AUTHOR

Borislav Kolev Marinov
B.K.Marinov-1@student.tudelft.nl

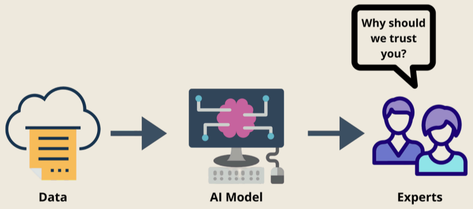
Evaluating Faithfulness of LLM Generated Explanations for Claims: Are Current Metrics Effective?

Analysing the Capabilities of Automatic Metrics to Represent the Difference in Faithfulness Between Explanations

SUPERVISORS

Pradeep Murukannaiah,
Shubhalaxmi Mukherjee

TU Delft



01. Introduction

- 🗒️ Data Access → Claims and evidence flow into LLMs
- 🧠 LLM → Produces outputs and explanations
  - Can sound convincing but be misleading
- 🤔 Experts ask → "Why should we trust this explanation?" → evaluation required
- 🎯 Our Goal
  - Evaluate if existing metrics (G-Eval, FactCC, UniEval, QAGs) truly capture faithfulness in fact-checking explanations

02. Research Question

How well do current evaluation metrics reflect the faithfulness of LLM-generated fact-checking explanations compared to journalist-written ones?

📘 Terminology

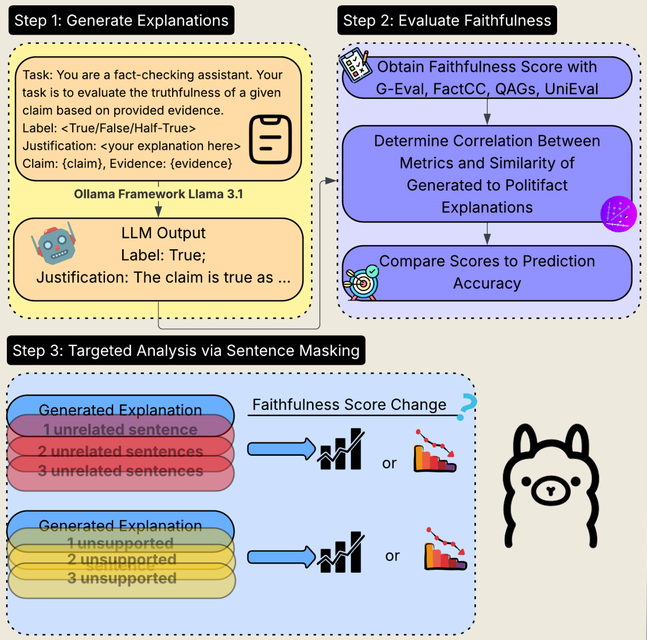
- 🗣️ Claim — Statement to be verified
- 🗂️ Evidence — Factual context for claim
- 🏷️ Label — Verdict (True / Half-True / False)
- 🔍 Faithfulness — Explanation aligns only with evidence (no extra or false info)

📊 Faithfulness Metrics

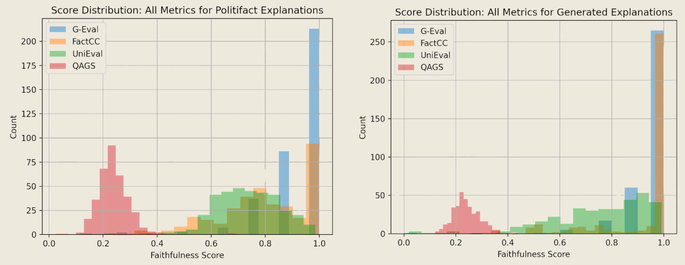
- 🧠 G-Eval — LLM grades explanations using reasoning prompts with form-filling and Chain-of-Thought
- 🧪 FactCC — Checks consistency with evidence using BERT
- ❓ QAGs — Converts explanation into questions, answers them with evidence
- ✅❌ UniEval — LLM uses boolean QA prompts to judge factual correctness
- Each metric gives a score between 0 (unfaithful) and 1 (faithful)

03. Methodology

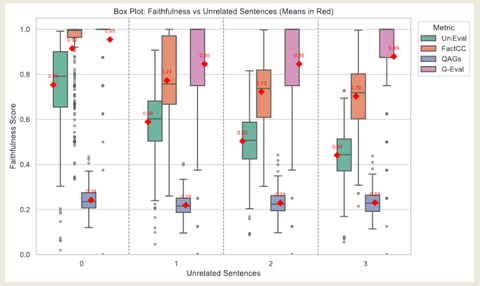
Datasets of Claims and Evidence:
- QuanTemp with Journalist explanations from PolitiFact
- HoVer with evidence from Wikipedia articles

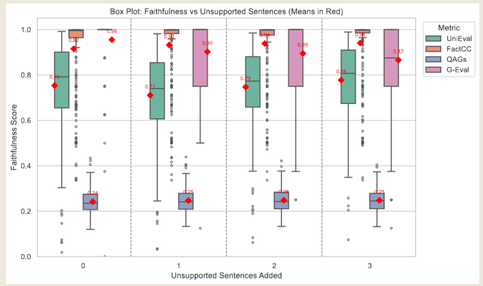


04. Results



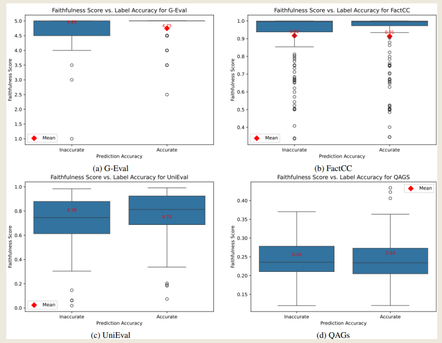
Metric Correlations: Metrics favor LLMs – Normal distribution of scores for expert, Normal distribution skewed towards higher scores for generated





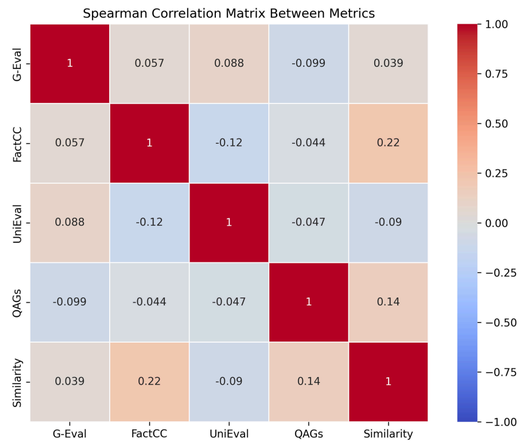
Targeted Tests: Some metrics detect hallucinations (FactCC, UniEval) 📉📉; others (G-Eval) fail to penalize unfaithful content consistently, while QAGs is robust to noise, but low in scores.

Accuracy to Fathfulness Scores: All metrics show slightly higher average scores for incorrect labels. Medians are higher for correct predictions, but not consistently.



05. Analysis

- 📊 Metric Correlation with Semantic Similarity
  - Weak correlations overall (ρ ≤ 0.23)
  - ✅ FactCC & QAGs: Best (though limited) alignment with expert explanations
  - ⚠️ UniEval & G-Eval: Low or even negative correlations
- 🧠 Score Bias
  - 🔺 All metrics assign slightly higher scores to incorrect predictions
  - 🤖 G-Eval shows the strongest bias favoring LLM outputs
- 🧪 Perturbation Sensitivity
  - 🧩❌ Unrelated Sentences: FactCC & UniEval show clear score drops
  - ⚠️🗒️ Unsupported Sentences: Metrics often fail to detect hallucinations, though some score reduction observed



06. Conclusion

- Existing metrics are unreliable and inconsistent
- Different metrics perform better under the experiments, but no metric consistently reflects faithfulness
- Have some positive aspects that show latent potential
- **Further Work**
  - 💡 Use more capable LLMs (e.g., GPT-4) for generation and evaluation.
  - 🛠️ Fine-tune metrics on the actual dataset (e.g., PolitiFact QuanTemp) to improve alignment.
  - 🧪 Test new perturbations — insert noise or hallucinations into the middle of explanations.