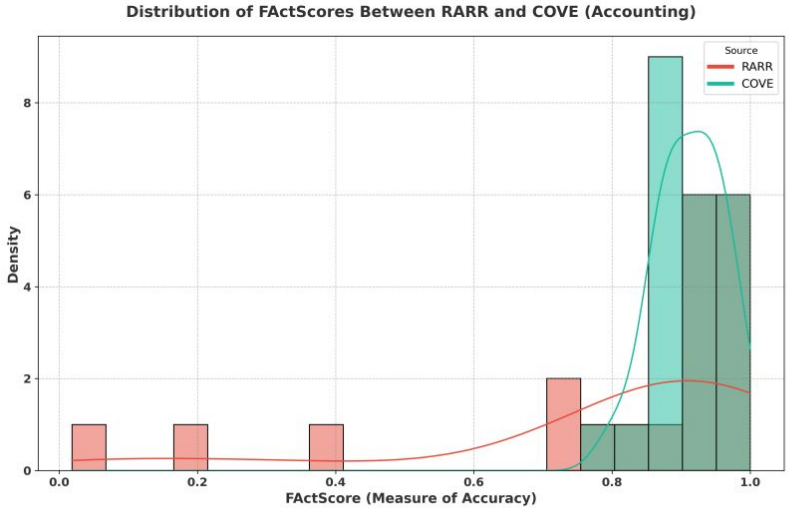
4 **education and outreach**: triumf is also involved in education and outreach activities, providing training for students and researchers, and engaging the public with science through tours, lectures, and community programs

5 **innovative technologies**: the facility is known for developing advanced technologies in accelerator physics, detector technology, and medical imaging, further enhancing its research capabilities and contributions to various fields

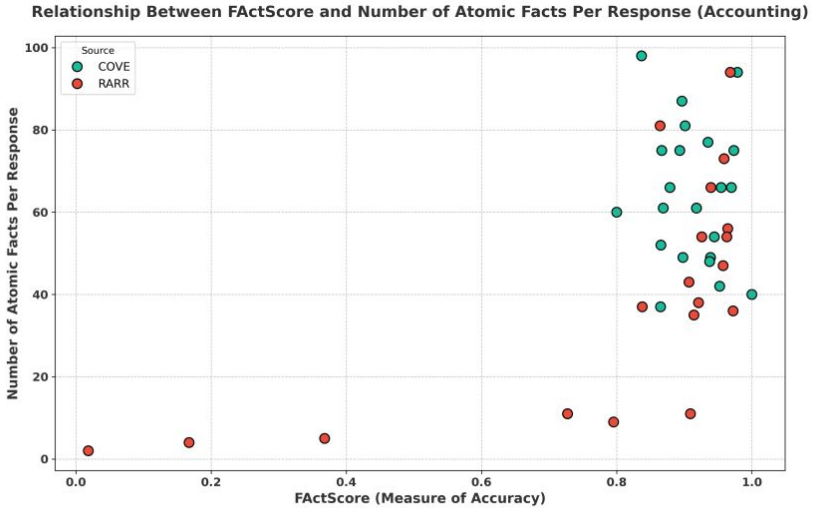
triumf plays a crucial role in advancing our understanding of the universe at the most fundamental level and is a key player in both national and international scientific communities

Analysis - LongFact Experiments

- LongFact w/ Factscore
- COVE vs. RARR



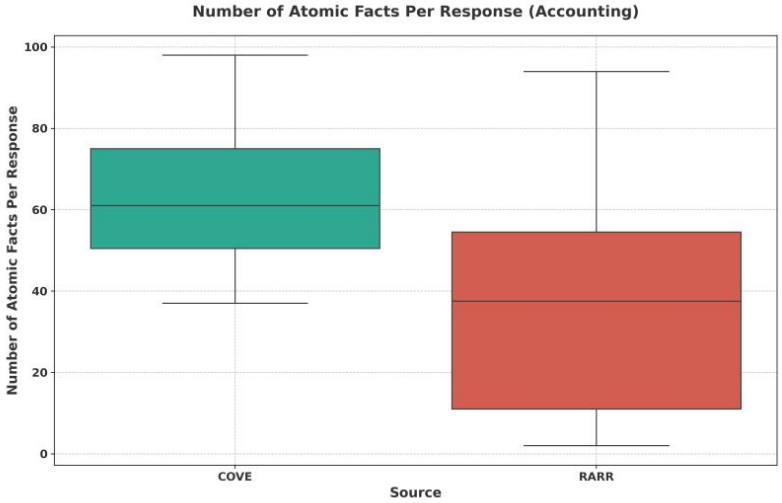
(a) FactScore Distribution



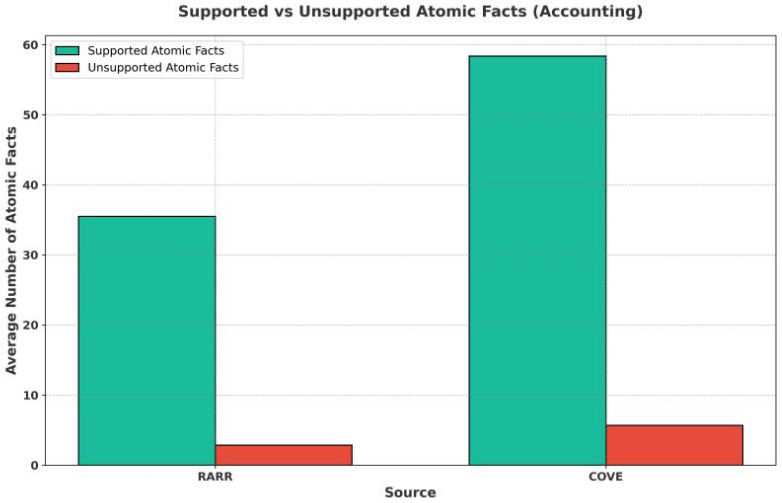
(b) FactScore vs. Atomic Facts per Response

Analysis - LongFact Experiments

- LongFact w/ Factscore
- COVE vs. RARR



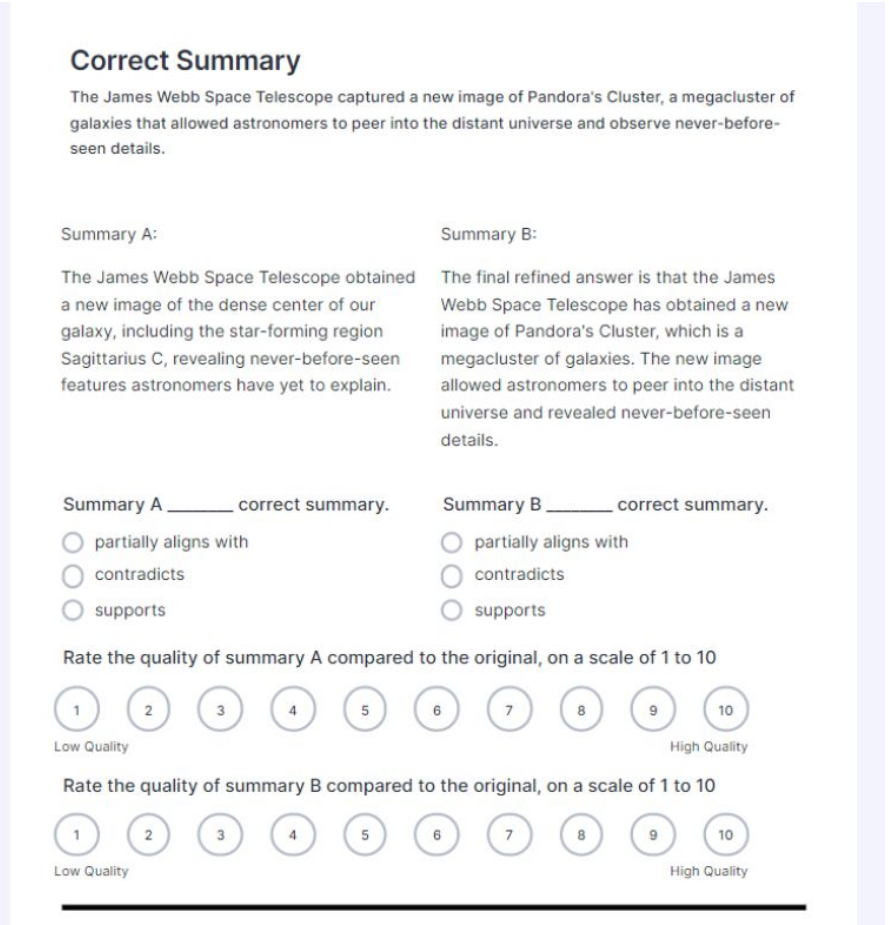
(a) Number of Atomic Facts per Response



(b) Supported vs. Unsupported Atomic Facts

Human Evaluation

- 25 participants
- Each user => 40 questions



Correct Summary

The James Webb Space Telescope captured a new image of Pandora's Cluster, a megacluster of galaxies that allowed astronomers to peer into the distant universe and observe never-before-seen details.

Summary A:

The James Webb Space Telescope obtained a new image of the dense center of our galaxy, including the star-forming region Sagittarius C, revealing never-before-seen features astronomers have yet to explain.

Summary B:

The final refined answer is that the James Webb Space Telescope has obtained a new image of Pandora's Cluster, which is a megacluster of galaxies. The new image allowed astronomers to peer into the distant universe and revealed never-before-seen details.

Summary A _____ correct summary.

- partially aligns with
- contradicts
- supports

Summary B _____ correct summary.

- partially aligns with
- contradicts
- supports

Rate the quality of summary A compared to the original, on a scale of 1 to 10

1 2 3 4 5 6 7 8 9 10

Low Quality — High Quality

Rate the quality of summary B compared to the original, on a scale of 1 to 10

1 2 3 4 5 6 7 8 9 10

Low Quality — High Quality

Human Evaluation

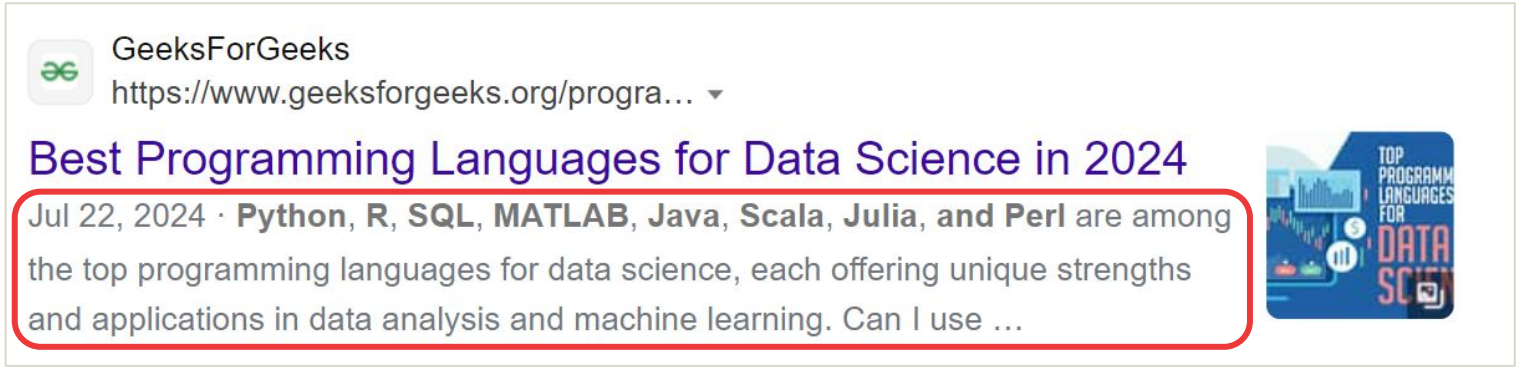
- In total, 1000 questions were answered

Method	Human Mean Score	G-Eval Score	Diff
RARR	0.68	0.65	0.03
COVE	0.54	0.52	0.02

Method	Human			NLI Model		
	Entailment	Neutral	Contradiction	Entailment	Neutral	Contradiction
RARR	45	40	15	30	49	21
COVE	31	37	32	28	47	25

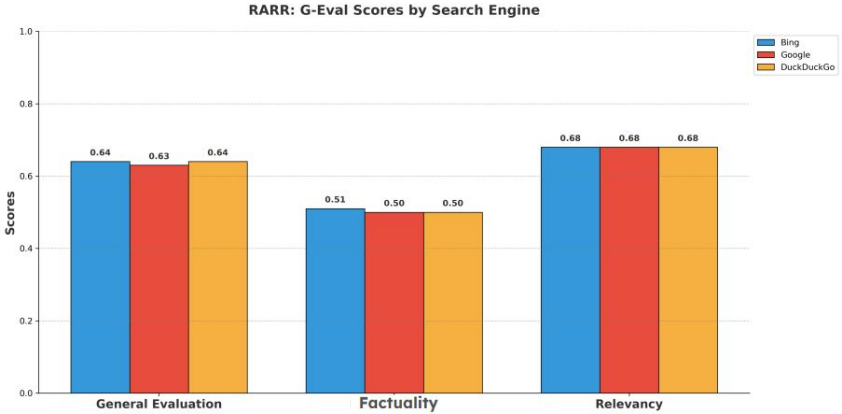
Improvements - Snippets over Full-Text Retrieval

- Search Snippets

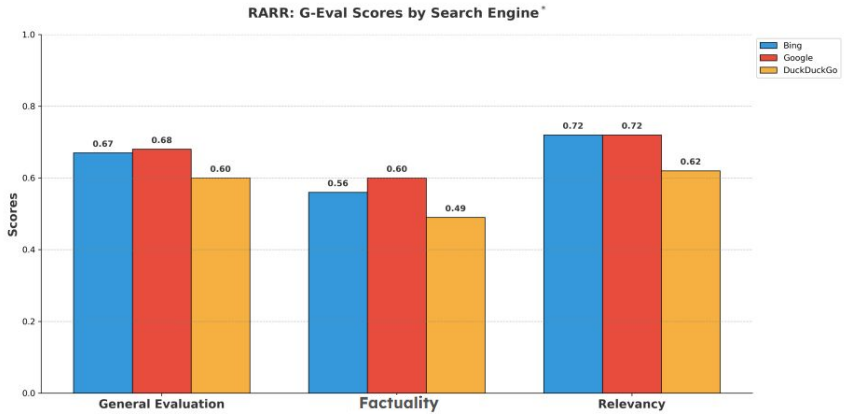


- Passages
  - URL
  - Full-text
  - Retrieval
  - Ranking
  - Top-5

Improvements - Snippets over Full-Text Retrieval

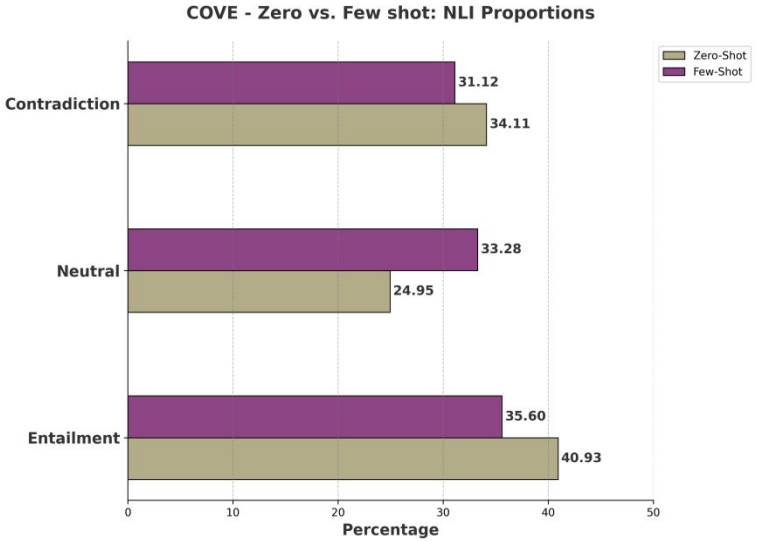


(a) Passage Retrieval - G-Eval Comparison



(b) Snippets - G-Eval Comparison

Improvements - Few-shot over Zero-Shot

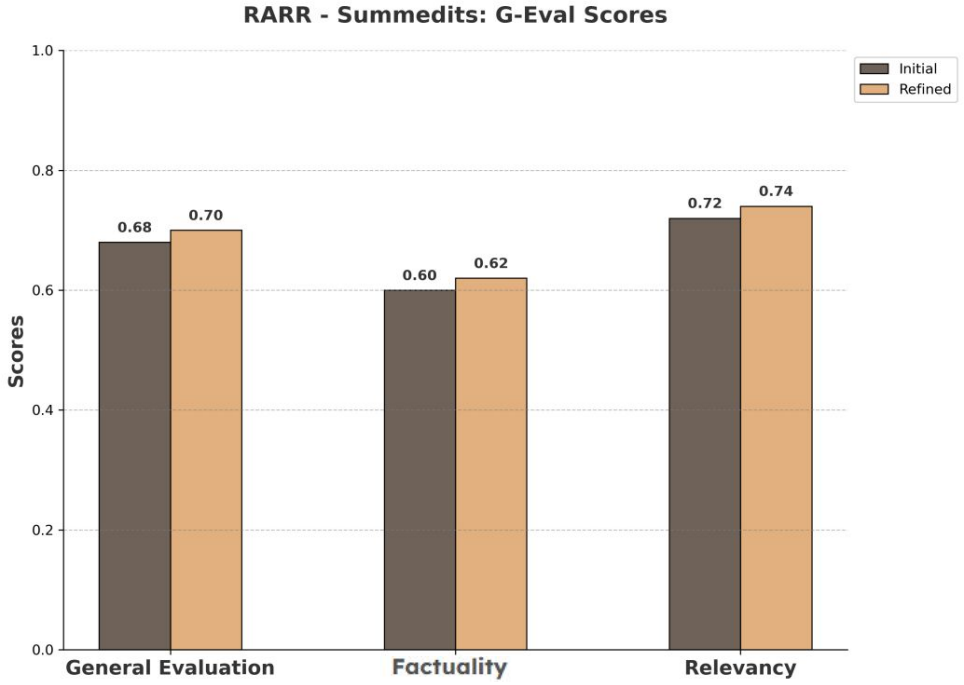


(a) NLI Proportions Comparison

Original	Correction (Zero-Shot)	Improved Correction (Few-Shot)
A Chinese mine collapse killed at least two people and left over 50 others missing. Chinese officials say numerous vehicles were also buried in the collapse.	**The final refined answer is:** A Chinese mine tragedy killed at least two people and left over 50 others missing. Chinese officials say numerous vehicles were also buried in the collapse. **The number of people reported missing after the incident is not provided in the given context.**	The Chinese mine tragedy led to the deaths of at least two individuals, with over 50 people reported missing. Chinese officials have also confirmed that multiple vehicles were buried in the collapse.

Improvements - Focused Context and Specificity

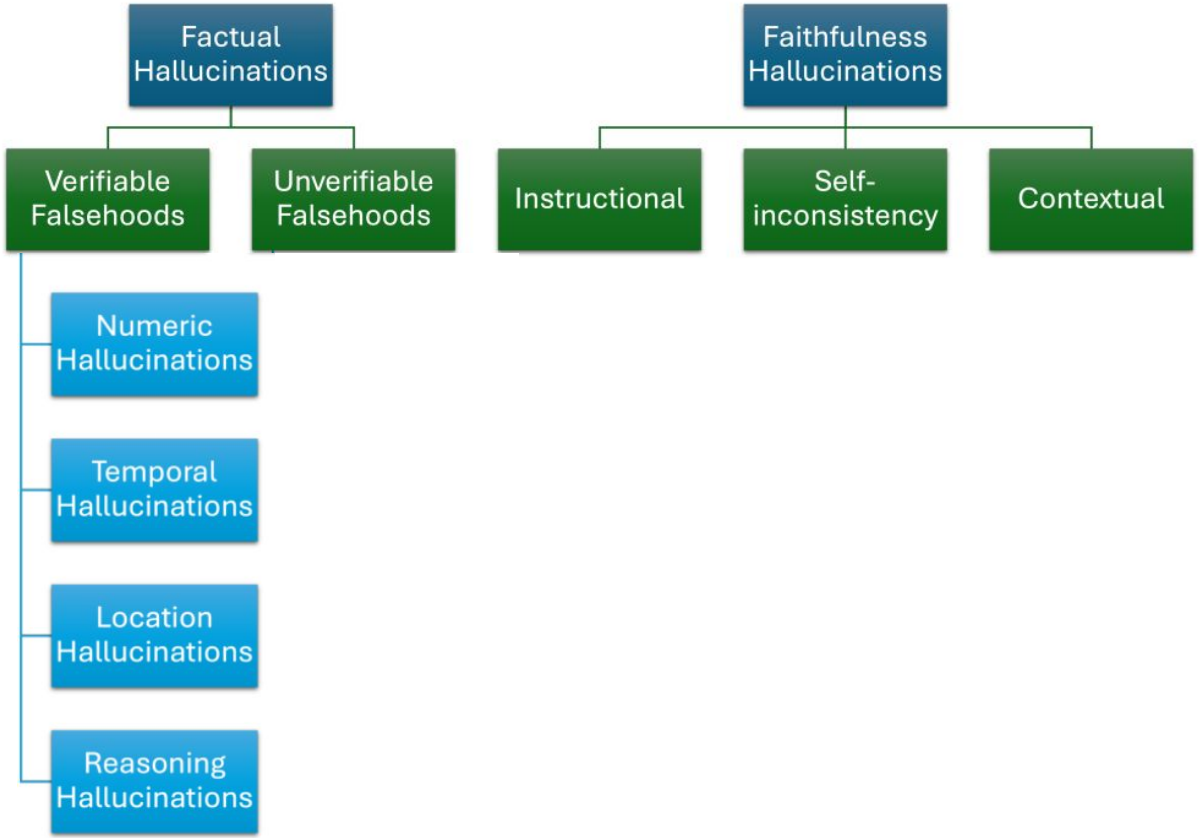
- Decontextualization Improvement
  - Refined prompts to fix the issue of some questions being too general and lacking context
- Clarity/Detail
  - Providing extra context when needed
- Fixed minor spelling and grammar errors
- Set a new record on SummEdits
  - >0.5 Entailment
  - >0.7 G-Eval



(a) G-Eval Scores for Summedits

Taxonomy of Hallucinations

- 2 initial classifications
- Factual
  - Incorrect claims/statements
- Faithfulness
  - Deviate from a given instructions, message history



Distribution of Hallucination Types

- Proposed taxonomy
- Better analysis for improving correction

Type	Category	Subcategory	RARR (%)	COVE (%)
Factual Hallucinations	1.1 Verifiable Falsehoods	1.1.1 Numeric Hallucinations	24	15
		1.1.2 Temporal Hallucinations	14	12
		1.1.3 Location Hallucinations	19	21
		1.1.4 Reasoning Hallucinations	28	37
	1.2 Unverifiable Falsehoods	-	15	16

Conclusion

Conclusion

- RQ1: What is an appropriate taxonomy for categorizing hallucinations in LLMs?

Type	Category	Subcategory	Example	Explanation
Factual	1.1 Verifiable Falsehoods	1.1.1 Numeric Hallucinations	Current population of Munich is **less than 1 million**.	Actual population is over 1.5 million.
		1.1.2 Temporal Hallucinations	Galatasaray won the UEFA Cup in **1999**.	They actually won in 2000.
		1.1.3 Location Hallucinations	Eiffel Tower is located in **Berlin**.	It is actually in France.
		1.1.4 Reasoning Hallucinations	If it rains, the ground gets wet; the ground is wet, so it **must have rained**.	There can also be another cause such as a spill.
	1.2 Unverifiable Falsehoods		Joe woke up at **7 am** today.	Simply can not be proven wrong, and does not have to be an imaginary concept.
Faithfulness	2.1 Instructional		User: "How can I make dairy-free **pasta** sauce?" LLM: "Use mozzarella to make a dairy-free **pizza** sauce."	It is not following the instructions and gives the recipe for the pizza instead of pasta.
	2.2 Self-inconsistency		Message 1: "Hamburger is a **healthy** food." Message 2: "Fast-food can be **harmful to health**."	LLM does not follow its own reasoning in multi-turn conversation — hamburger is a fast-food can not be healthy if it is harmful.
	2.3 Contextual		Context: "Tomatoes are widely used **fruits** in Mediterranean countries." LLM: Tomatoes are commonly used **vegetables** in the Mediterranean region.	LLM contradicts the context and uses vegetables instead of fruits. Real-world or common knowledge is not relevant here — context is what matters.

Conclusion

- RQ2: How can the Retrieval Augmented Generation (RAG) technique handle long contexts efficiently?
    - Using Atomic Parts
    - Using Search Snippets
    - Custom Knowledge Bases
    - Optimizing Connections
    - Advanced search techniques (e.g. Google Dorks)

Conclusion

- RQ3: How can we balance editing and faithfulness when improving generated text?
  - Ensure editing:
    - Use zero-shot
  - Keeping Relevant:
    - Avoid excessive refinement via text distance based rules
    - Adding Few-Shot Examples

Future Work

Future Work

- Leverage full datasets
  - ExpertQA
  - LongFact
- Optimization for industrial use cases
  - Small Language Models (SLMs)
  - Caching
  - Merging prompts
- Recent post-hoc fact-checkers
  - e.g., Factcheck-GPT

Ihsan Soydemir

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for
Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.umt.de
wwwmatthes.in.tum.de