Pragmatic Evaluations of Large Language Models Deployed in Hospital Workflows: Two Examples from the Children's Hospital of Philadelphia

Ashley Oliver[1], Dhineshvikram Krishnamurthy[1*], Hojjat Salmasian[1,2], Abdul Tariq[1]*

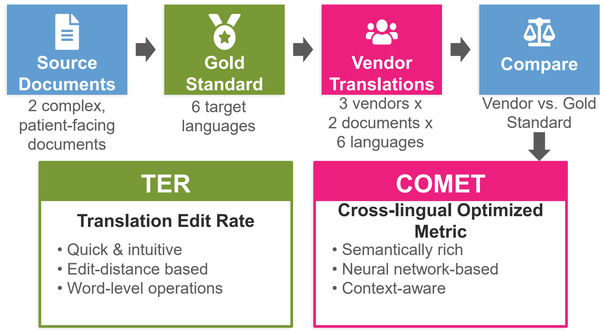
[1]Children's Hospital of Philadelphia; [2]University of Pennsylvania; [*]Equal contributions to this work

INTRODUCTION

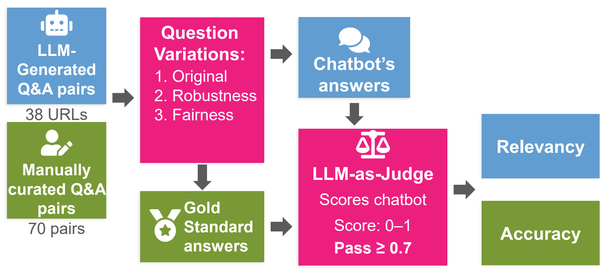
- Large Language Models (LLMs) are increasingly embedded in hospital workflows.
- Decisions to deploy these tools are often based on vendor claims rather than real-world performance.
- Health systems need **pragmatic, resource-aware evaluation approaches** that inform deployment decisions.
- We present two real-world evaluations to illustrate how LLM tools can be assessed in practice.

METHODS

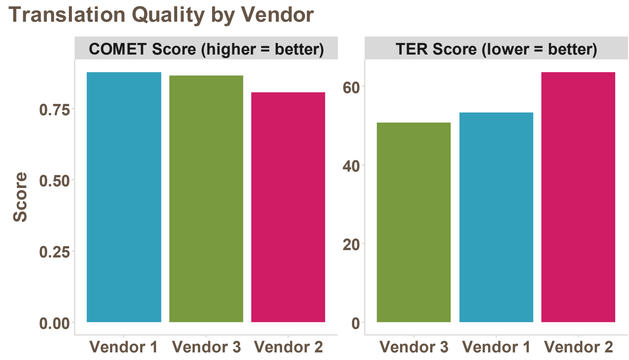
LLM-based translation platform

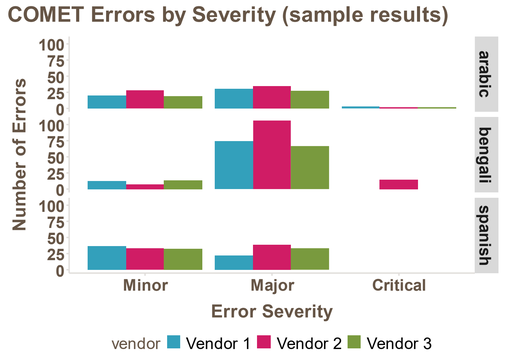


Patient-facing web chatbot



LLM-BASED TRANSLATION PLATFORM

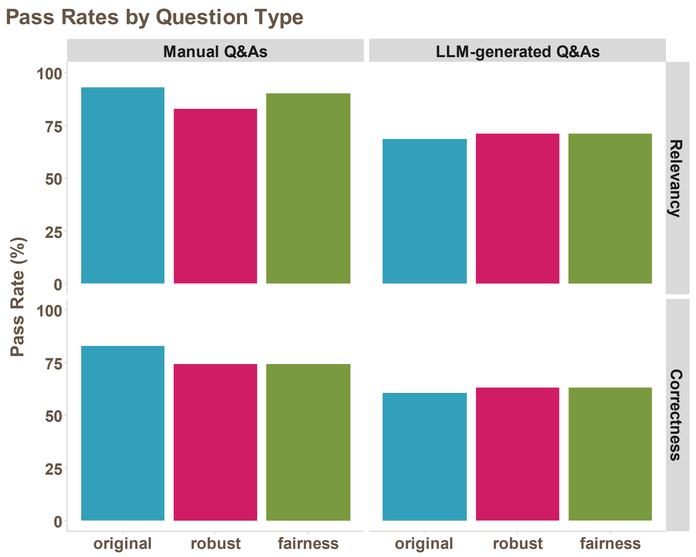




- Overall average scores across all languages and documents showed vendor performance at a high level, with one vendor (Vendor 2) scoring worst on both metrics
- COMET's error severity identification illustrated disparities in vendor performance across languages

PATIENT-FACING WEB CHATBOT

- Chatbot performed better for Q&As programmed in its knowledge base (Manual Q&As) than new Q&As (LLM-generated)
- Slight performance differences in Manual Q&As across question types indicate some degradation with re-worded questions



DISCUSSION

Our Evaluation Approach

1. **Automated & Scalable**
   Used cloud-based automated approaches that scale to larger datasets
2. **Gold Standard Comparison**
   Model outputs assessed against established human curated benchmarks
3. **Operationally Actionable**
   Results structured to directly support decision-making

KEY TAKEAWAYS

! Evaluate Independently

Real-world performance of LLM-based tools is consistently **lower than vendor claims.**

↕ Start Where You Can

Evaluations can be conducted with **varying levels of rigor and intensity** based on available resources.

✓ Even Limited Evaluations Add Value

They reveal issues that can be mitigated through workflow guardrails and provide feedback to vendors.

Children's Hospital of Philadelphia®
Center for Healthcare Quality & Analytics