Comprehensive Entailmennt Reasoning For Hallucination Detection

Ron Eliav, Arie Cattan, Eran Hirsch, Shahaf Bassan,
Elias Stengel-Eskin, Mohit Bansal, Ido Dagan

Main Contirbutions

- **Instruct LRMs to reason in a semantics-grounded way**
- **Evaluate reasoning steps**

Motivaton

Hallucination Detection

- Hallucination detection is commonly approached through NLI
- Most common approaches uses either a FT model or direct prompting of an LLM.
- Entailment classification inherently requires reasoning

Reasoning Models

- LRMs exhibit impressive reasoning capabilities.
- However, users typically do not control or guide the reasoning steps.

Solution – LACER

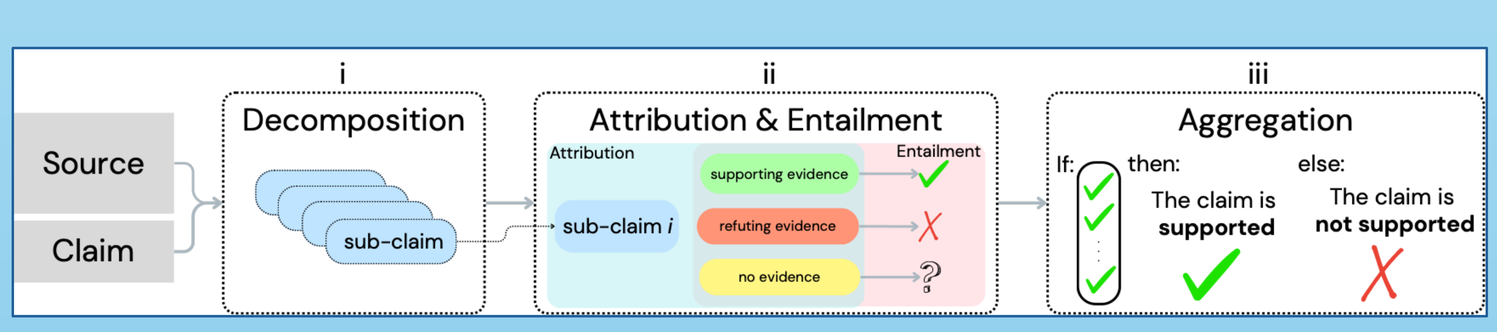
Localized & Attributed Comprehensive Entailment Reasoning

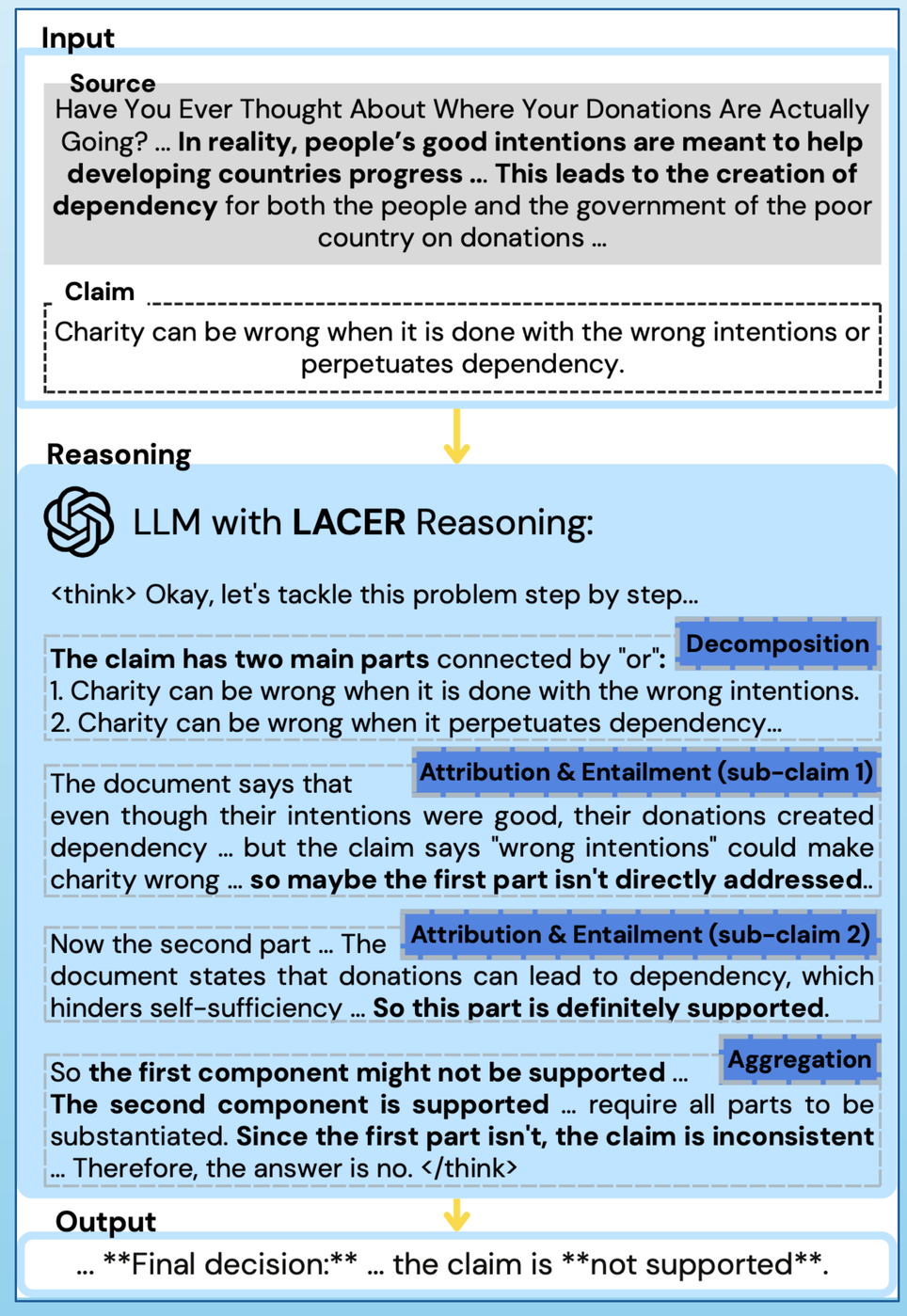
Instruct the model to reason through three steps:

(i) Decomposition
Break the hypothesis into sub-claims.

(ii) Attribution & Entailment
For each sub-claim, find supporting or refuting evidence in the source (if available), and classify it as *entailed*, *contradicted*, or *neutral*.

(iii) Aggreagtion
If all sub-claims are entailed, the claim is **supported**;
otherwise, it is **not supported**.





Metrics

For a decomposition of $\mathcal{C}$ into $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$,

$$\boldsymbol{A_{ttomicity}} := |\mathcal{H}|$$

$$\boldsymbol{S_{oundness}} := \frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathbb{1}\{h_i \; is \; sound\}$$

$$\boldsymbol{C_{ompleteness}} := \begin{cases} 1 & if \; \mathcal{H} \subseteq \bigcup_i h_i \\ 0 & otherwise \end{cases}$$

$$\boldsymbol{A_{ttribution}} := \frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathbb{1}\{h_i \; is \; correctly \; attributed\}$$

$$\boldsymbol{E_{ntailment}} := \frac{1}{|\mathcal{H}|}\sum_{i=1}^{|\mathcal{H}|} \mathbb{1}\{\hat{y}_i = y_i\}$$

Experiment Setup

Domains and Datasets:

- Fact Verification (ClaimVerify)
- Question Answering (LFQA-Verification)
- Summarization (TofuEval)

Models:

- Large Reasoning Models (LRMs)
  - DeepSeek-R1, O4-mini, Gemini-2.5-Pro, QWQ-32B-Preview
- Standard LLMs
  - DeepSeek-V3, GPT-4o-mini, Gemini-2.0-Flash, Qwen-Plus

Instructions

- **Baeline**:
  - No explicit instruction; model generates reasoning freely.
- **QA-based**
  - Generate questions and compare the claims' answers to the sources' answers.
- **Ours**:
  - Decompose -> Attribution & Entialment for each fact -> Aggregation

Insights

Manual Analysis (240 samples)

☑ Our approach improves performance in hallucination detection.

☑ Atomicity values remain relatively low.

☑ There is a tradeoff between atomicity and completeness.

☑ The attribution step often occurs even without explicit instruction.

☑ Models are nearly perfect in soundness and aggregation.

☑ Higher attribution correlates with higher accuracy.