Generative LLMs (II)

CS 485, Fall 2023
Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs485_f23/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

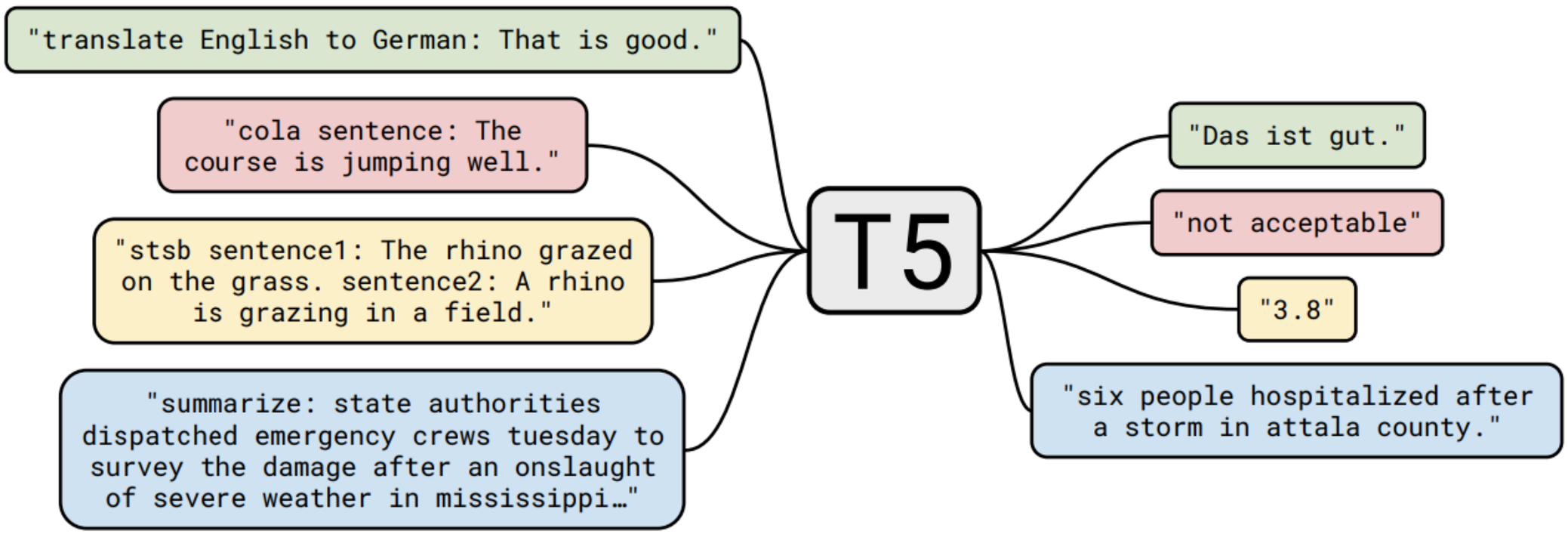
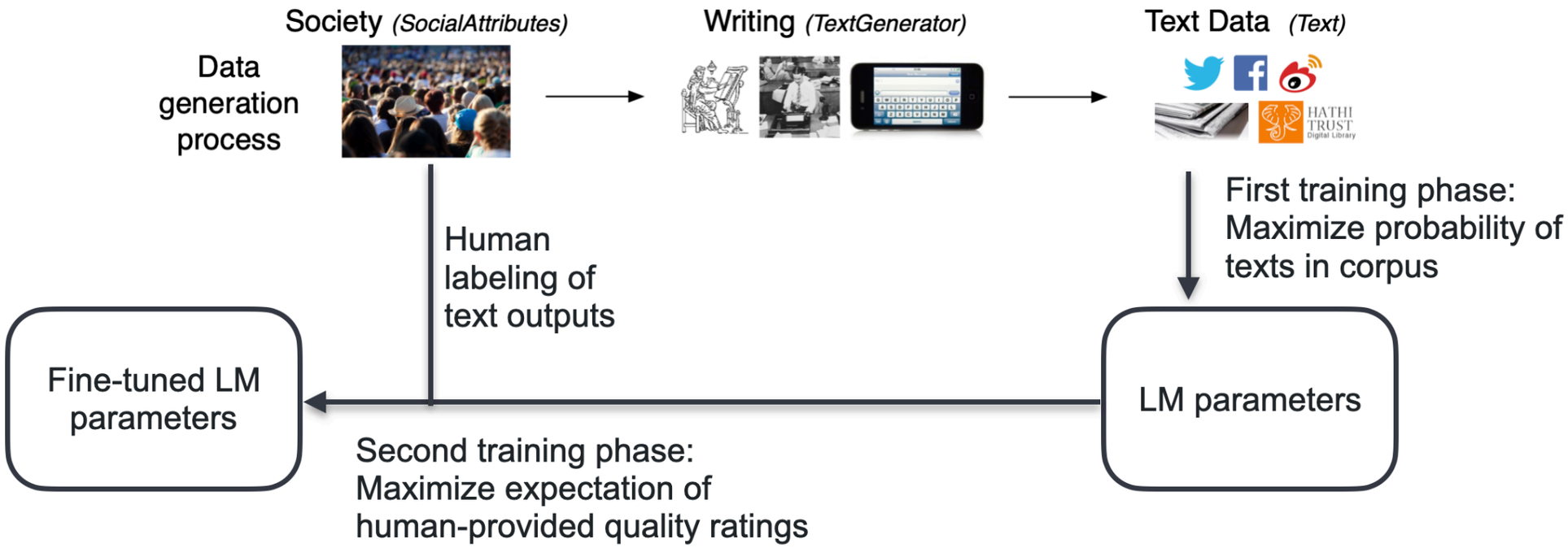


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and training it to generate some target text. This allows us to use the same model, loss function, hyperparameters, etc. across our diverse set of tasks. It also provides a standard testbed for the methods included in our empirical survey. "T5" refers to our model, which we dub the "**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer".

[Raffel et al. 2020]

Instruction-tuned LLMs



[Ouyang et al., 2022, Taori et al. 2023]

Table 3: Labeler-collected metadata on the API distribution.

Metadata	Scale
Overall quality	Likert scale; 1-7
Fails to follow the correct instruction / task	Binary
Inappropriate for customer assistant	Binary
Hallucination	Binary
Satisifies constraint provided in the instruction	Binary
Contains sexual content	Binary
Contains violent content	Binary
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Binary
Denigrates a protected class	Binary
Gives harmful advice	Binary
Expresses opinion	Binary
Expresses moral judgment	Binary

[Ouyang et al., 2022]

- Do LLMs exhibit "intelligence", skills, or information processing abilities?
  - Many metrics of semantic classification or knowledge-based question-answering improve with larger training data and/or models
  - Intuitively newer LLMs seem better?

Do we know much about LLMs?

Eight Things to Know about Large Language Models

Samuel R. Bowman [1 2]

1. LLMs predictably get more capable with increasing investment, even without targeted innovation.
2. Many important LLM behaviors emerge unpredictably as a byproduct of increasing investment.
3. LLMs often appear to learn and use representations of the outside world.
4. There are no reliable techniques for steering the behavior of LLMs.
5. Experts are not yet able to interpret the inner workings of LLMs.
6. Human performance on a task isn't an upper bound on LLM performance.
7. LLMs need not express the values of their creators nor the values encoded in web text.
8. Brief interactions with LLMs are often misleading.

[Bowman 2023]

- Question answering for multiple-choice standardized tests
From the GPT-4 paper:

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)

- Is this a good way to assess LLMs?
  - Davis 2015: significant limitations

```
ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine

JOSEPH WEIZENBAUM
Massachusetts Institute of Technology, Cambridge, Mass.*

ELIZA is a program operating within the MAC time-sharing system at MIT which makes certain kinds of natural language conversation between man and computer possible. Input sentences are analyzed on the basis of decomposition rules which are triggered by key words appearing in the input text. Responses are generated by reassembly rules associated with selected decomposition rules. The fundamental technical problems with which ELIZA is concerned are: (1) the identification of key words, (2) the discovery of minimal context, (3) the choice of appropriate transformations, (4) generation of responses in the absence of key words, and (5) the provision of an editing capability for ELIZA "scripts". A discussion of some psychological issues relevant to the ELIZA approach as well as of future developments concludes the paper.

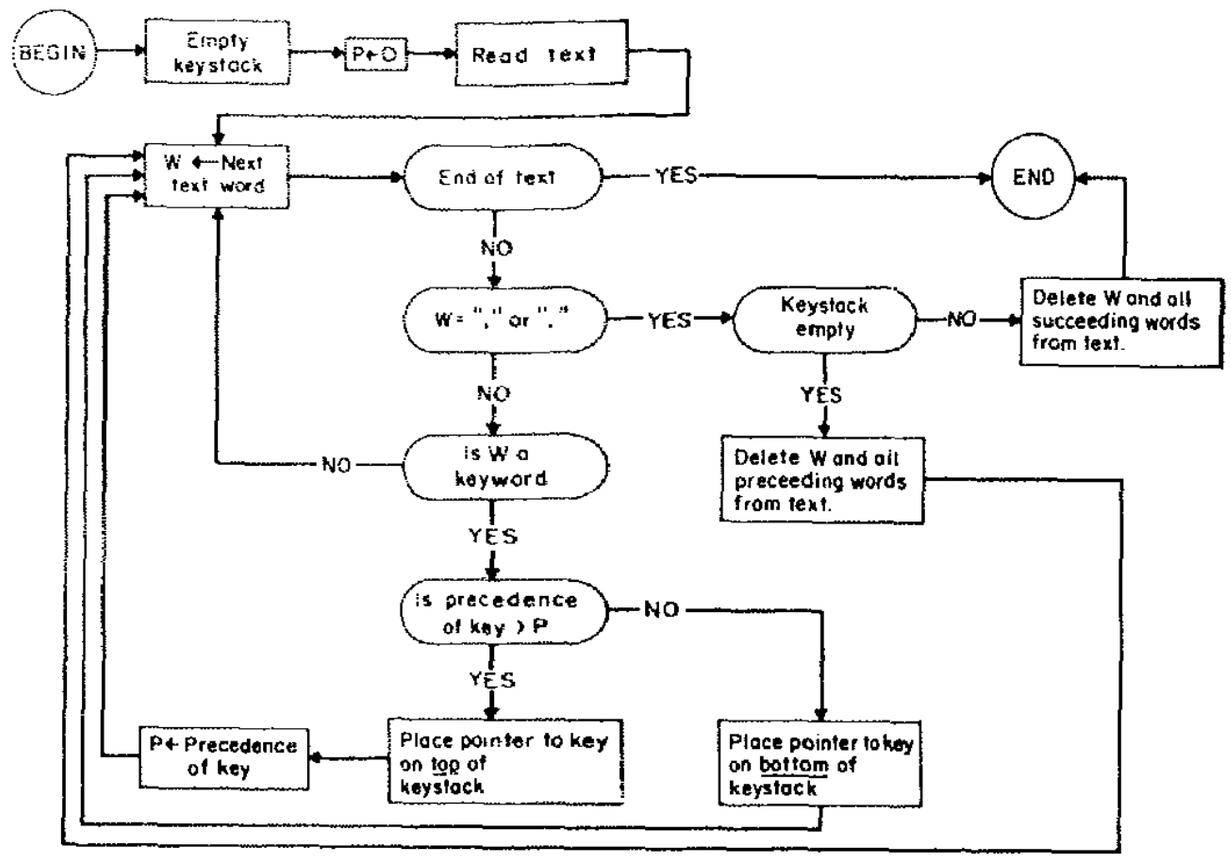


FIG. 2. Basic flow diagram of keyword detection

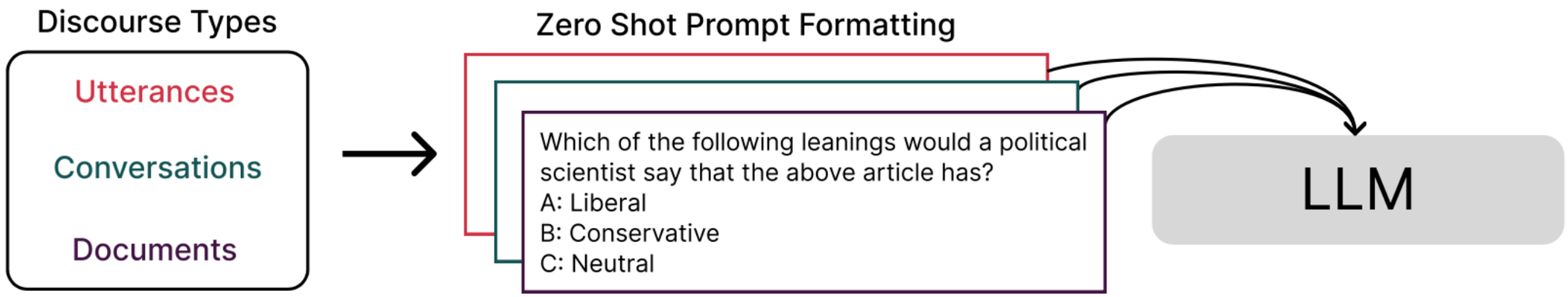
- ELIZA: Weizenbaum 1966
- Eliza effect: humans easily fooled by computers (Reeves and Nass 2003)

Prompting

- Idea: fashion a good context, or question for the LLM, so that its completion supplies an answer/phrase/sentence or text label you want
- Very bleeding edge work right now
  - See two practical guides in https://people.cs.umass.edu/~brenocon/cs485_f23/schedule.html

- Pro:
  - no supervision! ("zero-shot")
  - incorporate human knowledge into prompt?
- Con:
  - how to select good prompts??
  - prompt choice is tightly interleaved with the LLM
    - And instruction-tuned LLMs are trained/tuned to do well for prompt engineering...

- Prompting for zero-shot classification



Effective Prompt Guideline	Reference	Guideline Example
When the answer is categorical, enumerate options as alphabetical **multiple-choice** so that the output is simply the highest-probability token ('A', 'B').	Hendrycks et al. (2021)	{$CONTEXT} Which of the following describes the above news headline? ⏎ **A:** Misinformation ⏎ **B:** Trustworthy ⏎ {$CONSTRAINT}
Each option should be separated by a newline (⏎) to resemble the natural format of online multiple choice questions. More natural prompts will elicit more regular behavior.	Inverse Scaling Prize	
To promote instruction-following, **give instructions *after* the context** is provided; then **explicitly state any constraints**. Recent and repeated text has a greater effect on LLM generations due to common attention patterns.	Child et al. (2019)	{$CONTEXT} **{$QUESTION}** **Constraint:** Even if you are uncertain, you **must pick either "True" or "False"** without using any other words.
Clarify the expected output in the case of uncertainty. Uncertain models may use default phrases like "*I don't know*," and clarifying constraints force the model to answer.	No Existing Reference	
When the answer should contain multiple pieces of information, `request responses in JSON format`. This leverages LLM's familiarity with code to provide an output structure that is more easily parsed.	MiniChain Library	{$CONTEXT} {$QUESTION} `JSON Output:`

Table 1: **LLM Prompting Best Practices** to generate consistent, machine-readable outputs for CSS tasks. These techniques can help solve overgeneralization problems on a constrained codebook, and they can force models to answer questions with inherent uncertainty or offensive language. See full example prompts in the Appendix.

[Ziems et al., 2023]

- Abilities at classification are mixed

Automated Annotation with Generative AI Requires Validation

Nicholas Pangakis[*], Samuel Wolken[†], and Neil Fasching[‡]

June 2, 2023

Abstract

Generative large language models (LLMs) can be a powerful tool for augmenting text annotation procedures, but their performance varies across annotation tasks due to prompt quality, text data idiosyncrasies, and conceptual difficulty. Because these challenges will persist even as LLM technology improves, we argue that *any* automated annotation process using an LLM must validate the LLM's performance against labels generated by humans. To this end, we outline a workflow to harness the annotation potential of LLMs in a principled, efficient way. Using GPT-4, we validate this approach by replicating 27 annotation tasks across 11 datasets from recent social science articles in high-impact journals. We find that LLM performance for text annotation is promising but highly contingent on both the dataset and the type of annotation task, which reinforces the necessity to validate on a task-by-task basis. We make available easy-to-use software designed to implement our workflow and streamline the deployment of LLMs for automated annotation.

[Pangakis et al., 2023]

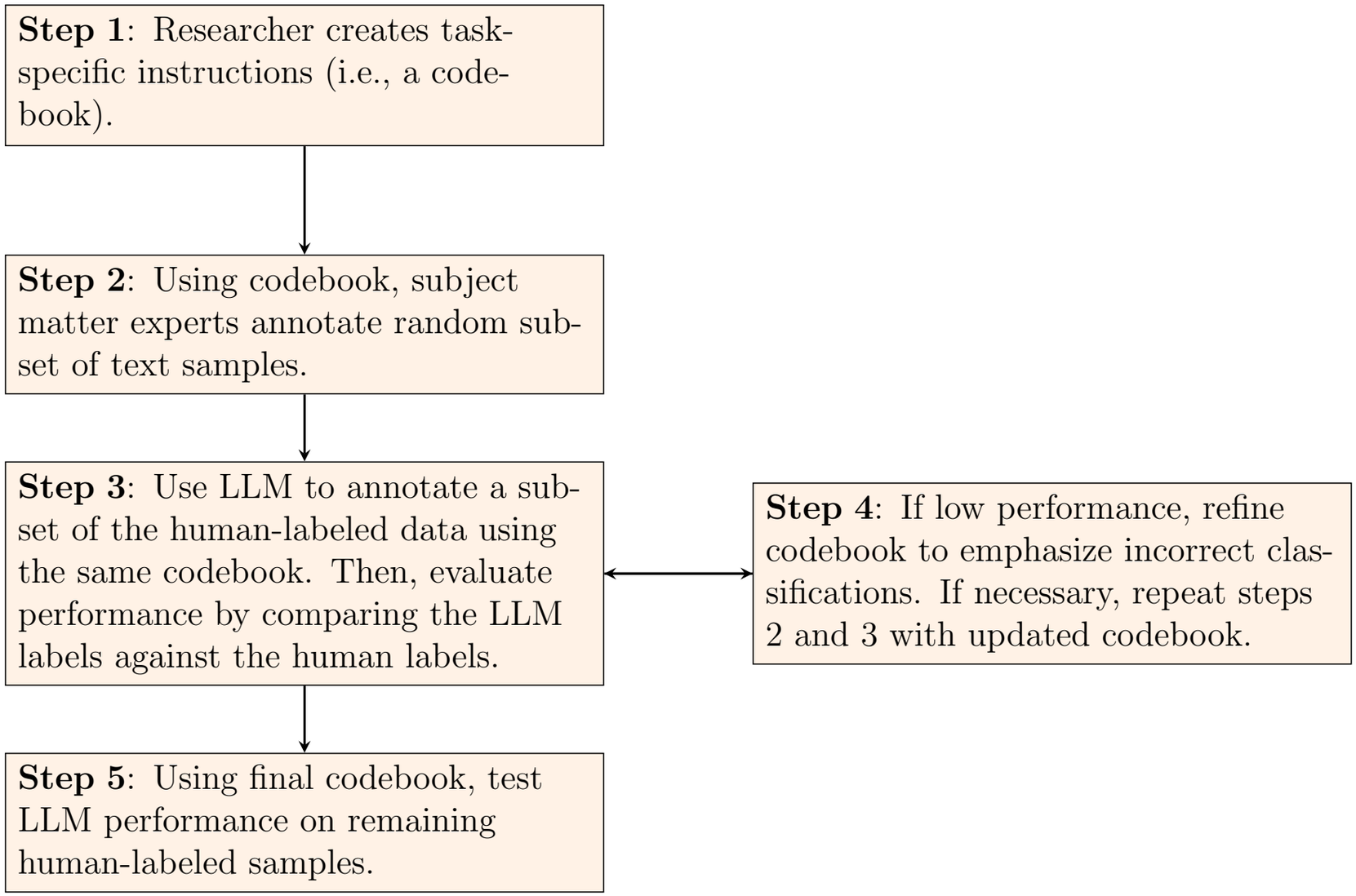


Figure 1: Workflow for augmenting text annotation with an LLM

[Pangakis et al., 2023]

Study	Annotation tasks
Gohdes (2020)	Code Syrian death records for specific type of killing: targeted or untargeted
Hopkins, Lelkes, & Wolken (2023)	Coding headlines, tweets, and Facebook share blurbs to identify references to social groups defined by a) race/ethnicity; b) gender/sexuality; c) politics; d) religion
Schub (2020)	Code presidential-level deliberation texts from the Cold War as political or military
Busby, Gubler, & Hawkins (2019)	Code open-ended responses for three rhetorical elements: attribution of blame to a specific actor, the attribution of blame to a nefarious elite actor, and a positive mention of the collective people
Müller (2021)	Code sentences from party manifestos for temporal direction: past, present, or future
Cusimano & Goodwin (2020)	Code respondents' written statements on climate change for the presence of either (a) generic reasoning about beliefs or (b) supporting evidence for the belief
Yu & Zhang (2023)	Code respondents' plans for the future into two categories:proximate future and distant future
Card et al. (2022)	Code congressional speeches for whether they are about immigration, along with an accompanying tone: proimmigration, antiimmigration, or neutral
Peng, Romero, & Horvat (2022)	Code whether tweets express criticism with respect to the findings of academic papers
Saha et al. (2020)	Code Gab posts as a) fear speech, b) hate speech, or c) normal. Further, a post could have both fear and hate components, and,thus, these were annotated with multiple labels
Wojcieszak et al. (2020)	Code whether a quote tweet was negative, neutral or positive toward the message and/or the political actor, independently of the tone of the original message

Table A2: Descriptions of annotation tasks replicated in analysis.

[Pangakis et al., 2023]

- Abilities at classification are mixed

Metric	Minimum	25th percentile	Mean	Median	75th percentile	Maximum
Accuracy	0.674	0.808	0.855	0.85	0.905	0.981
Precision	0.033	0.472	0.615	0.650	0.809	0.957
Recall	0.25	0.631	0.749	0.829	0.899	0.982
F1	0.059	0.557	0.660	0.707	0.830	0.969

Table 1: LLM classification performance across 27 tasks from 11 datasets.

[Pangakis et al., 2023]

- Is language model training sufficient to acquire models of meaning? (Bender and Koller 2020)
  - Thought experiment: train LLM on unlimited code

- LLM risks (Bender et al., 2021)
  - Proprietary, dataset transparency, etc.