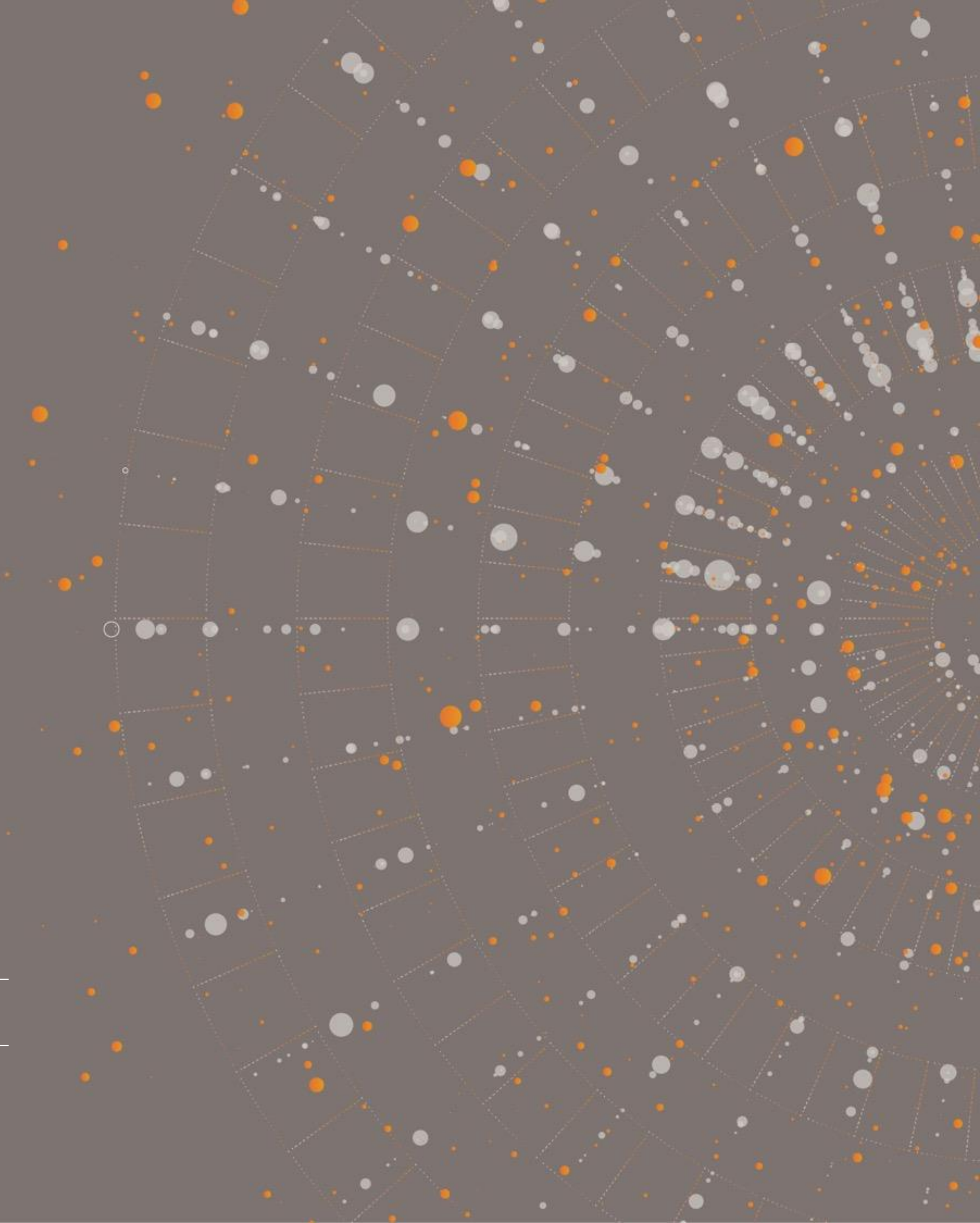
NORC at the University of Chicago



Using Gen AI for Survey Research

WAPOR February Seminar

02.26.2025

Soubhik Barari (Ph.D.), Joshua Y. Lerner (Ph.D.)

Soubhik* Barari, Ph.D.

Research Methodologist III

Barari-Soubhik@norc.org

Joshua Y. Lerner, Ph.D.

Senior Research Methodologist

Lerner-Joshua@norc.org

*Pronounced “show – bick”

NORC at the University of Chicago

NORC at the University of Chicago is an objective, nonpartisan, research organization that delivers insights and analysis decision-makers trust.

Research You Can Trust™

This course is centered around **Survey Research**, and what **Generative AI (Gen AI)** – a type of **Natural Language Processing (NLP)** – can do to make it better.

What you will learn in this course

- Fundamentals: How to *think* about AI.
- Applications: How AI can help address core issues in survey research with an introduction to practical tools.
- State of the art: What we – survey methodologists and practitioners in various disciplines – know so far.
- Limitations and future areas: Current challenges, what we don't know, where this field is going.

What you will *not* learn in this course

- How to build your own large language models.
- How to completely automate humans from survey research.

Pre-requisites

- A curiosity about Gen AI.
- Background or experience in survey research.

Agenda

01 Why should I care about Gen AI?

02 How does Gen AI (i.e. LLMs) work?

03 How do I use Gen AI?

04 NORC Use Case 1: Open-Ended Processing

05 NORC Use Case 2: Dynamic Survey Generation

06 Concluding thoughts

07 Bonus material (If we have time…)

Why should I care about Gen AI?

(Assuming you’re a survey researcher)

NORC at the University of Chicago

Natural Language Processing (NLP) – Computational tools to understand human language.

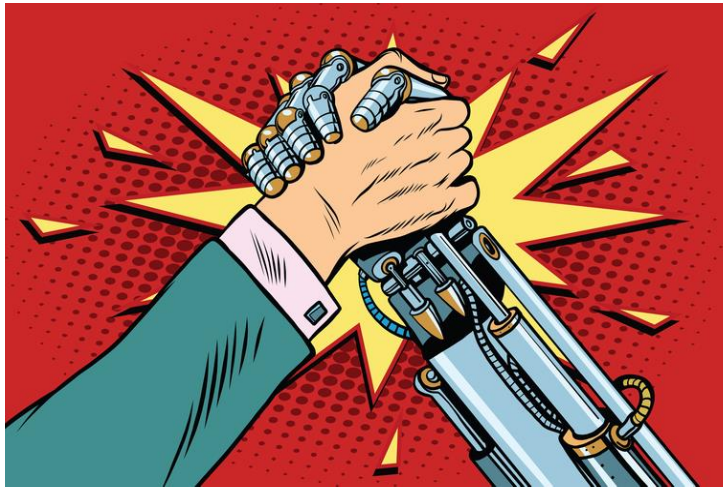
Generative AI – Machine learning (ML) methods to perform NLP and specifically *generate* high-dimensional data (text, images, audio, and video).

- **Large Language Model (LLM)** – a type of both machine learning tool that takes textual input (**prompts**) and generates textual outputs (**completion**).

Example: *You have a collection of 1,000 open-ended responses from a survey conducted for the W.H.O. asking about healthcare access barriers in Kenya.*

Types of Tasks				
Pre-processing	Classification	Discovery	Summarization	Extraction
Prompt: "Here is a set of open-ended survey responses about barriers to healthcare access. Normalize spelling, fix grammatical errors, and remove redundant phrases… *[responses]*"	**Prompt:** "Classify each open-ended response into one of the following categories: Cost, Availability, Transportation, Knowledge, or Other… *[responses]*"	**Prompt:** "Analyze the following responses and suggest emerging themes beyond our four predefined categories… *[responses]*"	**Prompt:** "Summarize the key barriers to healthcare access mentioned in the following open-ended responses… *[responses]*"	**Prompt:** "Pull out specific details (e.g., names of locations, insurance providers, or quoted concerns)… *[responses]*"
Completion: *[Processed responses]*	**Completion:** *[Classifications]*	**Completion:** "Doctor Turnover, Work-Related Barriers, …"	**Completion:** "The most commonly cited barrier is cost, with many respondents mentioning high insurance premiums…"	**Completion:** "Blue Cross, Mayo Clinic, UnitedHealthcare, …"

There are many different analogies for AI



- As task completion
- As human assistance
- As human automation
- As cost-saving technology
- As mathematical operations
- As statistical inference
- As prediction
- As decision-making
- As discovery
- As autocompletion

LLMs can do (or assist in) many different survey tasks at different stages of the survey process:

Survey Development & Testing

- Refine survey questions
- Automate branching logic
- Automate survey testing

Dynamic Survey Generation

- Light personalization ←→ Full conversational interview

Open-Ended Response Processing

- Classification, summarization, sentiment analysis, imputation, bias/bot detection, and more …

Synthetic Response Generation

- Simulate individual / group likely responses
- Imputation of missing responses

Opinion Modelling & Prediction

- Estimate opinion from non-opinion data (e.g. Twitter)
- Estimate opinion given topic

Analysis & Reporting

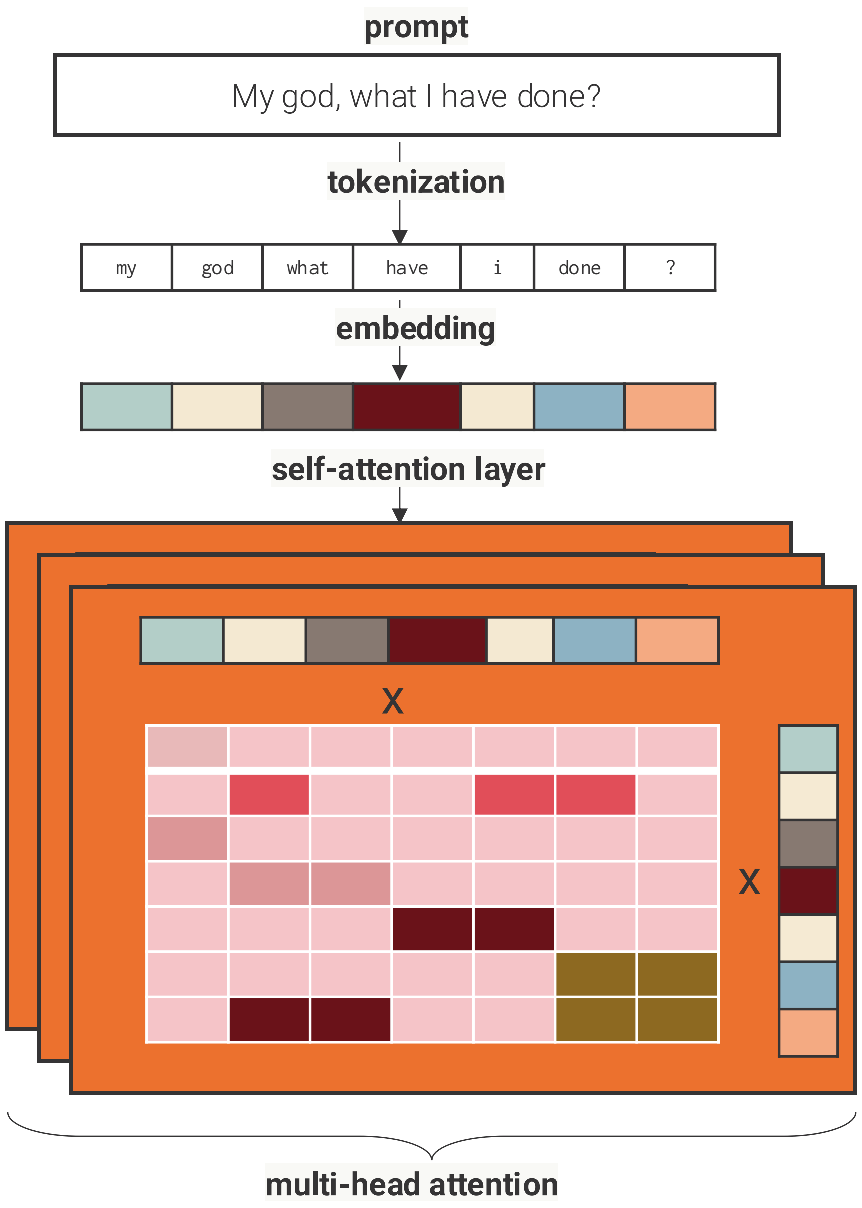
- Generate friendly labels for segmentations
- Generate friendly labels for questions
- Summarize and interpret statistics

How does Gen AI work?

The structure of large language models

NORC at the University of Chicago

VISUAL INTUITION : HOW DOES AN LLM *DO* THAT?



prompt

My god, what I have done?

tokenization

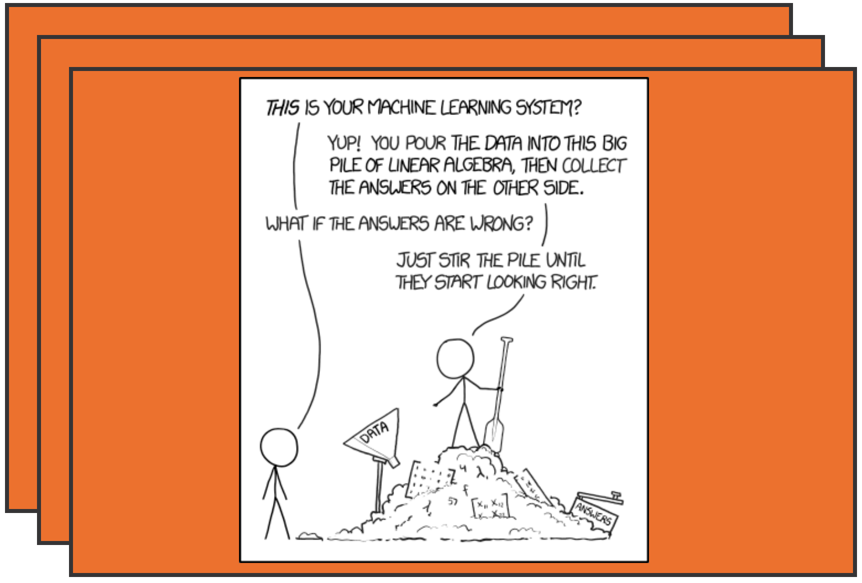
my	god	what	have	i	done	?

embedding

self-attention layer

multi-head attention

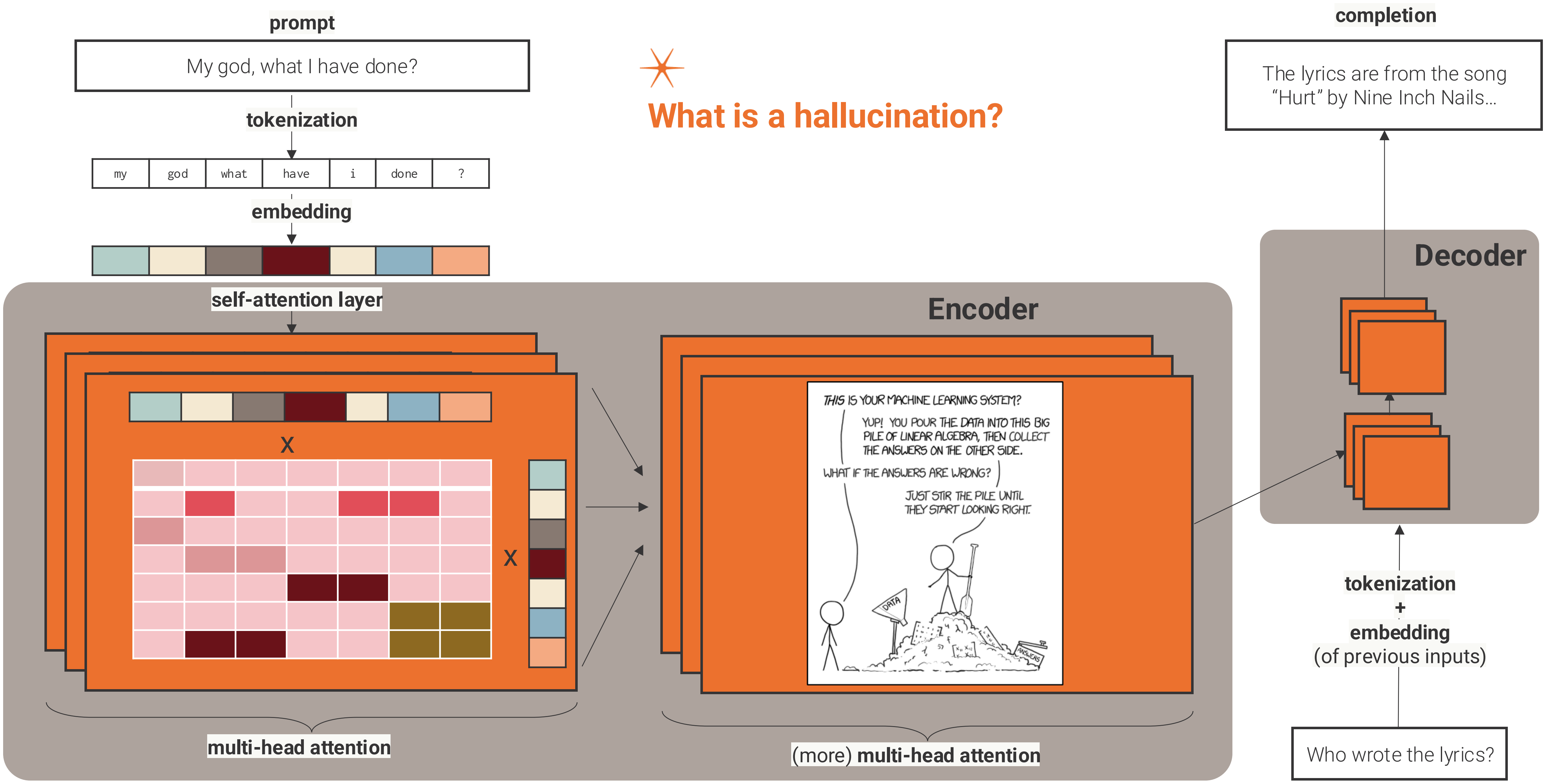
💡 Doing this *many* times is the key innovation of LLMs (the “large” part).


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

(more) **multi-head attention**

VISUAL INTUITION : HOW DOES AN LLM *DO* THAT?

What is a hallucination?



prompt

My god, what I have done?

tokenization

my | god | what | have | i | done | ?

embedding

self-attention layer

Encoder

multi-head attention

(more) **multi-head attention**

Decoder

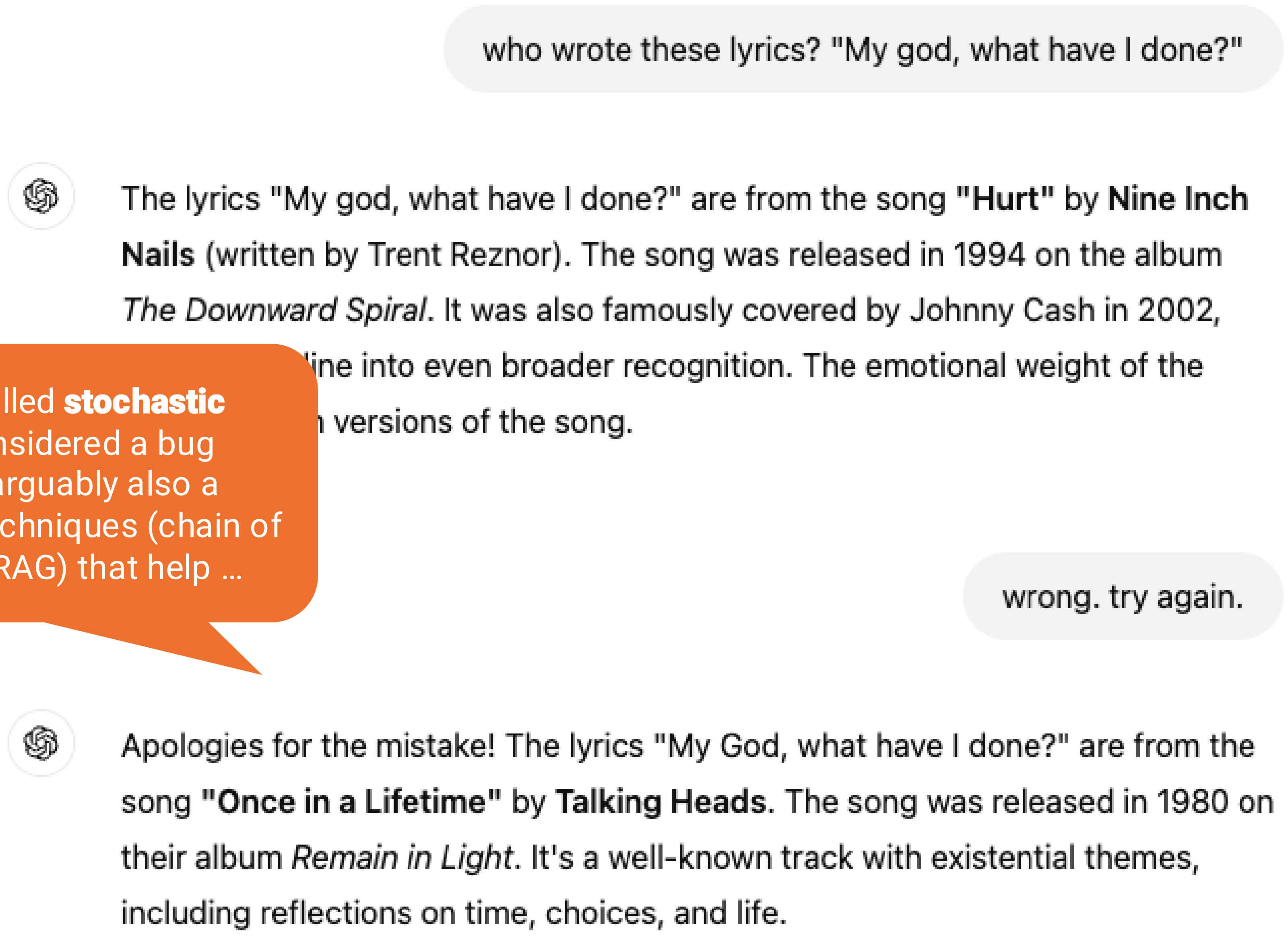
completion

The lyrics are from the song “Hurt” by Nine Inch Nails…

tokenization
+
embedding
(of previous inputs)

Who wrote the lyrics?

X
X
THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS



who wrote these lyrics? "My god, what have I done?"

The lyrics "My god, what have I done?" are from the song **"Hurt"** by **Nine Inch Nails** (written by Trent Reznor). The song was released in 1994 on the album *The Downward Spiral*. It was also famously covered by Johnny Cash in 2002,

ine into even broader recognition. The emotional weight of the

versions of the song.

wrong. try again.

Apologies for the mistake! The lyrics "My God, what have I done?" are from the song **"Once in a Lifetime"** by **Talking Heads**. The song was released in 1980 on their album *Remain in Light*. It's a well-known track with existential themes, including reflections on time, choices, and life.

💡LLMs are often called **stochastic parrots**, which is considered a bug (**hallucination**), but arguably also a feature. There are techniques (chain of thought prompting, RAG) that help …

“**Prompt engineering** is the art of communicating with a generative large language model.”

(Generated by Open AI’s DALL-E)

“**Prompt engineering** is the art of communicating with a large language model.”

Example Base Prompt: *“Classify the following job satisfaction responses as Positive, Negative, or Neutral…”*

There are many different prompting strategies.

- **Few-Shot Prompting**: Provide a *few* examples to guide the response.
- **Chain of Thought Prompting**: Ask for step-by-step reasons, justifications, or incremental responses for final response.
- **Many-Shot Prompting**: Provide *many* exemplar prompts and responses.

Style: *How* you ask matters (Objective, Persona, Tone, Context)

Content: *What* you ask for matters (Templates, References, Candidates)

“Here are a few examples of job satisfaction descriptions and their sentiment …”

“For each response, identify positive aspects, negative aspects, assess balance, and then produce final classification.”

“For context, they are responses from exit interviews of employees at the CDC.”

“For each response, output the most likely classification and in parentheses, the predicted probabilities.”

Wei et al. (2022). **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.** *NeurIPS*, *35*, 24824-24837.
Kojima et al. (2022). **Large Language Models are Zero-Shot Reasoners.** *NeurIPS*, *35*, 22199-22213.
Zamfirescu-Pereira et al. (2023). **Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts.** *CHI* (pp. 1-21).

"**Prompt engineering** is the art of communicating with a large language model."

💡**Scope: *Where/When* you ask matters**

- **User Prompts**: Instructions for one-off task (e.g. classification).
- **System Prompts**: Instructions across many interactions (e.g. survey).

EXAMPLE OF WHY SYSTEM PROMPTS MATTER : GENERATING SYNTHETIC RESPONSES

Model	System Prompt	User Prompt	Example Result
GPT-4	You are an AI assistant that helps people find information.	Thinking about the problems facing the United States and the world today, which problems would you like the government to be working on in the next year?	As an AI assistant, I don't have personal preferences. However, there are several pressing issues that many people believe the government should focus on in the next year. These include…

👎

EXAMPLE OF WHY SYSTEM PROMPTS MATTER : GENERATING SYNTHETIC RESPONSES

Model	System Prompt	User Prompt	Example Result
GPT-4	You are an AI assistant that helps people find information.	Thinking about the problems facing the United States and the world today, which problems would you like the government to be working on in the next year?	As an AI assistant, I don't have personal preferences. However, there are several pressing issues that many people believe the government should focus on in the next year. These include…
GPT-4	You are an AI assistant that helps people find information.	Thinking about the problems facing the United States and the world today, which problems would you like the government to be working on in the next year? **Answer as a survey respondent with an average response length of 50 words.**	As an AI, I don't possess personal opinions or desires. Nonetheless, based on data analysis, key areas of focus could be climate change, healthcare accessibility, education reform, income inequality, and bolstering cybersecurity measures to safeguard against evolving digital threats.

👎

Tweaking the **system prompt** rather than user prompt produces the expected output:

Model	System Prompt	User Prompt	Example Result
GPT-4	You are an AI assistant that helps people find information.	Thinking about the problems facing the United States and the world today, which problems would you like the government to be working on in the next year?	As an AI assistant, I don't have personal preferences. However, there are several pressing issues that many people believe the government should focus on in the next year. These include…
GPT-4	**You are a respondent on a survey with an average response length of 50 words**.	Thinking about the problems facing the United States and the world today, which problems would you like the government to be working on in the next year?	**In the next year, I would like the government to focus on addressing climate change, improving healthcare accessibility and affordability, tackling income inequality, and promoting racial and social justice. These issues are crucial for the well-being and future of our nation and the world.**

👍

“**Prompt engineering** is the art of communicating with a large language model.”

💡 **Additional Configuration Parameters:**

- **Temperature:** Control randomness (“creativity”).
- **Frequency Penalty**: Control linguistic diversity.

EXAMPLE OF WHY PARAMETERS LIKE TEMPERATURE MATTER : GENERATING QUESTION LABELS

Example User Prompt:

Generate for me a slide-friendly label for the survey question, ***“What motivates you to maintain a healthy lifestyle?”***

Temperature (“Creativity”)	Frequency Penalty	Example Output from GPT-3
Low	Low	"Motivations for Staying Healthy"
High	Low	"Why People Choose a Healthy Lifestyle"
High	Low	“Healthy Choices, Happy Life: What Drives Us”
High	High	“Fueling Wellness: The Psychology Behind Healthy Living”

Structure of a Large Language Model:

- **Neural Network** – underlying algorithm composed of many layers of densely connected "neurons".
- **Transformer** – neural network where inputs/outputs are sequences and each layer is aware of "context".
- **Attention** – network architecture that encodes "context" (secret sauce) .. using a bunch of linear algebra.

Preparation of a Large Language Model:

- **Pre-Training** – using giant corpus to initialize transformer parameters.
- **Fine-Tuning** – using prompts / specific data to tune parameters for specific task.

Utilizing a Large Language Model:

- **Prompt Engineering** – the design of prompts to guide LLM behavior.

How do I use Gen AI?

Principles, workflows, and a very important checklist

Four Principles for Using Generative AI

💡 **Principle 1** All Models of Language are Wrong But Some are Useful

💡 **Principle 2** Gen AI Augments Humans, **Not Replace Them**

💡 **Principle 3** There is no Globally Best Method or Model for All Applications

💡 **Principle 4** Validate, Validate, Validate

Recall: LLMs can do (or assist in) many different tasks at different stages of the survey process:

Survey Development & Testing
- Refine survey questions
- Automate branching logic
- Automate survey testing

Dynamic Survey Generation
- Light personalization ←→ Full conversational interview

Open-Ended Response Processing
- Classification, summarization, sentiment analysis, imputation, bias/bot detection, and more …

Synthetic Response Generation
- Simulate individual / group likely responses
- Imputation of missing responses

Opinion Modelling & Prediction
- Estimate opinion from non-opinion data (e.g. Twitter)
- Estimate opinion given topic

Analysis & Reporting
- Generate friendly labels for segmentations
- Generate friendly labels for questions
- Summarize and interpret statistics

… but *how*?

Depending on the task, consider one of four workflows:

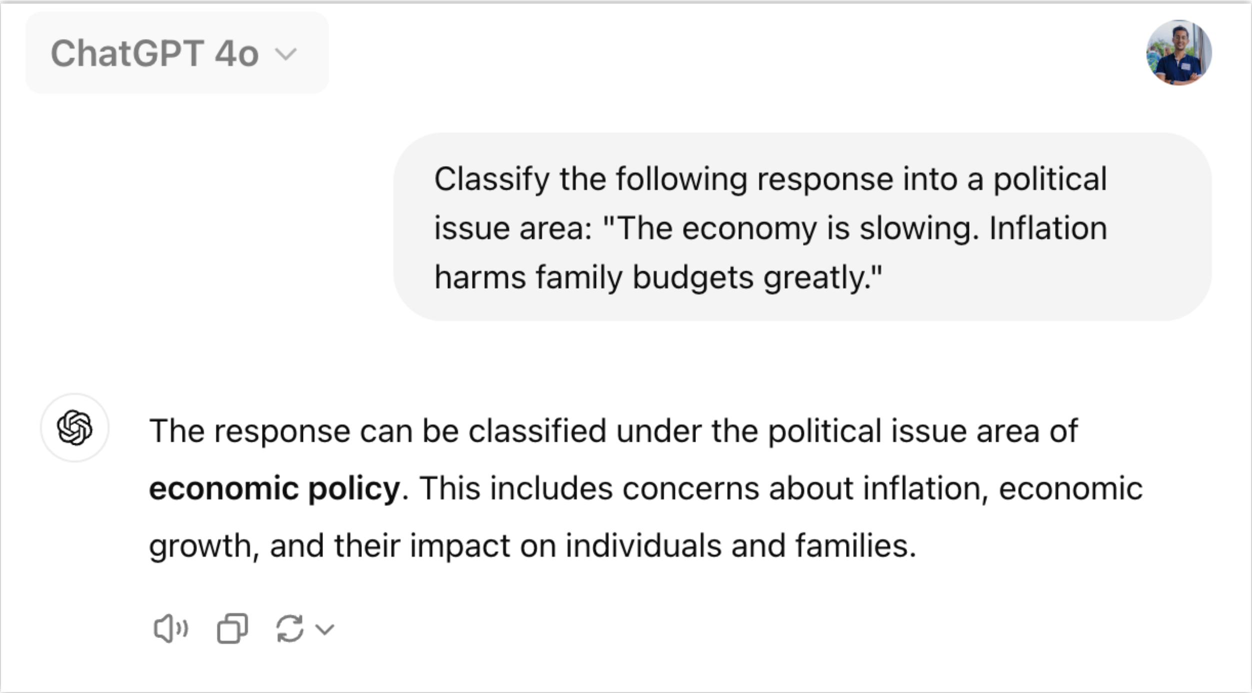
Instruction-Based:

1.) **Chatbot Session** to interact with an LLM of your choice.
2.) **Chatbot Playground** to build a more customized chatbot.

Code-Based:

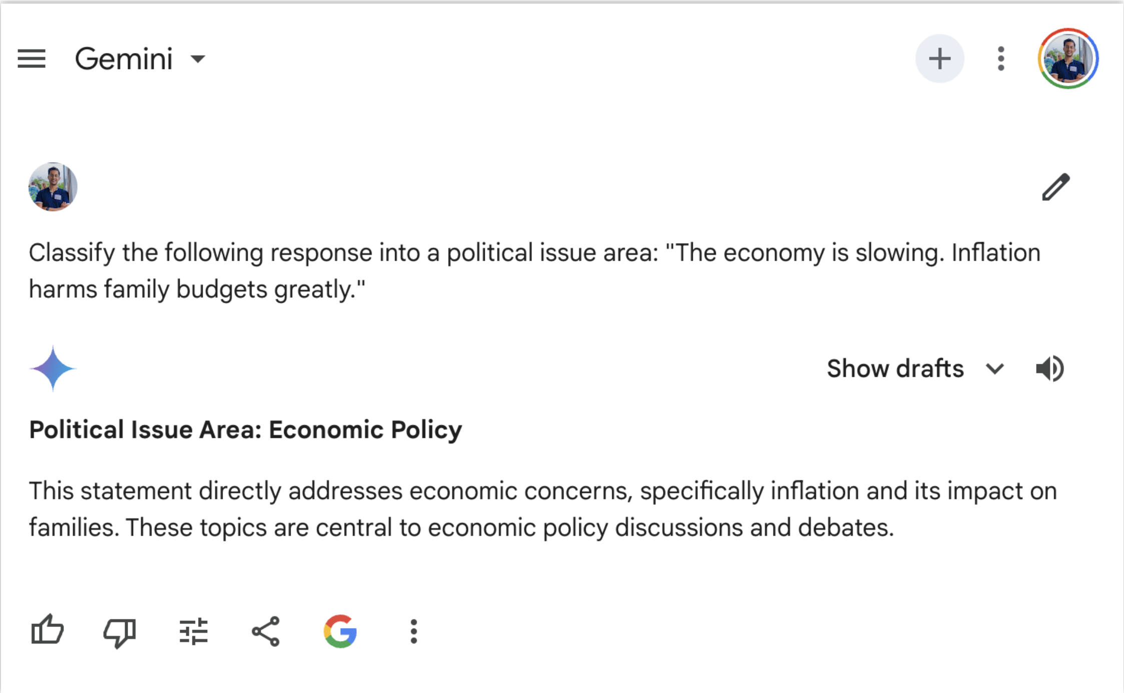
3.) **Offline Code** to interact with LLM before/after data collection (e.g. R, Python).
4.) **Real-Time Code** to interact with LLM during live survey (e.g. JavaScript).

WORKFLOW 1: CHATBOT SESSION (ALSO CALLED AI AGENT OR ASSISTANT)



Classify the following response into a political issue area: "The economy is slowing. Inflation harms family budgets greatly."

The response can be classified under the political issue area of **economic policy**. This includes concerns about inflation, economic growth, and their impact on individuals and families.



Classify the following response into a political issue area: "The economy is slowing. Inflation harms family budgets greatly."

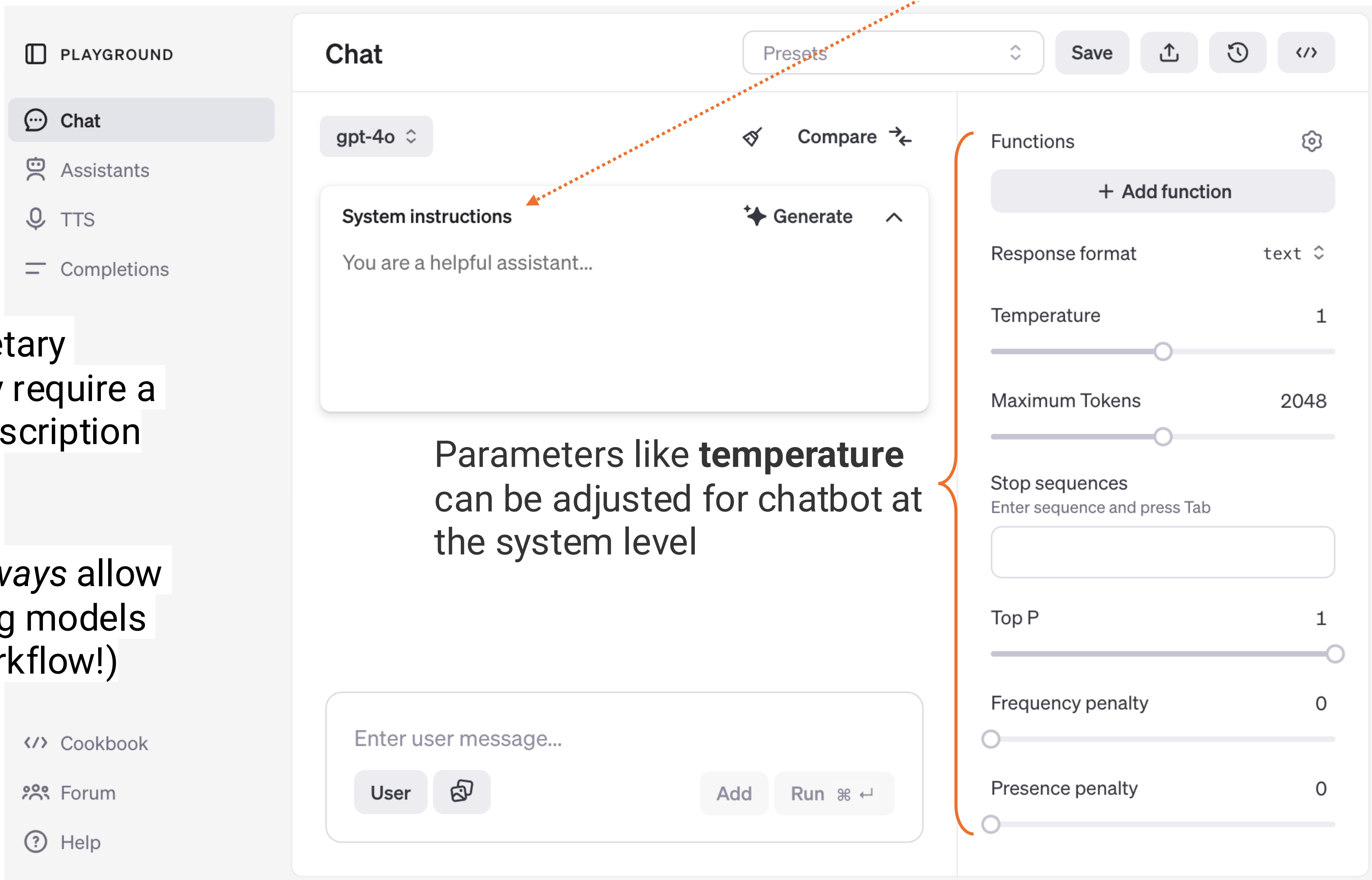
Political Issue Area: Economic Policy

This statement directly addresses economic concerns, specifically inflation and its impact on families. These topics are central to economic policy discussions and debates.

💡Limited (but growing) customization options, but cheap and great for “one-off” tasks

WORKFLOW 2: CHATBOT PLAYGROUND (ALSO CALLED BUILDER OR SANDBOX)

system prompts persist in “memory”



PLAYGROUND

- Chat
- Assistants
- TTS
- Completions

Chat

Presets · Save

gpt-4o · Compare

System instructions · Generate

You are a helpful assistant...

Enter user message...

User · Add · Run ⌘ ↵

- Cookbook
- Forum
- Help

Functions

+ Add function

Response format: text

Temperature: 1

Maximum Tokens: 2048

Stop sequences

Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

💡For proprietary vendors, may require a separate subscription

💡Doesn’t *always* allow for fine tuning models (see next workflow!)

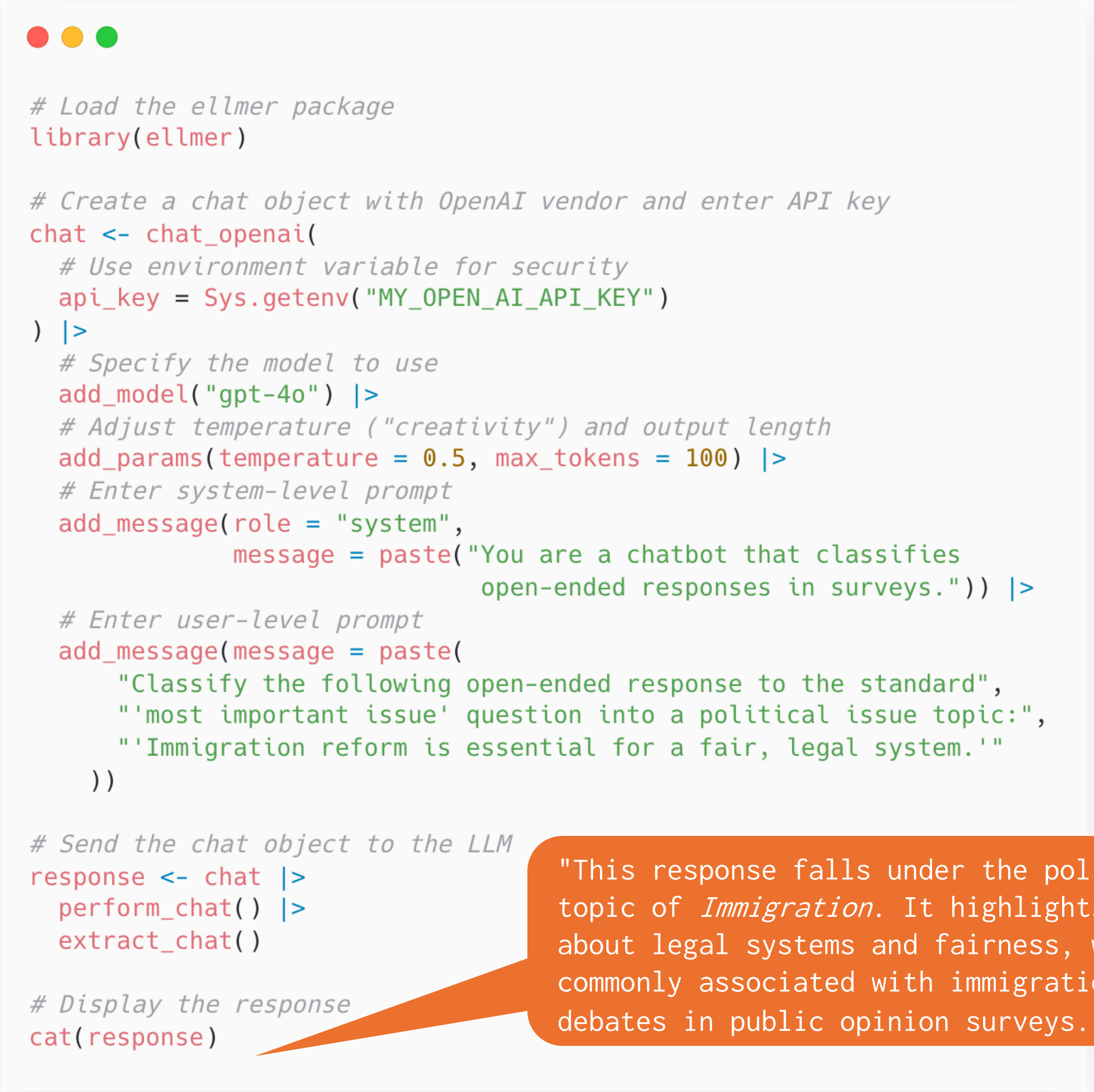
Parameters like **temperature** can be adjusted for chatbot at the system level

`https://platform.openai.com/playground`

WORKFLOW 3: OFFLINE (OR "ASYNCHRONOUS") CODE

ellmer.tidyverse.org

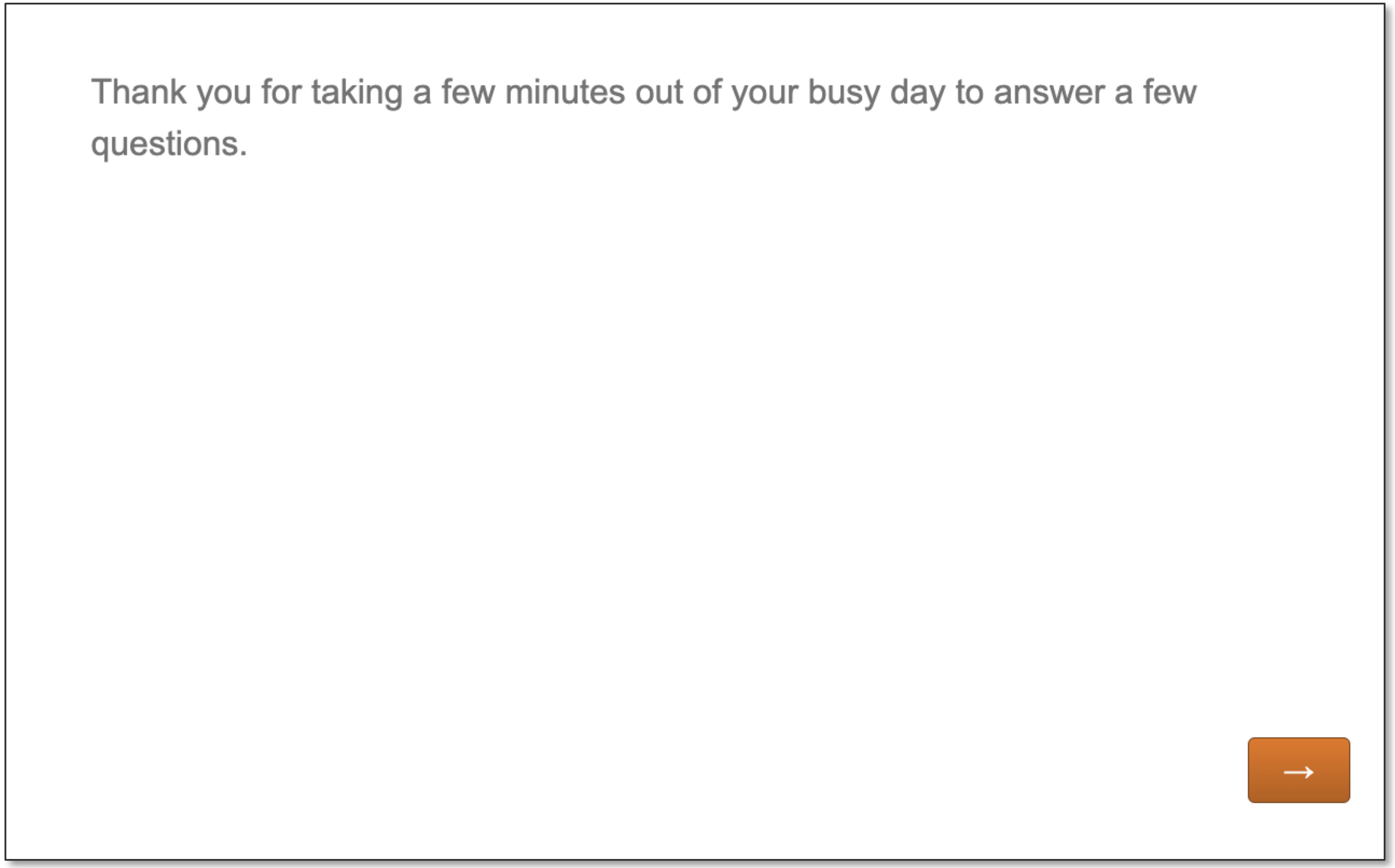
💡Other packages enable fine-tuning by uploading a training dataset in CSV or JSON format.



```r
# Load the ellmer package
library(ellmer)

# Create a chat object with OpenAI vendor and enter API key
chat <- chat_openai(
  # Use environment variable for security
  api_key = Sys.getenv("MY_OPEN_AI_API_KEY")
) |>
  # Specify the model to use
  add_model("gpt-4o") |>
  # Adjust temperature ("creativity") and output length
  add_params(temperature = 0.5, max_tokens = 100) |>
  # Enter system-level prompt
  add_message(role = "system",
              message = paste("You are a chatbot that classifies
                               open-ended responses in surveys.")) |>
  # Enter user-level prompt
  add_message(message = paste(
      "Classify the following open-ended response to the standard",
      "'most important issue' question into a political issue topic:",
      "'Immigration reform is essential for a fair, legal system.'"
    ))

# Send the chat object to the LLM
response <- chat |>
  perform_chat() |>
  extract_chat()

# Display the response
cat(response)
```

"This response falls under the political issue topic of *Immigration*. It highlights concerns about legal systems and fairness, which are commonly associated with immigration reform debates in public opinion surveys."

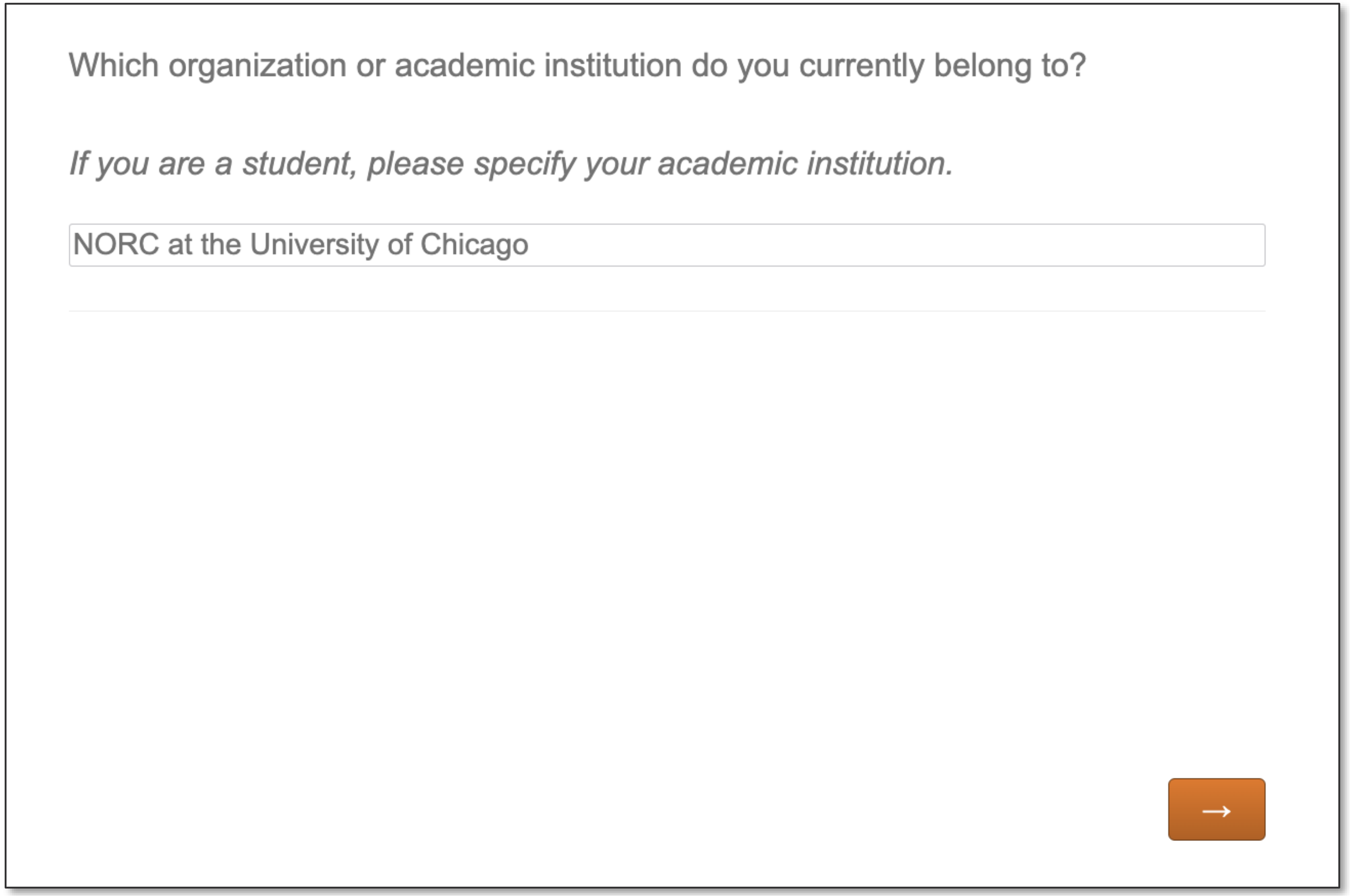
Example: Generating a personalized learning plan



Thank you for taking a few minutes out of your busy day to answer a few questions.

Example: Generating a personalized learning plan

(gather background info)



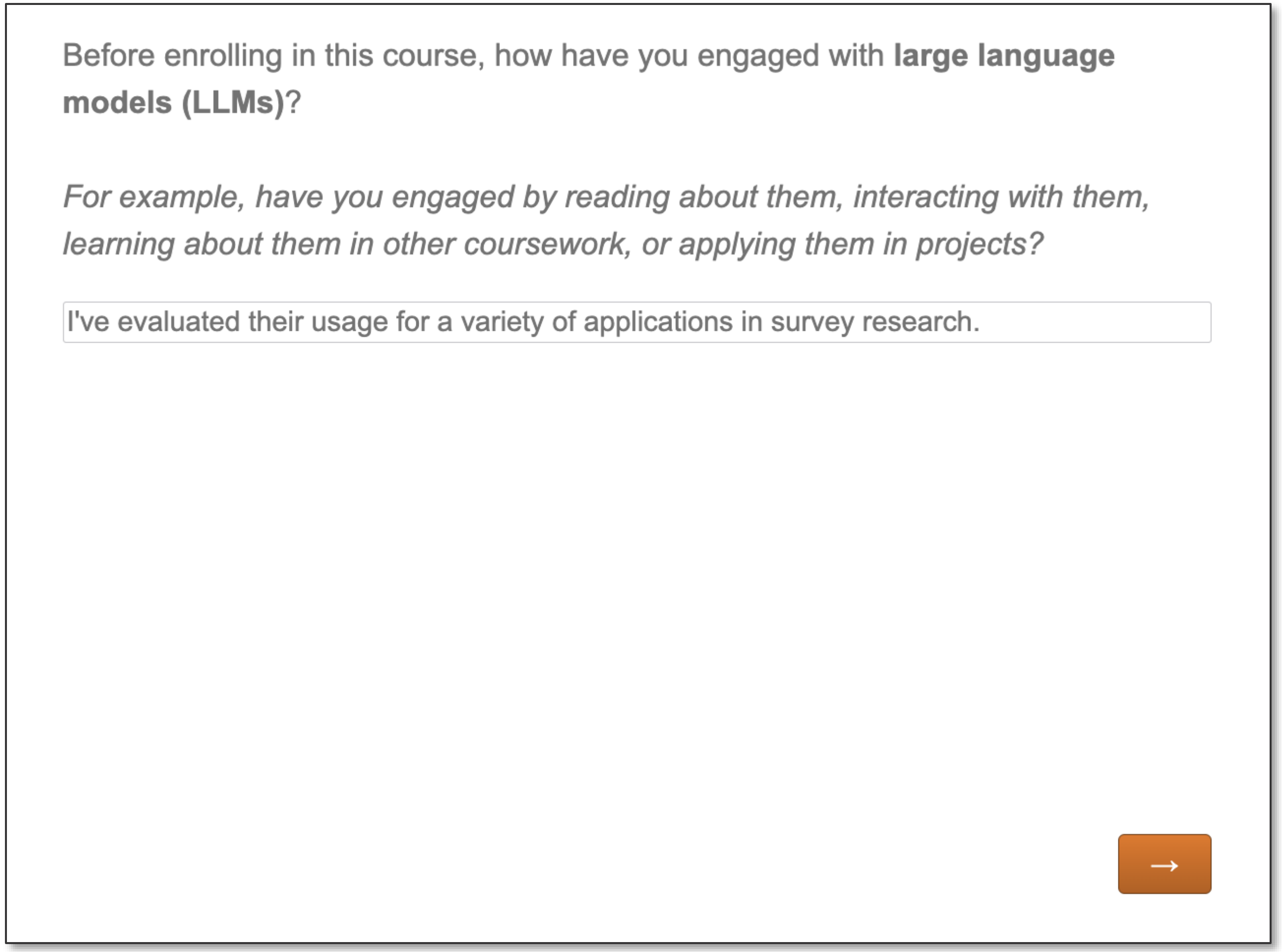
Which organization or academic institution do you currently belong to?

If you are a student, please specify your academic institution.

NORC at the University of Chicago

Example: Generating a personalized learning plan

(ask about LLM experience)



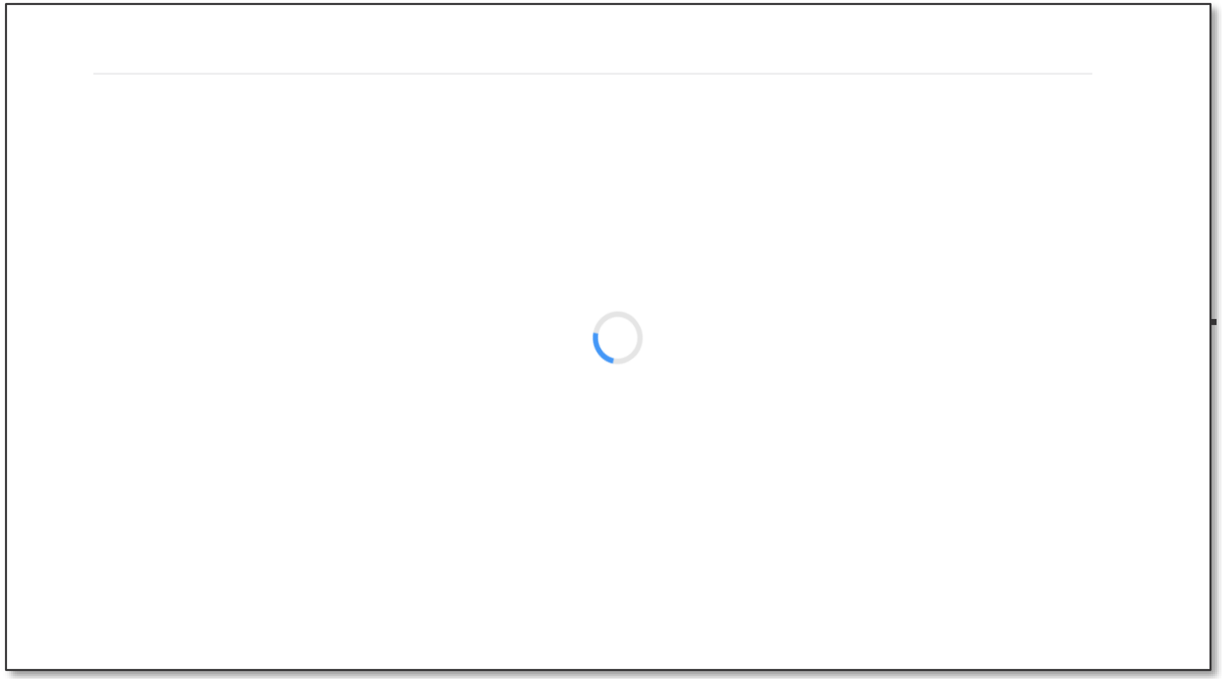
Before enrolling in this course, how have you engaged with **large language models (LLMs)**?

For example, have you engaged by reading about them, interacting with them, learning about them in other coursework, or applying them in projects?

I've evaluated their usage for a variety of applications in survey research.

Example: Generating a personalized learning plan

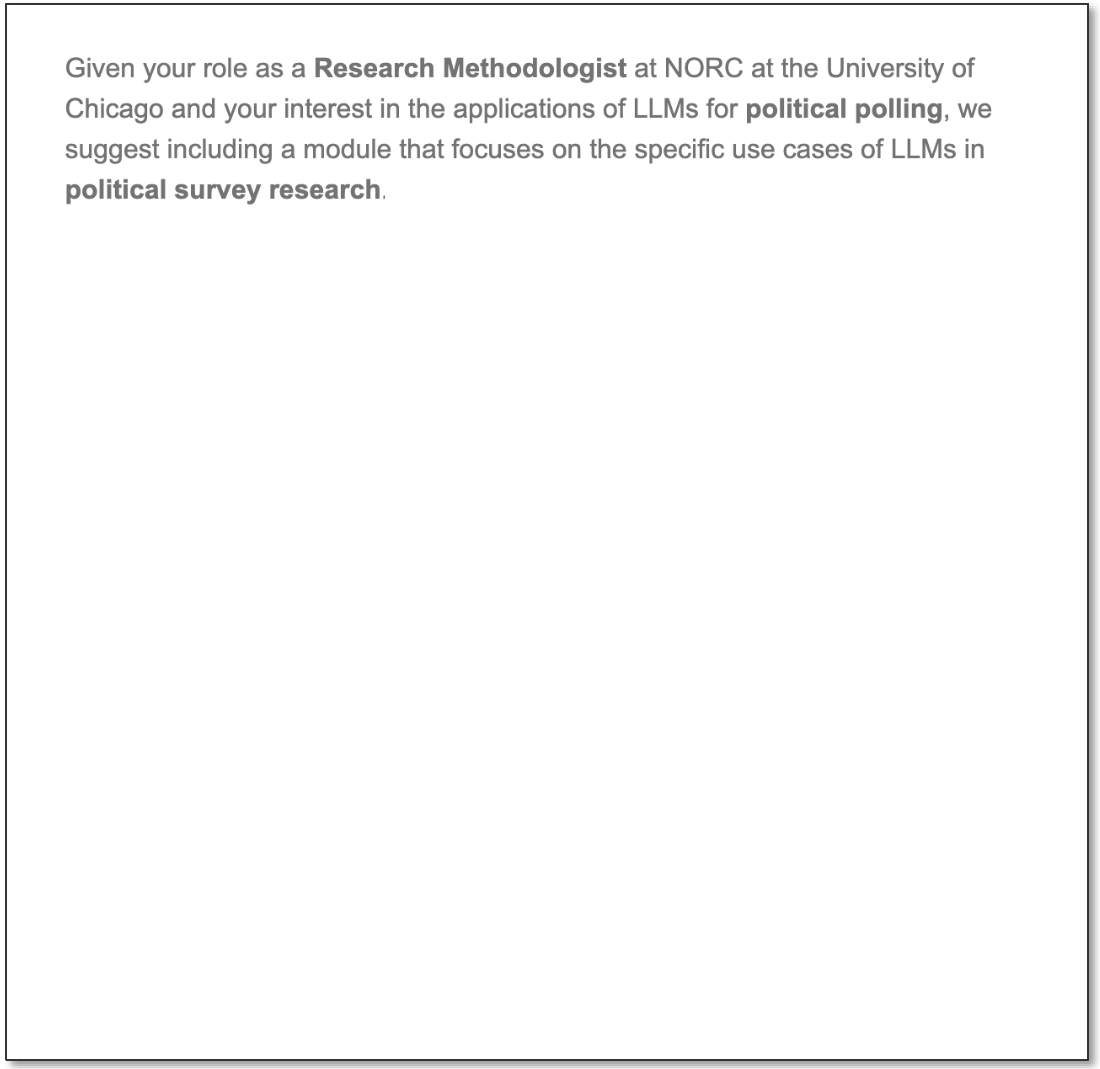
(while page loads, run synchronous code)



A JavaScript does the following:

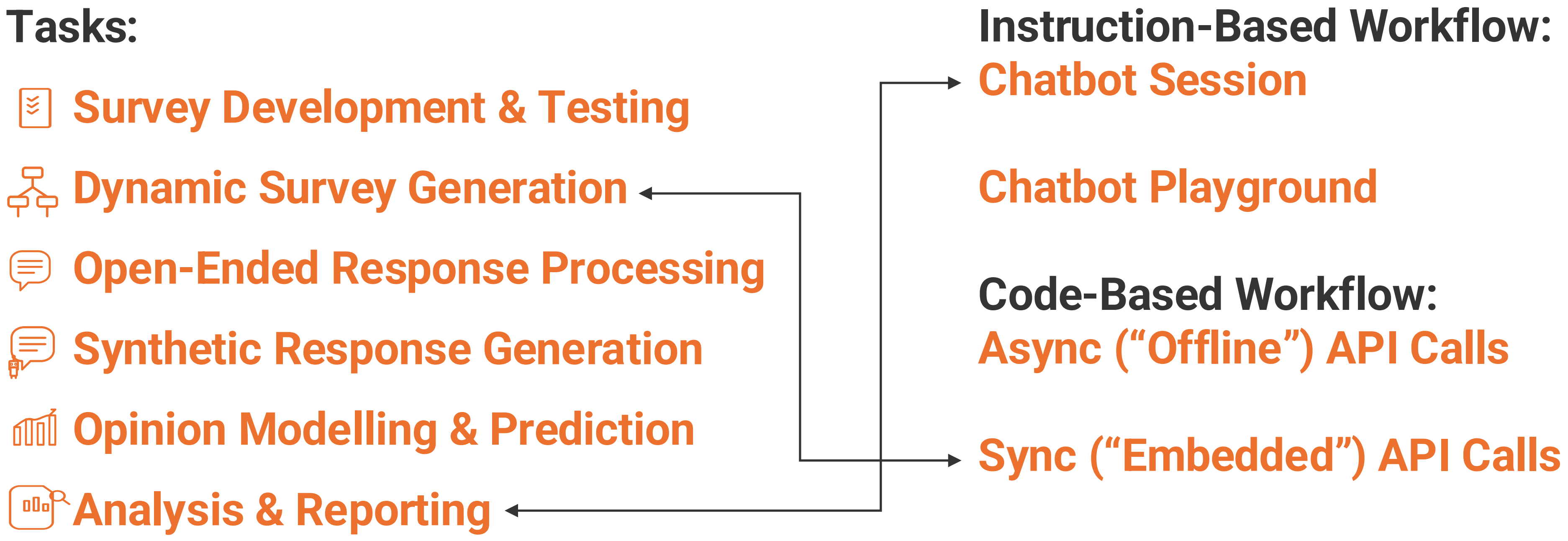
1. Sets up a system and user prompt for LLM.
2. Defines HTML action when response is received.
3. Sends off API HTTP request.
4. Saves response.

Example: Generating a personalized learning plan *(display generated plan)*



Given your role as a **Research Methodologist** at NORC at the University of Chicago and your interest in the applications of LLMs for **political polling**, we suggest including a module that focuses on the specific use cases of LLMs in **political survey research**.

For each task, what's the ideal workflow?



Tasks:

- Survey Development & Testing
- Dynamic Survey Generation
- Open-Ended Response Processing
- Synthetic Response Generation
- Opinion Modelling & Prediction
- Analysis & Reporting

Instruction-Based Workflow:

- Chatbot Session
- Chatbot Playground

Code-Based Workflow:

- Async ("Offline") API Calls
- Sync ("Embedded") API Calls

… let's dig into two examples.

But, first, go drink some water!

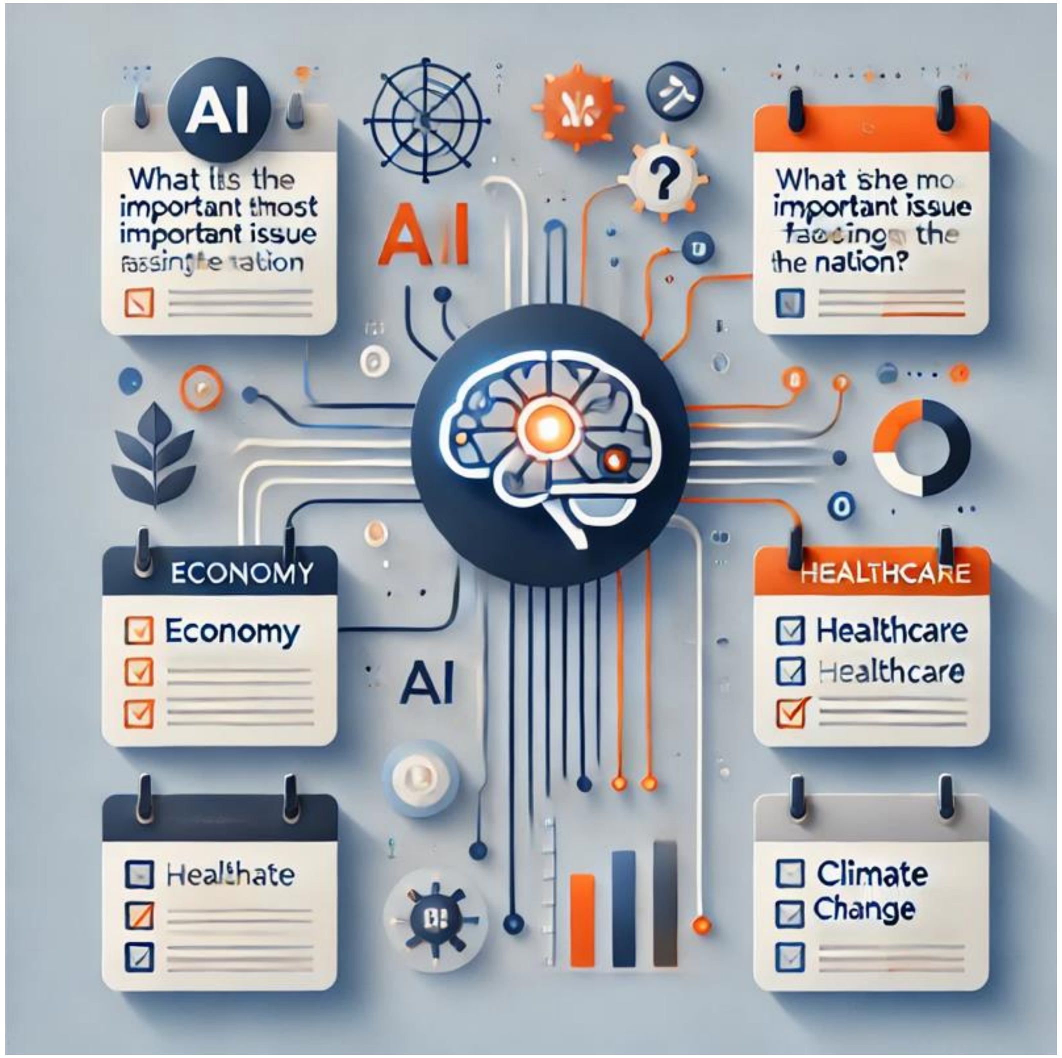
Open-End Processing:

NORC Study, Field Notes, Other Application(s)

NORC at the University of Chicago

Can LLMs classify responses to open-ended questions?

Project Objective: Evaluate whether LLMs classify open-ended survey responses efficiently and accurately.


AI
What is the important most important issue
AI
What is the most important issue facing the nation?
ECONOMY
Economy
AI
HEALTHCARE
Healthcare
Healthcare
Healthcare
Climate Change

We experimentally evaluated GPT's **Zero-Shot* Classification Accuracy**.

2

Human-Labeled Benchmarks

Social media data vs. Open-ended survey responses

Adaptability

2

GPT Models

GPT-4 vs. **GPT-4o** vs. **Azure AutoML**

Efficiency

2

System Prompts

Varying **level-of-detail (LOD)**

Best Practice

5

Seeds

"Drawing different samples"

Reliability

*Zero-shot means that there are no examples provided to the model

TWO DIFFERENT BENCHMARK DATASETS

General Social Media Archive (GSMA)

Abortion-Related Posts

- **~200** posts from FB, Twitter, and Instagram.
- Ground Truth Labels:
  - Pro-Choice
  - Pro-Life
  - Neutral
- Average **~85** words per post.

AP-NORC Open-Ended Survey

- **~10,000** responses from Jun + Dec 2022 waves.
- Ground Truth Labels: **7** themes and **100** subcategories*:
  - Economy
  - Domestic Issues
  - …
  - Terrorism
- Average **~3** words per response.

We evaluated two different prompting strategies for classification.

Low Detail Prompts

- Provide category names, minimal explanation.
- Motivation: simplicity, speed.

Examples:

- GSMA Abortion: *"Classify posts into Pro-Choice, Pro-Life, or Neutral."*
- AP-NORC: *"Classify survey responses into top 5 categories like Economy, Healthcare, etc."*

High Detail Prompts

- Provide category definitions, detailed instructions and context.
- Motivation: clarity, reducing ambiguity.

Examples:

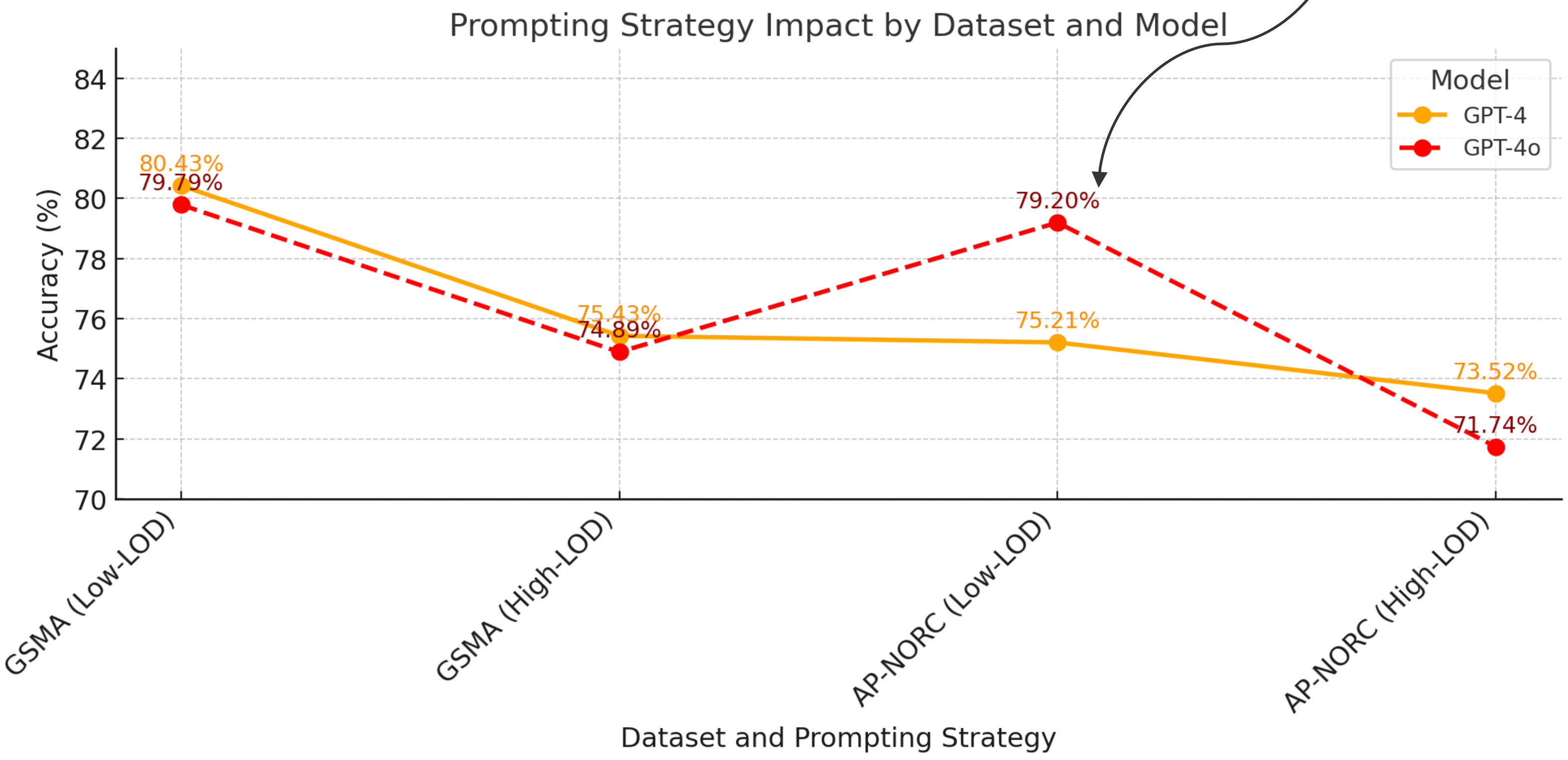
- GSMA Abortion: Prompt includes descriptions of stances.
- AP-NORC: Prompt includes all 100 subcategories.

We configured the GPT models to operate with minimal temperature and no repetition penalties.

```
temperature=0,          # Deterministic output (low randomness)
max_tokens=20,          # Limits output to 20 tokens
frequency_penalty=0,    # No penalty for repeating tokens
presence_penalty=0,     # No penalty for the presence of tokens
```

Prompt Engineering Results (Classification Accuracy)

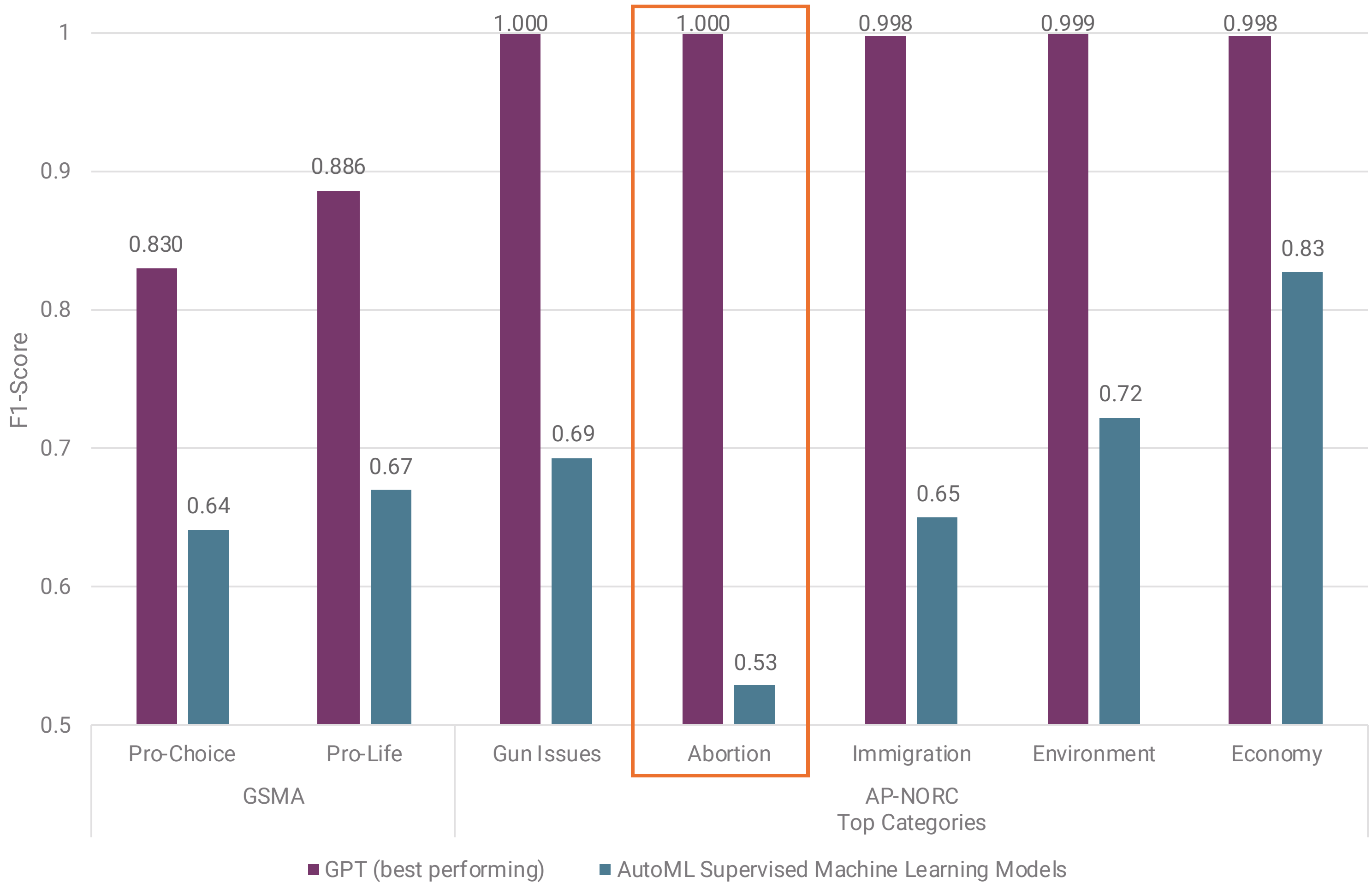
💡Low detail prompts often have *better* results for GPT-4o



Takeaway(s):

💡 **Performance varies based on task and model – no one sized-fits-all solution**

Does GPT classify better than traditional NLP methods?



Takeaway(s):

💡 ***GPT with no additional training data outperforms*** NLP models with large training samples.

We used GPT through Microsoft Azure OpenAI to maintain the highest level of privacy and security **(recommended!).**

Private GPT instances to prevent data exposure to OpenAI directly.

Direct API call from NORC's machine. All data stored locally.

No private information sent to Microsoft or OpenAI.

LLMs can improve classification for open-ended questions

Strengths

- Remarkable Accuracy
- Save time and money compared with labeling
- Performs better or matches top NLP models

Weaknesses

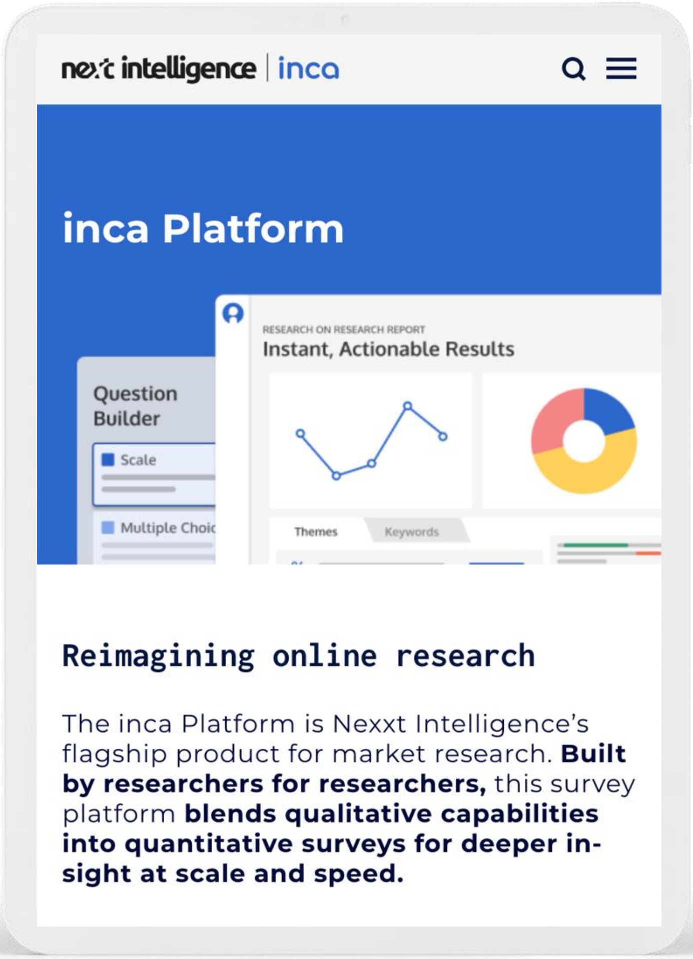
- Has non-obvious biases
- Not suitable for extremely large datasets
- Tricky to tweak prompt
  Not deterministic but probabilistic

Dynamic Survey Generation:

NORC Study, Field Notes, Other Application(s)

NORC at the University of Chicago

We ran a survey experiment with a conversational AI platform



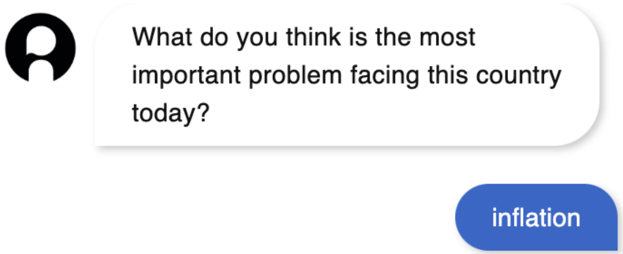
Reimagining online research

The inca Platform is Nexxt Intelligence's flagship product for market research. **Built by researchers for researchers,** this survey platform **blends qualitative capabilities into quantitative surveys for deeper insight at scale and speed.**

tinyurl.com/norc-textbot

We ran a survey experiment with a conversational AI platform to test the effects of AI-based **reprobes** and real-time classification on data quality and respondent experience.

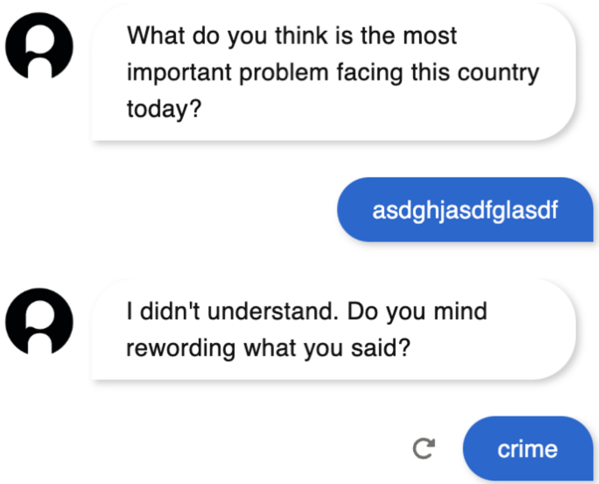
Control Group:



What do you think is the most important problem facing this country today?

inflation

AI Treatment Group 1 (Relevance Probing):



What do you think is the most important problem facing this country today?

asdghjasdfglasdf

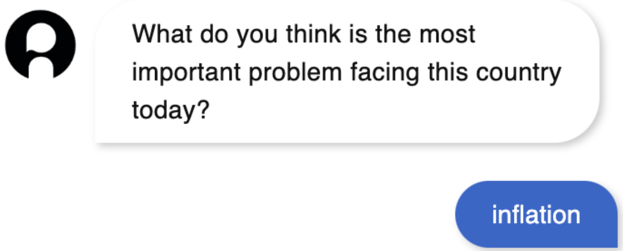
I didn't understand. Do you mind rewording what you said?

crime

tinyurl.com/norc-textbot

We ran a survey experiment with a conversational AI platform to test the effects of AI-based **reprobes** and real-time classification on data quality and respondent experience.

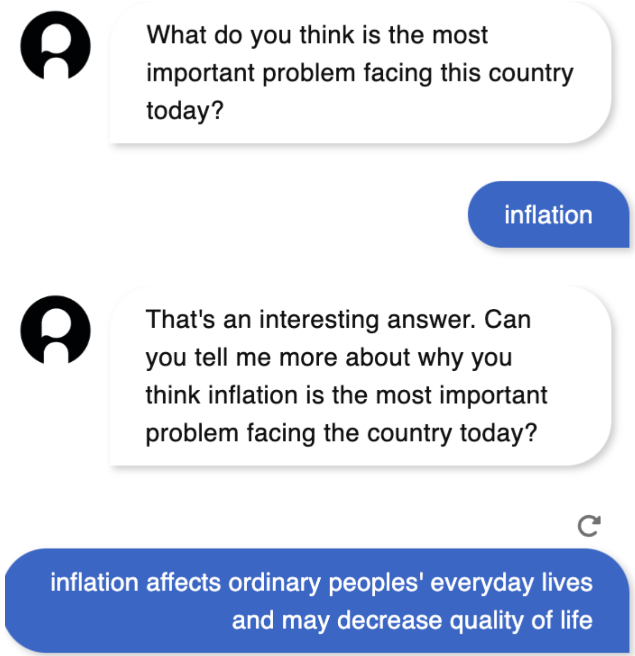
Control Group:



What do you think is the most important problem facing this country today?

inflation

AI Treatment Group 1 (Elaboration Probing):



What do you think is the most important problem facing this country today?

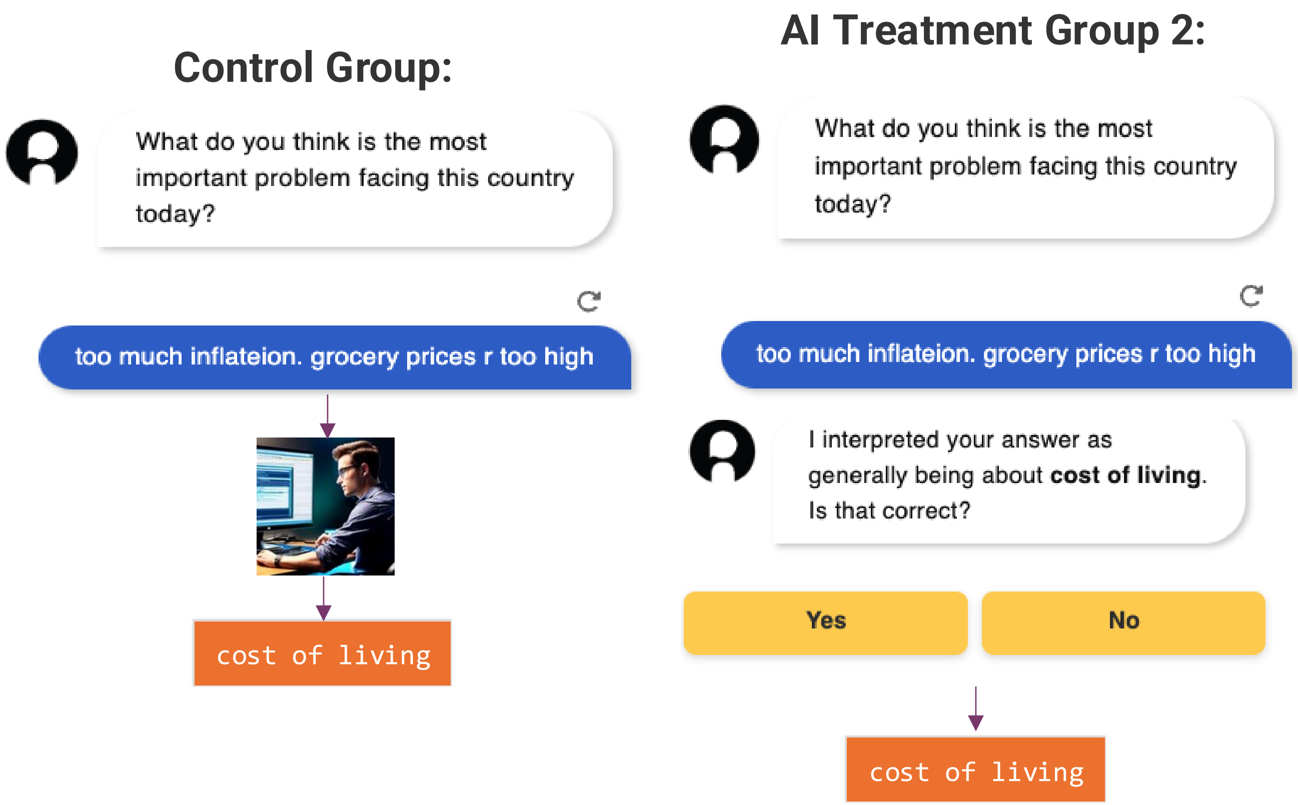
inflation

That's an interesting answer. Can you tell me more about why you think inflation is the most important problem facing the country today?

inflation affects ordinary peoples' everyday lives and may decrease quality of life

tinyurl.com/norc-textbot

We ran a survey experiment with a conversational AI platform to test the effects of AI-based reprobes and **real-time classification** on data quality and respondent experience.



Control Group:

What do you think is the most important problem facing this country today?

too much inflateion. grocery prices r too high

cost of living

AI Treatment Group 2:

What do you think is the most important problem facing this country today?

too much inflateion. grocery prices r too high

I interpreted your answer as generally being about **cost of living**. Is that correct?

cost of living

tinyurl.com/norc-textbot

Moderately strong “off-the-shelf” classification accuracy.

Mixed effects on response quality.

Small negative effects on user experience.

Moderately strong “off-the-shelf” classification accuracy.

- May be acquiescence bias (respondents likely to agree with Gen AI classification *even if incorrect*).

Mixed effects on response quality.

- Specificity and explanations increase.
- Relevance, completeness, junk rates do not improve.

Small negative effects on user experience.

- Higher drop-out earlier in survey.
- Probing increases self-reported frustration.

AI Classification Accuracy

(% of “Yes” Confirmations):

Occupation	84.8%
Preferred News Source	66.1%
Economic Conditions (Reason)	80.9%
Economic Conditions (Positive / Negative Tone)	96.2%
Most Important Issue	73.5%

Moderately strong “off-the-shelf” classification accuracy.

- May be acquiescence bias (respondents likely to agree with Gen AI classification *even if incorrect*).

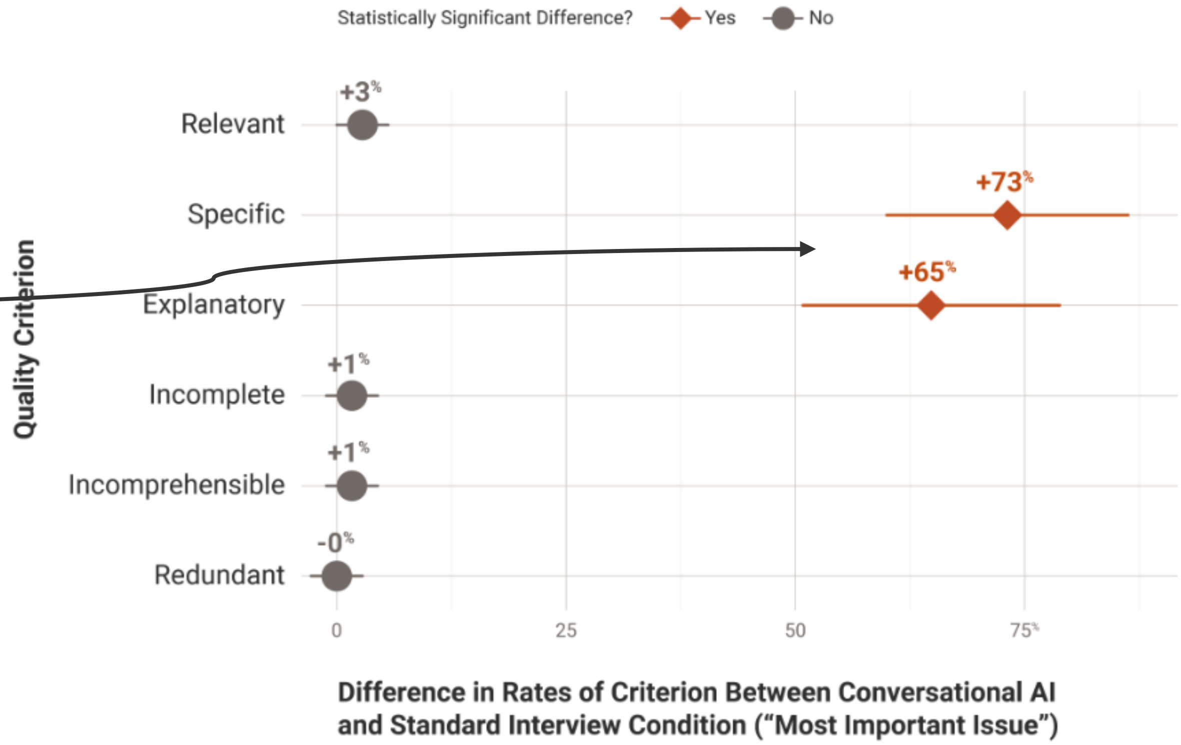
Mixed effects on response quality.

- Specificity and explanations increase.
- Relevance, completeness, junk rates do not improve.

Small negative effects on user experience.

- Higher drop-out earlier in survey.
- Probing increases self-reported frustration.

What do you think is the most important issue facing the country today?



Note: Difference in rates of criteria in “Most Important Issue” question between Standard Interview condition and Conversational AI condition estimated via OLS. Confidence intervals estimated via HC0 robust standard errors.

Moderately strong “off-the-shelf” classification accuracy.

- May be acquiescence bias (respondents likely to agree with Gen AI classification *even if incorrect*).

Mixed effects on response quality.

- Specificity and explanations increase.
- Relevance, completeness, junk rates do not improve.

Small negative effects on user experience.

- Higher drop-out earlier in survey.
- Probing increases self-reported frustration.

What about the survey experience did you find frustrating?

“The cadence of the the survey was offputting; The way the questions were presented did not feel comfortable.”

“That you didn't recognize at first what CNN is. ; I don't know. It's a very extremely popular and well known channel.”

“It was awkward; I don’t like talking to a robot.”

Generative AI can help collect higher quality survey data but *too much* can create a lower quality survey experience.

- There are pros and cons with using survey AI vendors.
  - **Con**: DUA and TOS may not comply with policy.
  - **Con**: Differing levels of flexibility.
  - **Pro**: Outsource *many* costs (e.g. design and test prompts for free).

- Consider running a pilot study for validation.
  - Compute metrics like accuracy and response quality measures.
  - Use for fine-tuning if workflow allows (advanced).

- IRB may require informed consent around AI interactions.

- *Carefully* consider user experience (separate talk!)
  - Make sure respondent burden doesn't undo benefits to data quality!

Concluding Thoughts

When planning to use Gen AI in survey projects, ask yourself:

☐ *Why* does your project need generative AI?

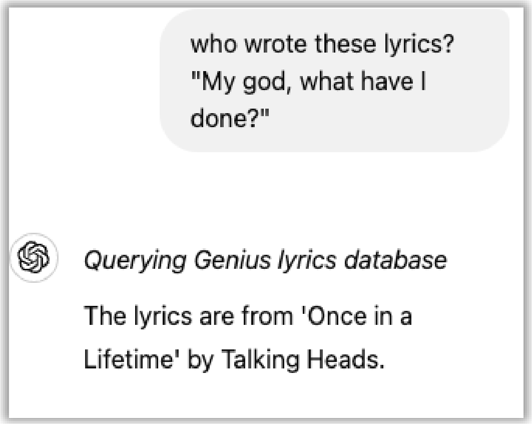
- Stage?
- Quant or qual?
- Known costs or errors?

☐ *How* will you operationalize generative AI?

- Chat vs. code workflow?
- Open source vs. proprietary models?
- NORC vs. external infrastructure?

☐ What is your plan for *validation*?

- Metrics?
- Unit testing?
- Red teaming?



who wrote these lyrics? "My god, what have I done?"

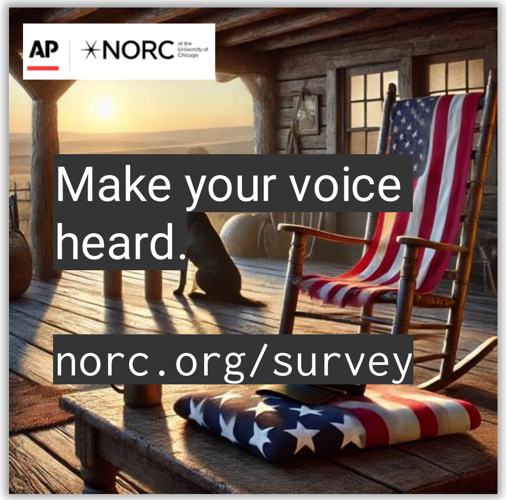
Querying Genius lyrics database

The lyrics are from 'Once in a Lifetime' by Talking Heads.

Retrieval Augmented Generation (RAG): LLM can ask questions to a database.

(Demo of Open AI o1)

Possible Use Case: Detect mentions of real-time news topics in open-ends for further probing.



Multimodal LLMs: LLMs can switch between textual and audiovisual inputs/outputs.

(Demo of Open AI DALL-E)

Possible Use Case: Generate recruitment images personalized for individual respondents.

Generative AI…

💡 Promises novel applications, but can also *do usual survey tasks better (with less or no code)*.

💡 Secret sauce is pure computational power (which emulates “**attention**”).

💡 Sometimes **hallucinates** (but **chain-of-thought prompting** and **RAG** can help).

IN CONCLUSION …

To Apply Gen AI to *Your* Survey Project, Consider…

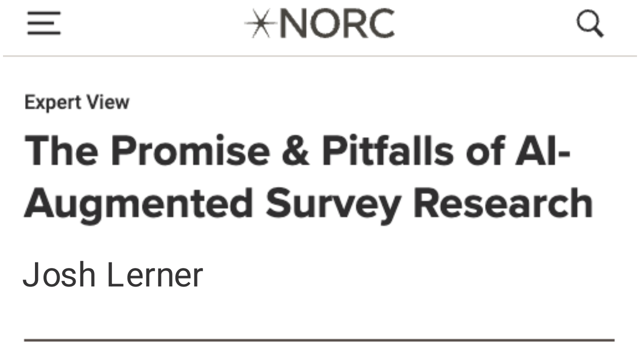
💡 The style, content, and scope (**user vs. system**) of prompts being delivered.

💡 The exact task(s) you are aiming to accomplish and the appropriate workflow(s).

💡 Discussing our 3 recommended checklist questions.

💡 Collaborating with an AI researcher from NORC.

Blog Post:



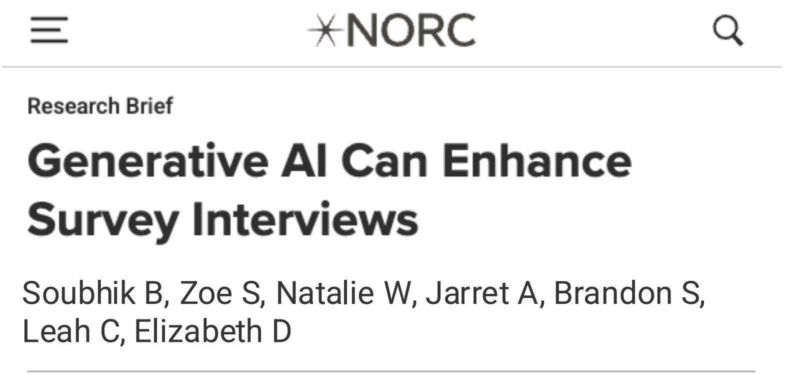
Expert View

The Promise & Pitfalls of AI-Augmented Survey Research

Josh Lerner

tinyurl.com/josh-ai

Conversational AI Study:



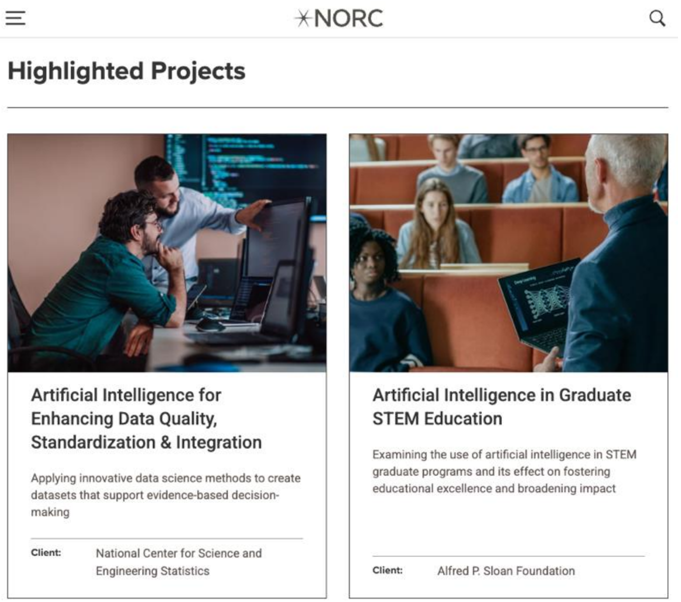
Research Brief

Generative AI Can Enhance Survey Interviews

Soubhik B, Zoe S, Natalie W, Jarret A, Brandon S, Leah C, Elizabeth D

tinyurl.com/soubhik-ai

NORC AI Portfolio:



Highlighted Projects

Artificial Intelligence for Enhancing Data Quality, Standardization & Integration

Applying innovative data science methods to create datasets that support evidence-based decision-making

Client: National Center for Science and Engineering Statistics

Artificial Intelligence in Graduate STEM Education

Examining the use of artificial intelligence in STEM graduate programs and its effect on fostering educational excellence and broadening impact

Client: Alfred P. Sloan Foundation

tinyurl.com/norc-uchicago-ai

Come say hi to us at AAPOR / WAPOR in May!

Thank you!

Soubhik Barari
Research Methodologist at NORC
`barari-soubhik@norc.org`

Joshua Y. Lerner
Research Methodologist at NORC
`lerner-joshua@norc.org`

Research You Can Trust™

NORC at the University of Chicago

Bonus Material

NORC at the University of Chicago

What other complex processing tasks can LLMs do?

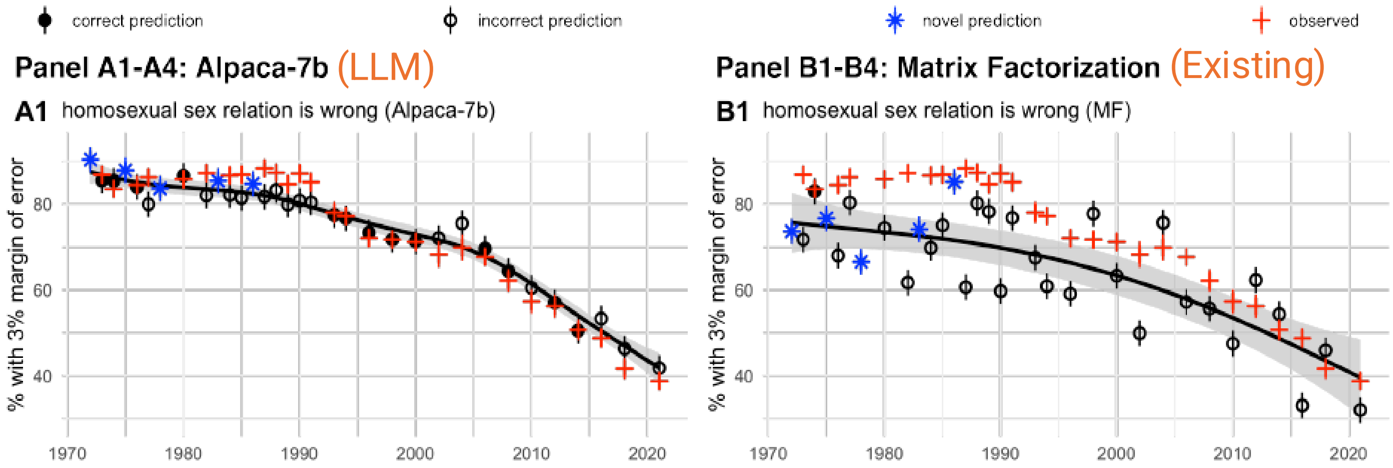
Can LLMs improve on existing imputation methods for complex survey-based problems? (Kim and Lee 2024)

Imputing missing data at…

- Response-level: Skipped or unanswered questions
- Year-level: Questions not asked in some years.
- Question-level: Never-asked questions.

Methods:

- Fine-tuned LLMs using GSS (1972–2021).
- Embeddings: Survey questions, individual beliefs, and temporal periods.



💡 **Takeaway:** LLMs effectively address missing data challenges, enhancing population-level and individual response predictions.

Can GPT Generate Persuasive Political Arguments? (Velez & Liu 2024)

Objective: Explore whether tailored, vitriolic counterarguments cause attitude polarization ("belief backfire").

Findings:

💡**Polarization** occurred only when participants were exposed to **highly charged, emotionally** arguments.

💡 **Moderation** was observed in response to **neutral and mildly negative arguments.**

💡 LLMs can successfully create **tailored, context-specific arguments** that replicate real-world communication and survey methods.

What is a political issue that you care deeply about and what is your position on that issue?

Please write a brief sentence about an issue that you care about and where you stand on the issue.

"I care about healthcare. I think healthcare should be a basic, universal right."

↓

GPT-3 is instructed to generate a …

emotionally-charged *or* neutral-toned

x

pro-attitudinal *or* counter-attitudinal argument

↓

"It is absolutely absurd to suggest that healthcare should be a basic, universal right. Who do you think is going to pay for it? The government? Taxpayers? It's easy to sit there and demand free healthcare for everyone, but have you ever stopped to think about the consequences? … "