



**IDM** INSTITUTE FOR  
DISEASE MODELING

## **Pressure-Testing and Prototyping AI Tools for Qualitative Analysis of Free Text Data: A Case Study on DRC Vaccination Surveys**

Roy Burstein (Gates Foundation/IDM), Eric Mafuta (KSPH), Joshua L. Proctor (Gates Foundation/IDM)

*December 2025. Better Data for Better Jobs and Lives: Innovations in Survey Measurement in the Age of AI*

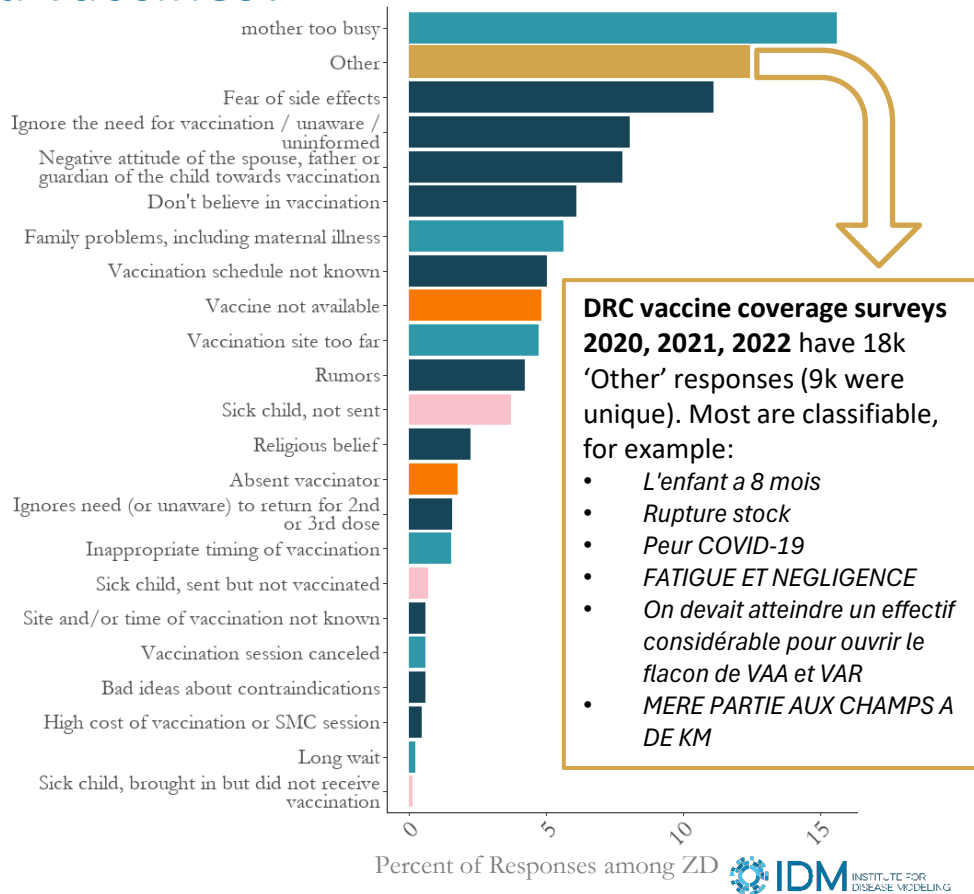


Ecole de Santé Publique de Kinshasa  
Kinshasa School of Public Health | KSPH  
UNIVERSITE DE KINSHASA

**Gates  
Foundation**

# Use case: Understanding barriers to vaccination. ‘Why has the child not received all the recommended vaccines?’

- Priority for immunization programs to better understand **why** children are not vaccinated (Zero-Dose = ZD).
- In surveys, simply asking *why* leads to many ‘Other: free-text’ responses. **This messy data** is often left untouched by busy researchers.
- Data from three **vaccine coverage surveys in DRC**.
- We ask: **Can we trust LLMs to do the job of categorizing** these responses for us?
- Precedence for this in NLP **topic modelling**.
  - As of late 2024: Only one paper so far in LLM space. (Mellon, et al. 2024). Done on English data on popular topic.
- We **tested various approaches, reanalyze the DRC survey data**, and comment on broader implications.



# Benchmarking and validating LLM workflows for assigning free-text responses to categories

- ChatGPT via the browser is a great way to quickly test hypotheses dropping in response data as a formative step, but accuracy needs to be more **rigorously tested** and compare across multiple approaches.
- **Manually benchmarked** 1000 responses
- **Tested three main analytic approaches**
  - Natural Language Processing (**NLP**): **Semantic Embeddings**, unsupervised clustering, and GPT-4o to summarize
  - “Just ask”/**Prompt-engineering**: Zero and Few Shot Learning, Chain of Thought (CoT)
  - **Fine-tuning** GPT-4o to this task
- **OpenAI** API functionality to control model requests for reproducibility and testing
- Assess **accuracy** scores between human intelligence (HI) and artificial intelligence (AI). **Can we rely on AI to do what a researcher (with a lot of time) would do?**

## Example Prompt:

You are an assistant helping to categorize responses from parents that have decided to not vaccinate their children.

The survey that was given to these parents had pre-defined categories for why they have decided to not vaccinate their children.

The following are the categories:

[1] Vaccination site too far

[2] Vaccination schedule not known

[3] mother too busy

...

[42] Issues with health workers (negative experience, no reminder, and distrust)

[43] No card to remind of appointment or dates not on card

Your role as the assistant is to assess the unstructured responses from the survey of parents.

If a reason in the response likely corresponds to one of the categories, provide only the category number

If a response does not fit into other categories and a new category needs to be created, provide the number 30

Only output one category number matching the most relevant category.

Here are two examples:

user: We don't live in this country.

assistant: 30

user: I could not find the time.

assistant: 3

# Accuracy increases with model complexity, engineering time, and researcher-effort to label examples for supervised approaches

- **Few-shot and fine-tuned** models have **comparable** accuracy but very different per-call token usage.
- Accuracy gains in few-shot **diminish** with more examples: 0->50: +8%, 50->800: +4%
- Explored ways to bring costs down via batching, but hallucinations and errors increased with bigger prompts
- **Costs** estimates are specific to this problem and are constantly **decreasing** (i.e., costs halved in the process of this project already!).
- Allows us to **deep-dive on differences** between HI and AI,

\* Total cost assuming 12k responses, per this specific dataset

\*\* Researcher Level of Effort assumes workflow is in place, as such future researcher cost primarily comes from labelling training examples

Approach	Precise Accuracy	Avg. prompt tokens/ API call	Upfront Cost (\$) (Oct 2024)	Avg. Cost (\$) / API call	Estimated Total Cost (\$) * (Oct 2024)	Researcher level of effort**
NLP + Clustering (n=200)	61.5%	605	0.0003	0.0015	0.30	Low (unsupervised)
GPT-4o Zero Shot	71.5%	398	0	0.0010	9.91	Low (unsupervised)
GPT-4o Chain of Thought	73.5%	472	0	0.0017	16.77	Low (unsupervised)
GPT-4o Few Shot (n = 20)	76.0%	621	0	0.0016	15.41	Medium
GPT-4o Few Shot (n = 50)	79.5%	943	0	0.0024	23.36	Medium
GPT-4o Few Shot (n=400)	81.5%	4486	0	0.0112	110.73	High
GPT-4o Few Shot (n = 800)	83.0%	8464	0	0.0212	208.84	Highest
GPT-4o Fine-tuned (n = 800)	83.5%	398	13	0.0010	22.91	Highest

# Accuracy increases with model complexity, engineering time, and researcher-effort to label examples for supervised approaches

- **Few-shot and fine-tuned** models have **comparable** accuracy but very different per-call token usage.
- Accuracy gains in few-shot **diminish** with more examples: 0->50: +8%, 50->800: +4%
- Explored ways to bring costs down via batching, but hallucinations and errors increased with bigger prompts
- **Costs** estimates are specific to this problem and are constantly **decreasing** (i.e., costs halved in the process of this project already!).
- Allows us to **deep-dive on differences** between HI and AI,

\* Total cost assuming 12k responses, per this specific dataset

\*\* Researcher Level of Effort assumes workflow is in place, as such future researcher cost primarily comes from labelling training examples

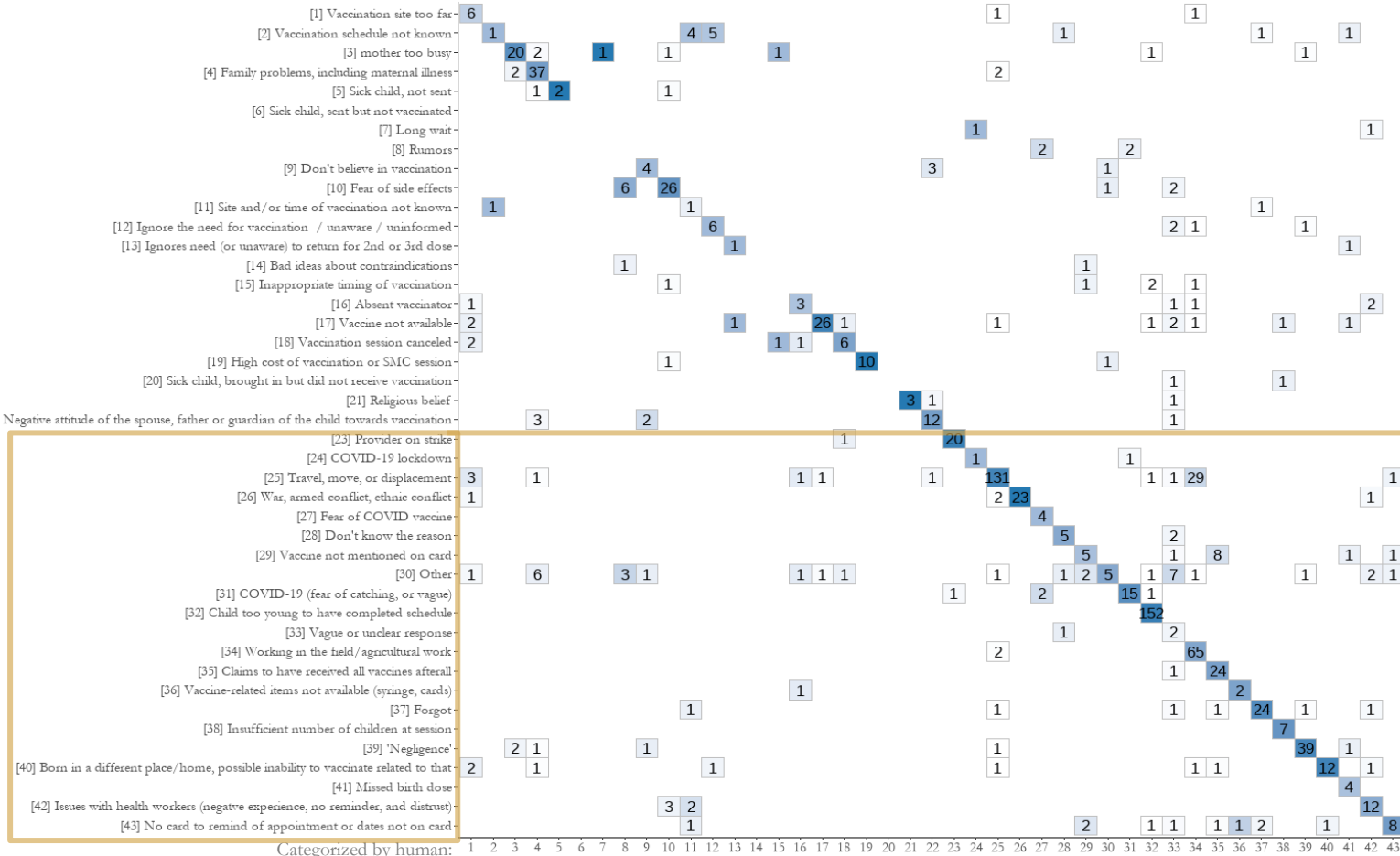
Approach	Precise Accuracy	Avg. prompt tokens/ API call	Upfront Cost (\$)	Avg. Cost (\$) / API call	Estimated Total Cost (\$)*	Researcher level of effort**
NLP + Clustering (n=200)	61.5%	605	0.0003	0.0015	0.30	Low (unsupervised)
GPT-4o Zero Shot	71.5%	398				(unsupervised)
GPT-4o Chain of Thought	73.5%	472				
GPT-4o Few Shot (n = 20)	76.0%	621				
GPT-4o Few Shot (n = 50)	79.5%	943				
GPT-4o Few Shot (n=400)	81.5%	4486	0	0.0112	110.73	High
GPT-4o Few Shot (n = 800)	83.0%	8464	0	0.0212	208.84	Highest
GPT-4o Fine-tuned (n = 800)	83.5%	398	13	0.0010	22.91	Highest

November 2025 Update

Approach	Precise Accuracy
<b>GPT-5.1 Zero Shot (no reasoning)</b>	<b>77%</b>

# A closer look at misclassifieds highlights nuances in comparing AI with HI

Categorized by LLM (GPT4 FS50):



Categorized by human:

Showing 950 untrained responses from the Few Shot (50) example.

Box darkness represents proportion of column-wise (HI) category assigned to each row (AI) category. Some columns are quite rare.

Categories 23-43 were not in the original survey question.

Strong diagonal indicates good overall agreement.

Some off-diagonal are true errors, but many are plausible.

# A closer look at misclassifieds highlights nuances in comparing AI with HI

Categorized by LLM (GPT4 FS50):



**AI: Fear of Side Effects**

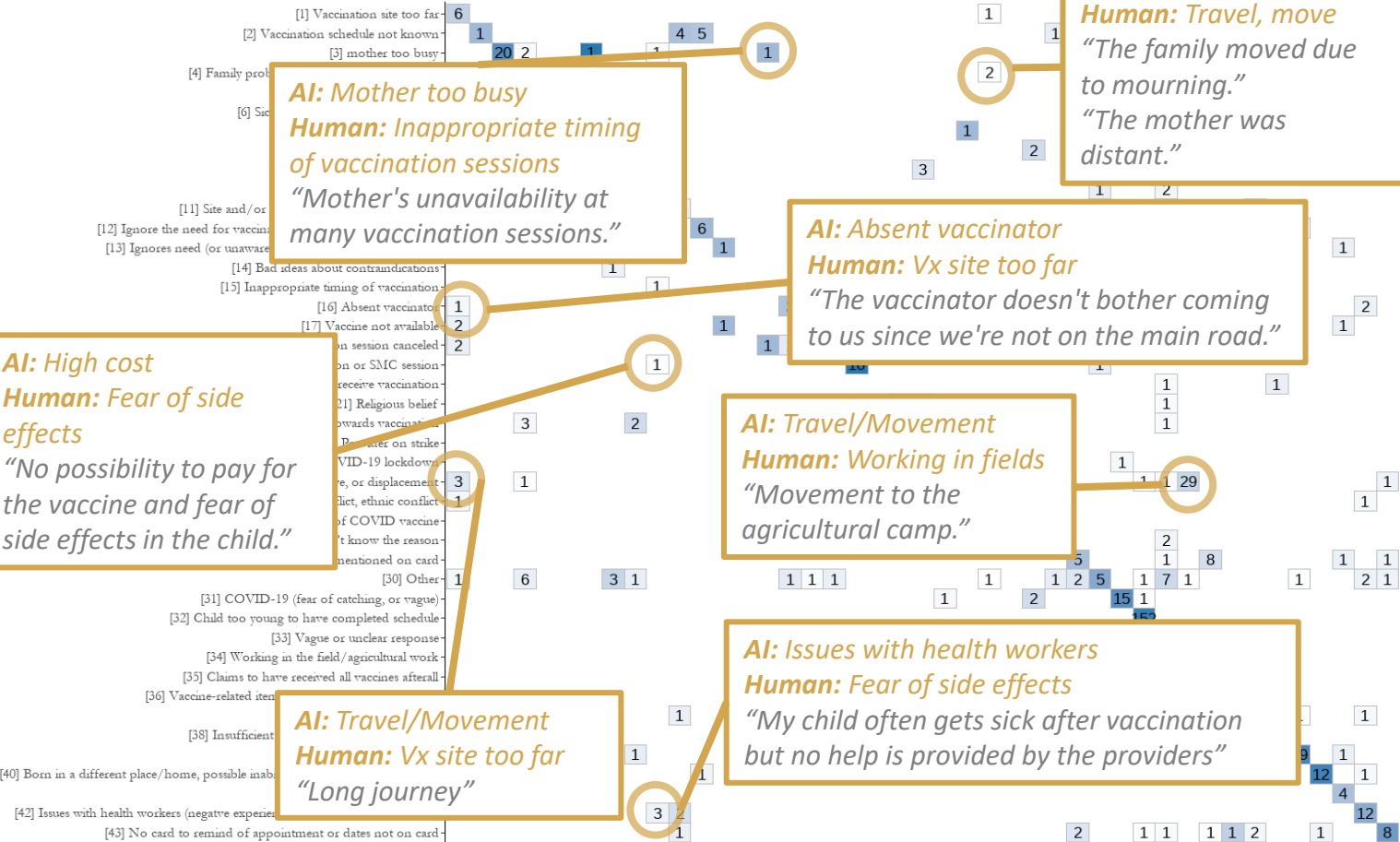
**HI: Rumors**

*Human interpreted as [8] Rumors while AI interpreted as [10] fear of side effects. Most of these hinge on the the interpretation of the term diseases/maladies in this context. Responses (translated):*

- "Vaccines cause diseases"
- "Vaccines cause disease in our children"
- "Vaccines give diseases to children"
- "I do not want to vaccinate my daughter because vaccines transmit diseases and I do not have the money to pay for medical care."
- "Injectable vaccines cause diseases in children, which is why I do not give them."
- "Contains Poison"

# A closer look at misclassifieds highlights nuances in comparing AI with HI

Categorized by LLM (GPT4 FS50):



**AI:** Mother too busy  
**Human:** Inappropriate timing of vaccination sessions  
 "Mother's unavailability at many vaccination sessions."

**AI:** Family problems  
**Human:** Travel, move  
 "The family moved due to mourning."  
 "The mother was distant."

**AI:** Absent vaccinator  
**Human:** Vx site too far  
 "The vaccinator doesn't bother coming to us since we're not on the main road."

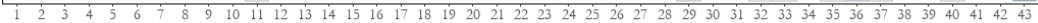
**AI:** High cost  
**Human:** Fear of side effects  
 "No possibility to pay for the vaccine and fear of side effects in the child."

**AI:** Travel/Movement  
**Human:** Working in fields  
 "Movement to the agricultural camp."

**AI:** Issues with health workers  
**Human:** Fear of side effects  
 "My child often gets sick after vaccination but no help is provided by the providers"

**AI:** Travel/Movement  
**Human:** Vx site too far  
 "Long journey"

Categorized by human:

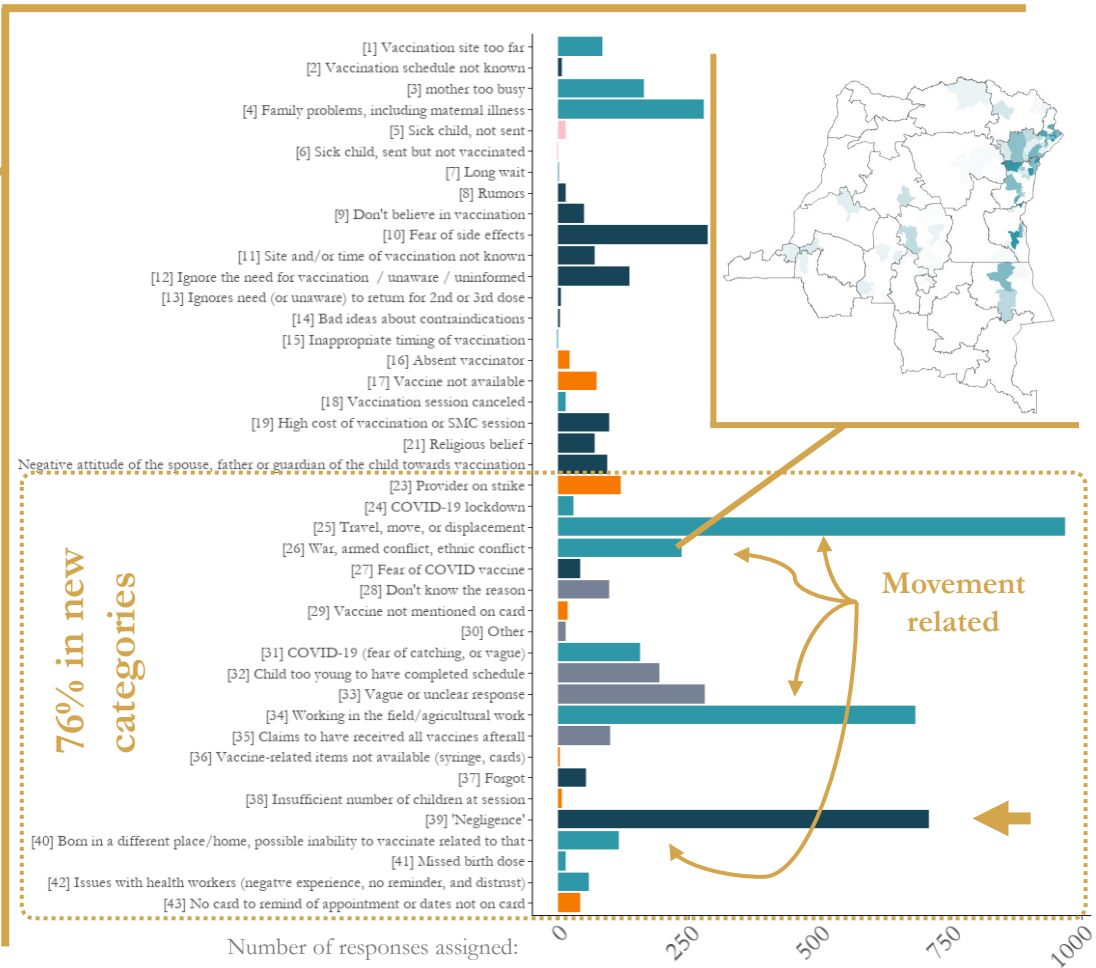
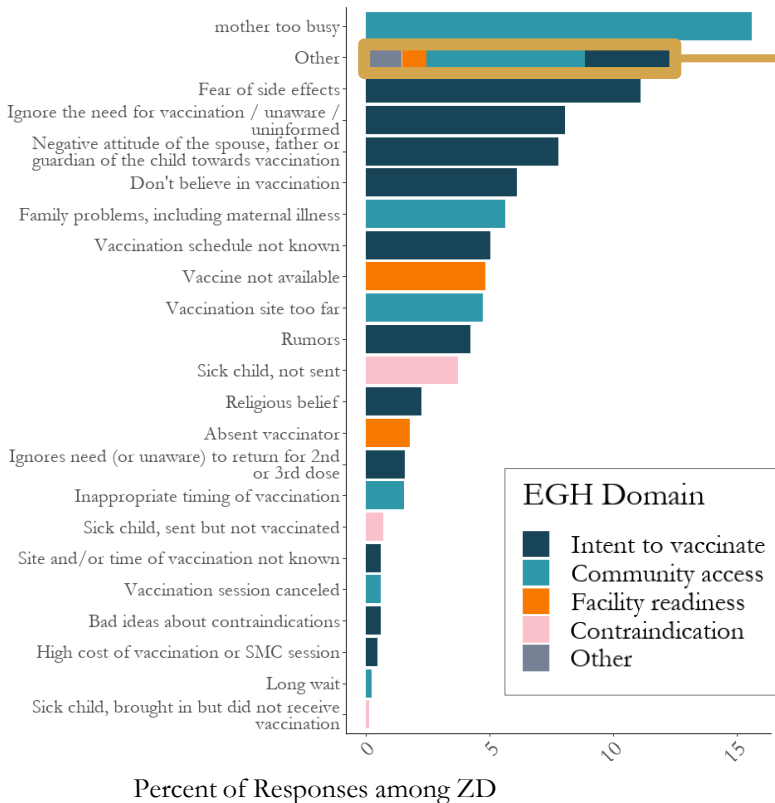


# Challenging the 'misclassified' observations yields higher potential accuracy

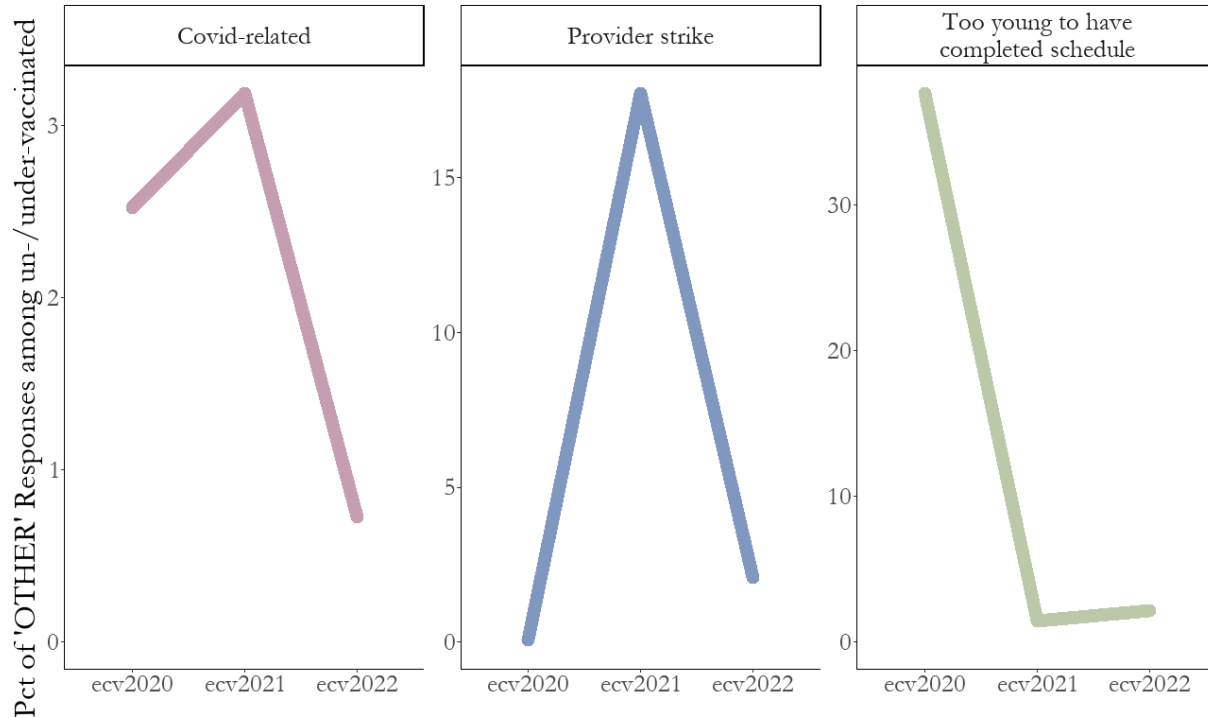
- Marking all plausible classifications yields an **'accuracy ceiling' as high as 96%** for the best approach
- As few as 20 few-shot examples achieves 93%.
- A single human coder with strict assignment rules is not a gold-standard.

Approach	Precise Accuracy	Accuracy Ceiling	Estimated Total Cost (\$)	Researcher level of effort
NLP + Clustering (n=200)	61.5%	74.5%	0.30	Low (unsupervised)
GPT-4o Zero Shot	71.5%	87.0%	9.91	Low (unsupervised)
GPT-4o Chain of Thought	73.5%	89.5%	16.77	Low (unsupervised)
GPT-4o Few Shot (n = 20)	76.0%	93.0%	15.41	Medium
GPT-4o Few Shot (n = 50)	79.5%	93.0%	23.36	Medium
GPT-4o Few Shot (n=400)	81.5%	92.5%	110.73	High
GPT-4o Few Shot (n = 799)	83.0%	95.0%	208.84	Highest
GPT-4o Fine-tuned (n = 799)	83.5%	96.0%	22.91	Highest

# Applying LLM-categorization to all survey responses verifies approach and yields new insights



# Applying LLM-categorization to all survey responses verifies approach and yields new insights: spotting trends



- Despite concerns and abundant rumors, **Covid related issues** (lockdowns, fear of infection, fear of the vaccine) were never a major reason given, and have now disappeared.
- **Transitory/unexpected major events** like health worker strikes can be captured via free text, without need for a priori assignment of these categories.
- **Free-text captures survey design issues.** 2020 survey included responses for children under 12 months, while future surveys did not.

# Integrating AI tools into survey analysis workflows has implications for future data collection and analysis efforts

- We explored the performance of different AI approaches, the associated costs, benchmarking requirements, and researcher/engineering time.
- We conclude that **LLMs are up to this classification task** and would likely be appropriate across many domains (but specifics, especially niche topics, should be tested). **Just a few examples** (FS or FT) improve accuracy. As models continue to improve, this will probably just be taken as a given (are we already there?)
- LLM and related AI tools have the **potential to change how we think about data** collection, large-scale surveys, and analysis
  - LLMs can help us **unlock and structure unstructured** data. Free-text responses in existing data sources, often ignored, can be revisited for new insights.
  - **Survey instrument design** is complex and challenging (ex. are all possible categories accounted for in a question?). Questions and answers are not **tailored** for individual language, culture, and other traits. LLMs could help mitigate these issues by scaling flexibility at point of collection.
  - Reframing free-text and audio transcriptions as **the new raw data** to help avoid interviewer effects and interviewer bias (ex. ‘Negligence’) and to gain richer qualitative insights.
- Technical next steps: Model uncertainty with consistency scoring; new approaches to making categories
-