

FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence

Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Louisa Göke, Monika Coers, Wei Xu, Byron C Wallace, Junyi Jessy Li



Northeastern University



PubMed
Original Abstract:



Plain Language Summary

Evidence Inference

Span from abstract:

3'AmNic-rEPA recipients with the highest serum anti-nicotine antibody response (top 30% by AUC) were significantly more likely to attain 8 weeks continuous abstinence from weeks 19 through 26 than the placebo recipients (24.6% vs. 12.0%, $p=0.024$, $OR=2.69$, 95% CI, 1.14–6.37).

Rationale: This result is somewhat vaguely mentioned by saying "who took the vaccine and produced lots of these anti-nicotine antibodies were more successful in quitting smoking".

Vague/Slightly inaccurate

Population

The population is mentioned, but **described inaccurately or missing critical descriptors**

Rationale: In multiple places the summary mentions "people" not smokers as those receiving the vaccine.

Intervention

The outcome is mentioned, and **described accurately**

Rationale: The intervention is correctly explained as a vaccine to help people stop smoking.

Comparator

The comparator is **missing in the model summary**

Rationale: The comparator "placebo" is not mentioned in the summary.

Added Information

creating antibodies against nicotine, the substance in cigarettes that causes addiction

Factual

Rationale: This is part of what "hemodynamic fluctuation" but does not cover the whole spectrum of hemodynamic.

a good way to help people stop smoking.

Factual

Rationale: A true fact, yet the abstract rather concludes that "This study demonstrates proof-of-concept that 3'AmNic-rEPA elicits antibodies to nicotine and is associated with higher continuous abstinence rates"

Outcome

The outcome is mentioned, and **described accurately**

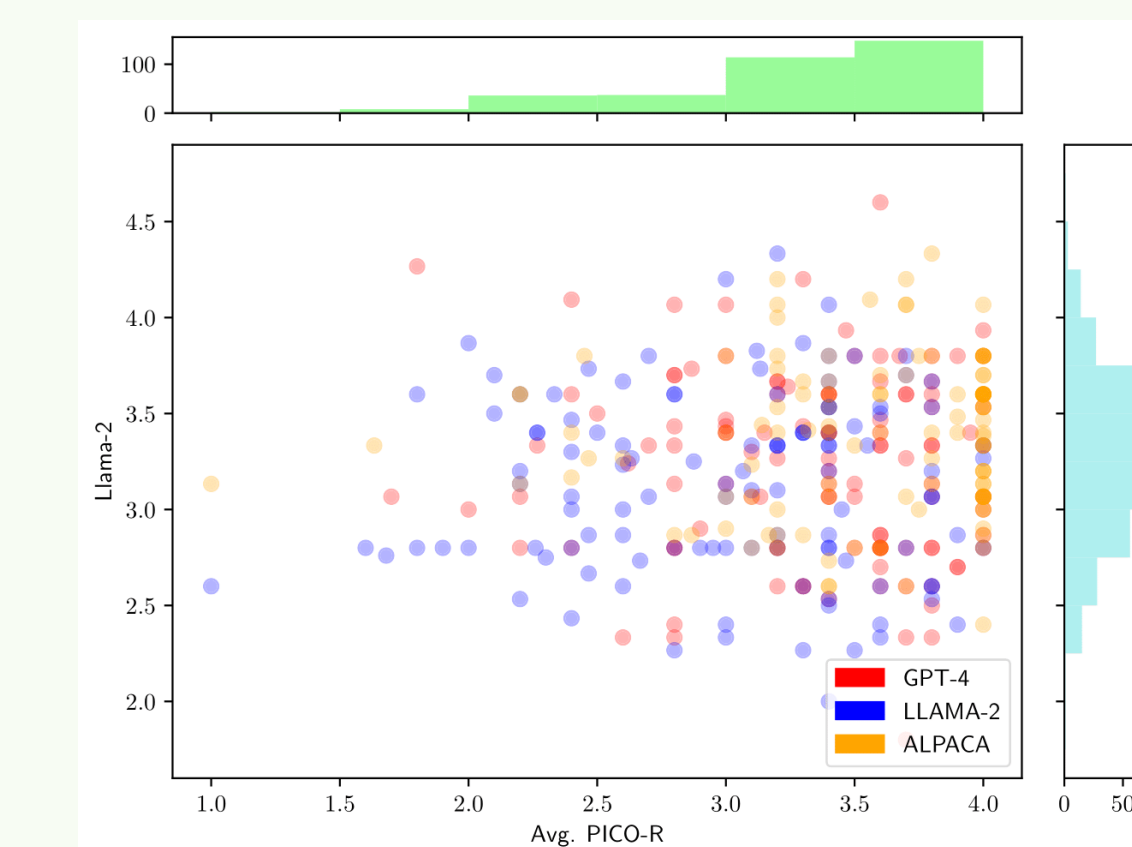
Rationale: The outcomes are correctly mentioned as number of created antibodies and abstinence rate.

This passage is about a vaccine called NicVAX that's designed to help people quit smoking. The vaccine works by creating antibodies against nicotine, the substance in cigarettes that causes addiction. The study found out that people who took the vaccine and produced lots of these anti-nicotine antibodies were more successful in quitting smoking. They also found that giving a bigger dose of the vaccine and giving it 5 times over 6 months was the best way to use it. So, the study shows that the vaccine could potentially be a good way to help people stop smoking.

Factuality Metric Evaluation

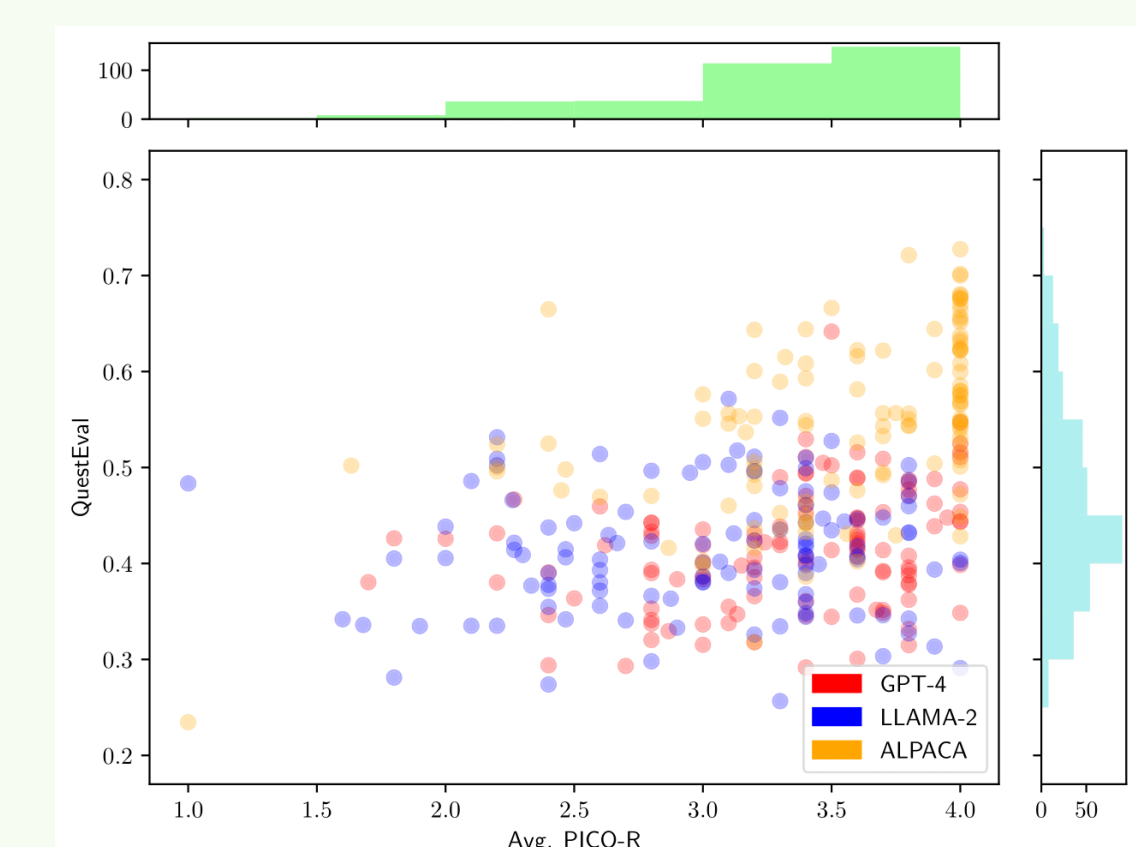
- We compared FactPICO ratings against other factuality metrics, including LLM-based evaluations with rationales modeled after the FactPICO framework.
- At a **system-level**, most metrics produced similar scores to those from experts.
- At an **instance-level**, scores from dedicated factuality models were more correlated with FP ratings compared to LLM-based evaluations, except for GPT-4 based evaluations.

Llama-2-Chat

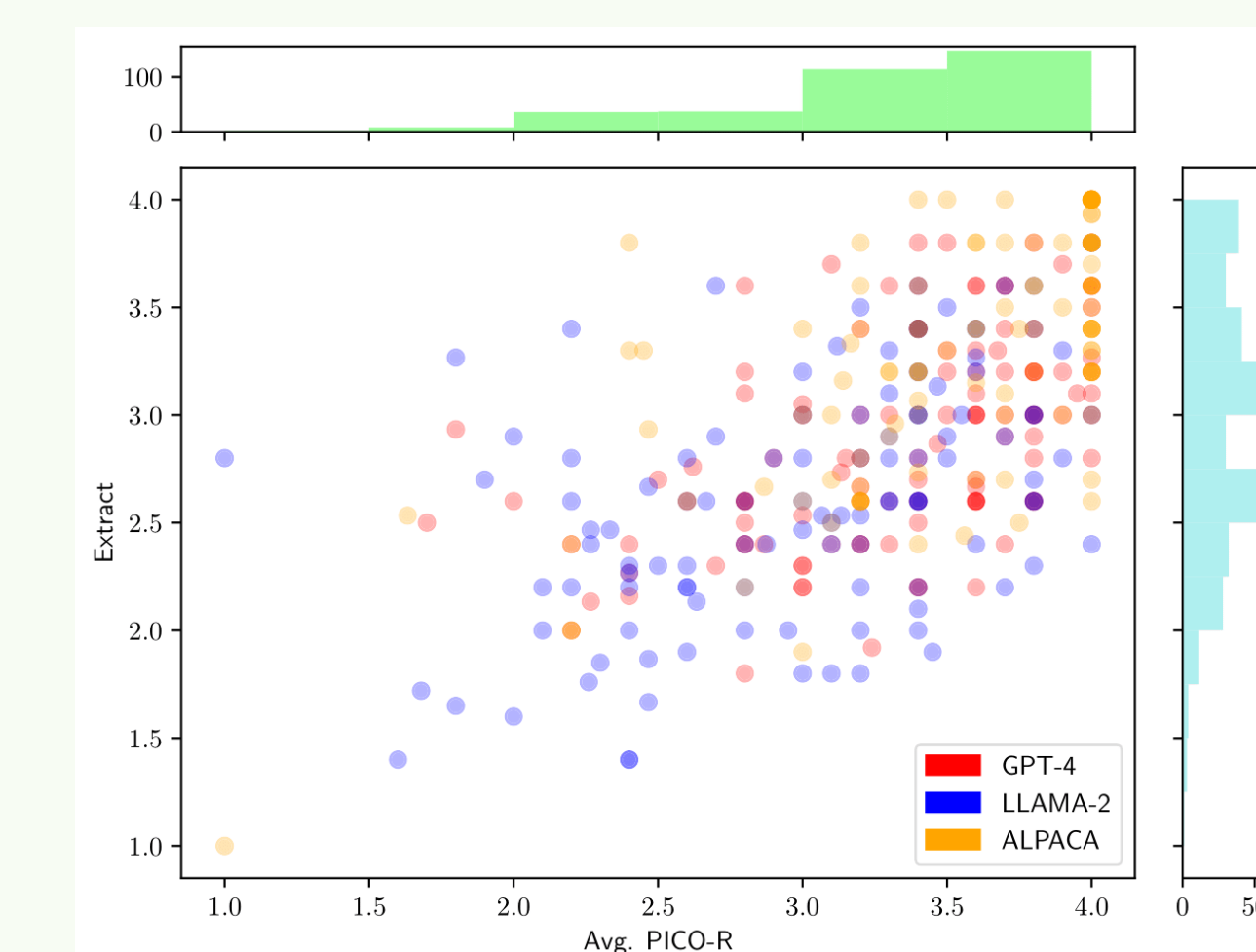


$\rho = 0.080$

QuestEval

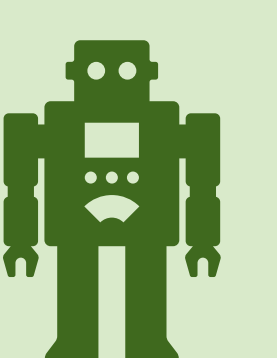


$\rho = 0.412$



$\rho = 0.633$

Most Correlated:
GPT-4 + PICO-R
Extraction



Mistral (7B)

Can Large Language Models (LLMs) correctly simplify key characteristics in medical texts?

We address this question by building **FactPICO**, an expert-annotated, fine-grained, factuality benchmark for the plain language summarization of technical abstracts describing randomized controlled trials (RCTs).

Experts evaluate key characteristics of medical trials: **PICO elements** (Population, Intervention, Comparator, Outcome), and **Evidence Inference**. FactPICO also includes an assessment of **added content**, and **expert-written rationales** to contextualize the evaluation.

Evaluation of LLM Simplifications

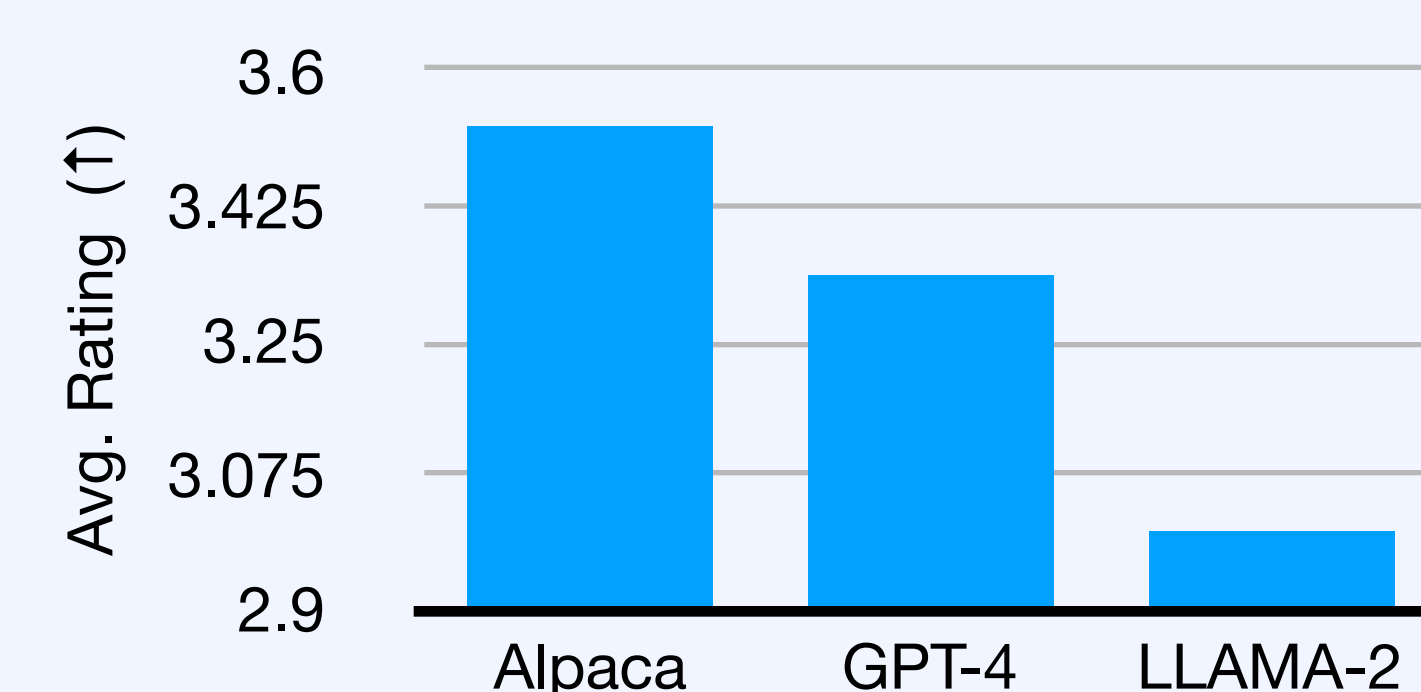
- We evaluated plain language summaries (PLSs) from: **Alpaca** (7B), **Llama-2-Chat** (7B), and **GPT-4**.
- Numerical ratings: 1 (missing) - 4 (accurately described)
- PLSs frequently made factual errors regarding *critical* RCT elements.
- We also noticed a tradeoff between factuality and simplicity.



human expert

It is mentioned that the patients included all (here called vaguely "people") had a kidney transplant, the size of population is missing.

Factuality



Simplicity

