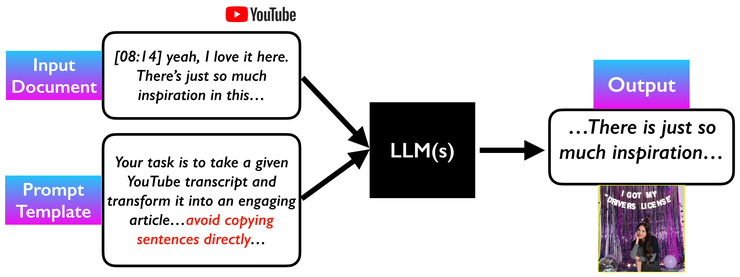
Scaling Up "Vibe Checks" for Large Language Models

Shreya Shankar, Parth Asawa, Madelon Hulsebos, Yiming Lin, J.D. Zamfirescu-Pereira, Bjoern Hartmann,
Aditya G. Parameswaran, Eugene Wu (& collaborators from HKUST, LangChain, and Université de Montréal)

EPIC DATA

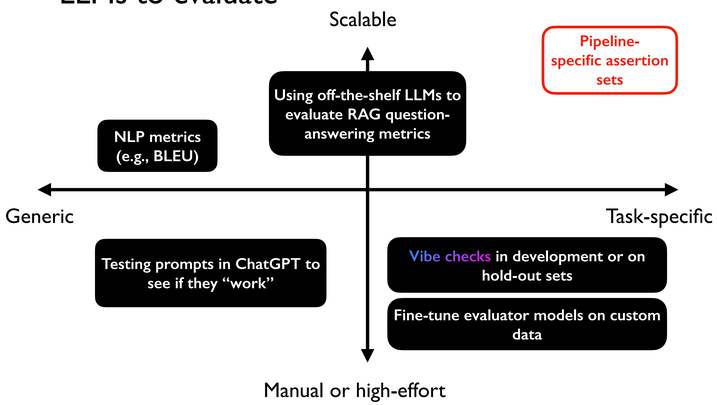
LLM Pipelines

- "Zero-shot" capabilities of LLMs enable intelligent data processing pipelines *without training models*.
- But LLMs make unpredictable mistakes, like hallucinations and ignoring instructions.



Vibe Checks, Rules, and Guardrails

- People rely on rules & guardrails to improve accuracy in traditional ML pipelines
- Hard to do for LLMs
  - What does "accuracy" mean for free-form text?
  - Metrics might be complicated, requiring humans or LLMs to evaluate



Generating Assertion Criteria: They're Hidden in Prompt Version Histories!

Summarize this document {doc_text}. Return your answer in markdown.

↓

Summarize this document {doc_text}. Return your answer in markdown. If the document has sensitive information, don't include it in the summary.

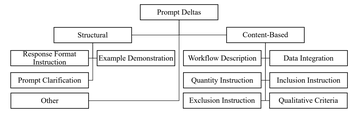
↓

Summarize this document {doc_text}. Return your answer in markdown. ~~If the document has sensitive information, don't include it in the summary.~~ DO NOT under any circumstances include sensitive information (e.g., race, ethnicity, gender).

↓

Summarize this document {doc_text}. Return your answer in markdown. ~~DO NOT under any circumstances include sensitive information (e.g., race, ethnicity, gender).~~ Don't include any sensitive information like race or gender. Have a professional tone.

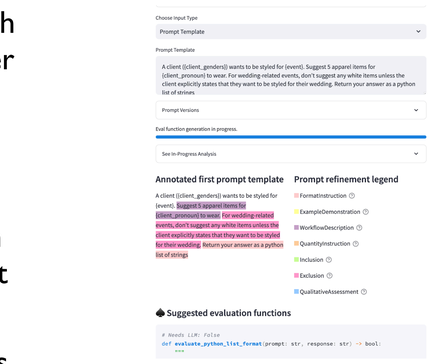
A Taxonomy of Prompt Deltas

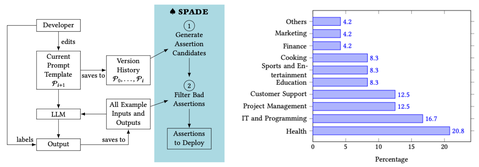


Category	Example Addition or Edit to a Prompt	Assertion Criteria
Response Format Instruction	*"Return your answer in Markdown"*	Parse to markdown correctly
Example Demonstration	*"Here is an example summary: # Medical History…"*	Infer detailed structure from example
Prompt Clarification	*"Return Give me a descriptive answer"*	N/A
Workflow Description	*"First, check for any tables or images. Then, …"*	Check for table summaries
Data Integration	*"The document info is {doc_info}"*	N/A
Quantity Instruction	*"The response should be at least 100 words"*	> 100 words
Inclusion Instruction	*"The title should be the same and end in Summary'"*	Assert same title + "Summary"
Exclusion Instruction	*"Do not include sensitive information"*	No name, race, gender, etc.
Qualitative Criteria	*"Your response should be in a professional tone"*	Professional tone

Large-Scale Deployment Insights

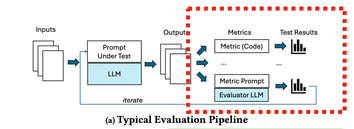
- Deployed a version with LangChain in November 2023
- Findings across 2000+ LLM pipelines
  - Inclusion & exclusion assertions were most common
  - Redundant assertions
  - Incorrect assertions
- See ArXiv preprint for how to solve these issues!





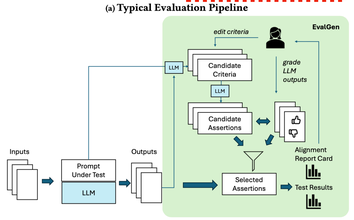
Interfaces for Evaluation Assistants

- To support iteration, we need to minimize wait time
- Can solicit human input *throughout* the assertion generation, filtering, and assessment workflows
  - Humans can edit criteria
  - Humans can grade LLM outputs



(a) Typical Evaluation Pipeline

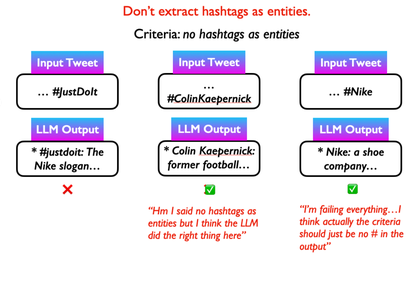
Lots of wait time here



(b) The EvalGen Evaluation Pipeline

Qualitative Study Insights

- Grading LLM outputs spurred changes or refinements to evaluation criteria
  - People *reinterpret* criteria to better fit the LLM's behavior
- Implications: grading must be a *continual* process, as prompts, LLMs, and pipelines change



Work Referenced

Shankar, Shreya, Haotian Li, Parth Asawa, Madelon Hulsebos, Yiming Lin, J. D. Zamfirescu-Pereira, Harrison Chase, Will Fu-Hinthorn, Aditya G. Parameswaran, and Eugene Wu. "SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines." *Under submission.*

Shankar, Shreya, J. D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya G. Parameswaran, and Ian Arawjo. "Who Validates The Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences." *Under submission.*