LLMs generate survey questions with quality comparable to those used in practice

—

prompting matters more than model (size)

AI for Survey Design: Generating and Evaluating Survey Questions with LLMs

Anna Fuchs, Anna-Carolina Haensch, Wiebke Weber

What we did

Study design: 5 LLMs + 3 prompting strategies + 4 survey domains + 15 concepts of interest = 900 LLM-generated survey items (612 valid)

Evaluation: Survey Quality Predictor (SQP 3.0)

What we find

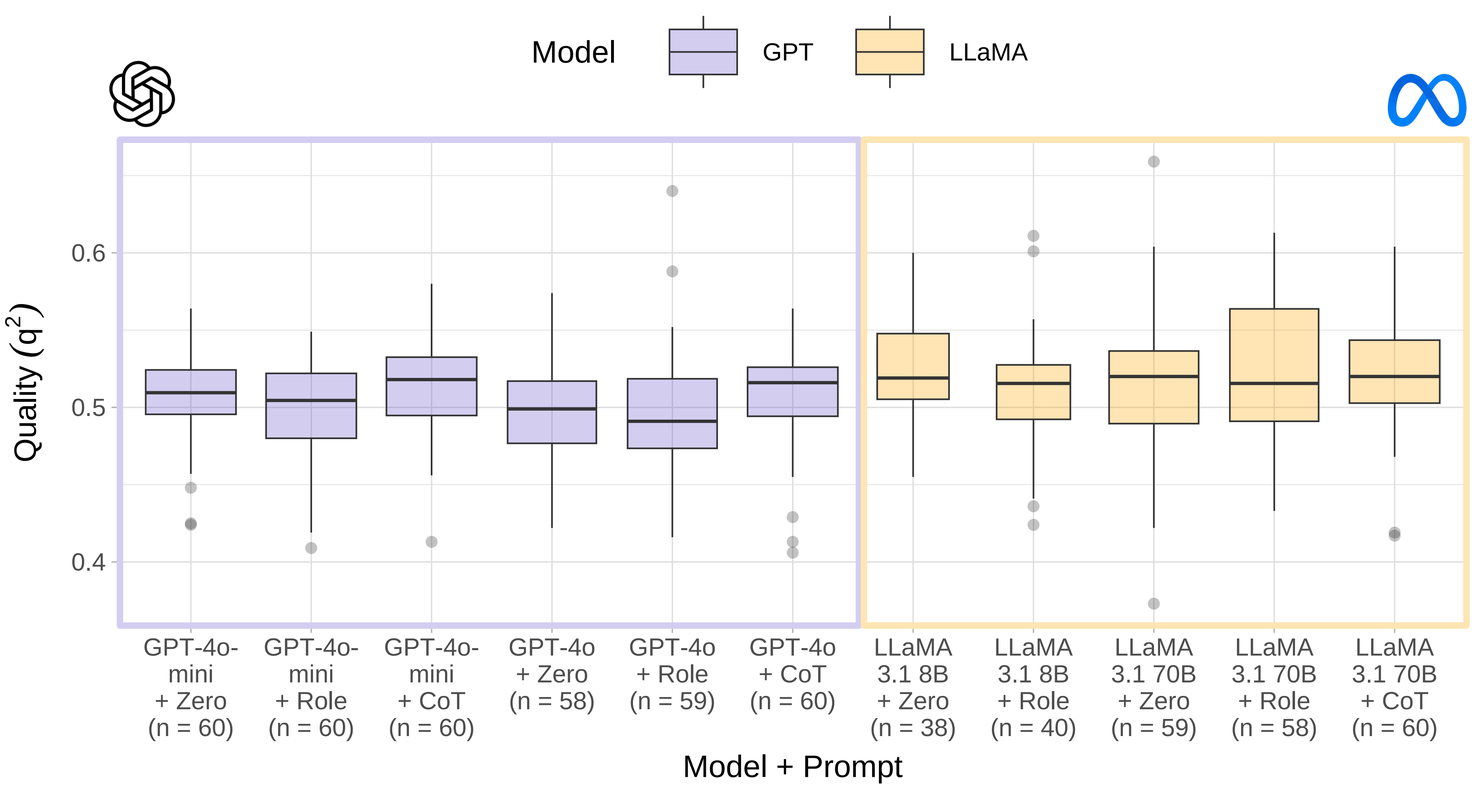


Figure: Quality of generated survey items across pre-trained LLMs. Measurement quality can theoretically range from 0, i.e., not measuring the concept at all, to 1, i.e., measuring the concept perfectly. In practice, in SQP the quality ranges from 0.2 to 0.7 with an average of 0.58.

➡ LLMs can be used as assistive tools in survey design.

➡ **Small GPT** and **large LLaMA** model perform best.

➡ **Chain-of-Thought** prompting achieves the highest quality.

➡ Item characteristics are consistent across GPT models but vary with LLaMA size and prompting.

➡ Open-source GPT and small LLaMA model fail to produce valid results.

SQP requires coding linguistic and formal survey item characteristics

Example of survey item characteristics from a total of ~60 automated + manually coded characteristics

26 words in the request | Theoretical range of the concept is bipolar | Interrogative question | Introduction text present | 11 words in the introduction text

The next few questions are about your views of other people.

A9 Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?

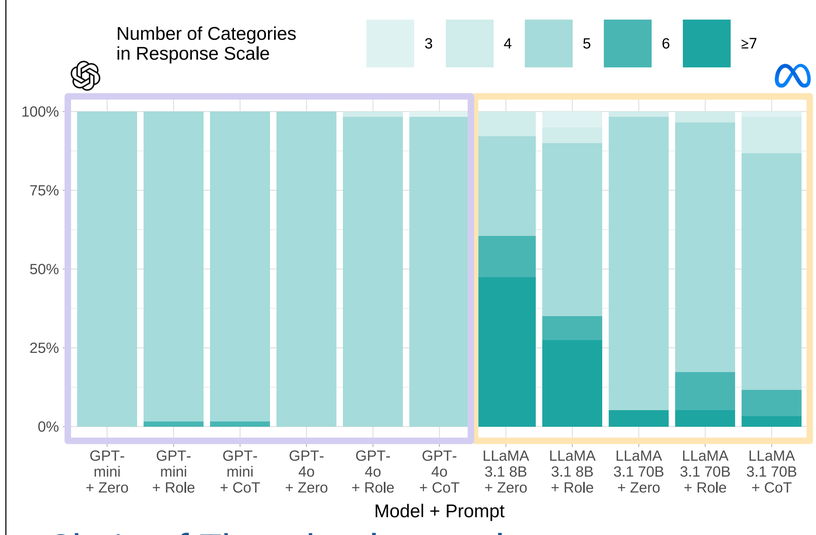
You can't be too careful — Most people can be trusted

1 — 2 — 3 — 4 — 5

5 categories | No 'Don't know' option present | Horizontal scale | Complete sentence labels | Neutral category present | Partially labeled categories

Example characteristics illustrating differences across models and prompting techniques

GPT models show a clear tendency toward 5-category scales



Chain-of-Thought almost always presents an introduction

Prompt	Introduction text present
CoT	90%
Role	33%
Zero	12%

Contact:
anna.fuchs@stat.uni-muenchen.de

Social Data Science + AI Lab | LMU Ludwig-Maximilians-Universität München | Munich Center for Machine Learning | KODAQS | University of Maryland