

SAFETY

Q3'24 Council Meeting

August 22, 2024

Antoine Tixier, Matt Hallowell

Agenda

- ChatSafetyAI Updates:
 - New LLM Engine (GPT-4o)
 - Regulations Mode Improvements
 - Chat History w/ Control Over Data
 - Off-topic Classifier Improvements
- Collect Feedback and Suggestions

This call will be recorded.

ChatSafetyAI Version 08-22-24

- Deployed this morning
- Based on GPT-4o (released 05/13/24), the latest, bleeding edge version of the GPT LLM!
- Many other improvements

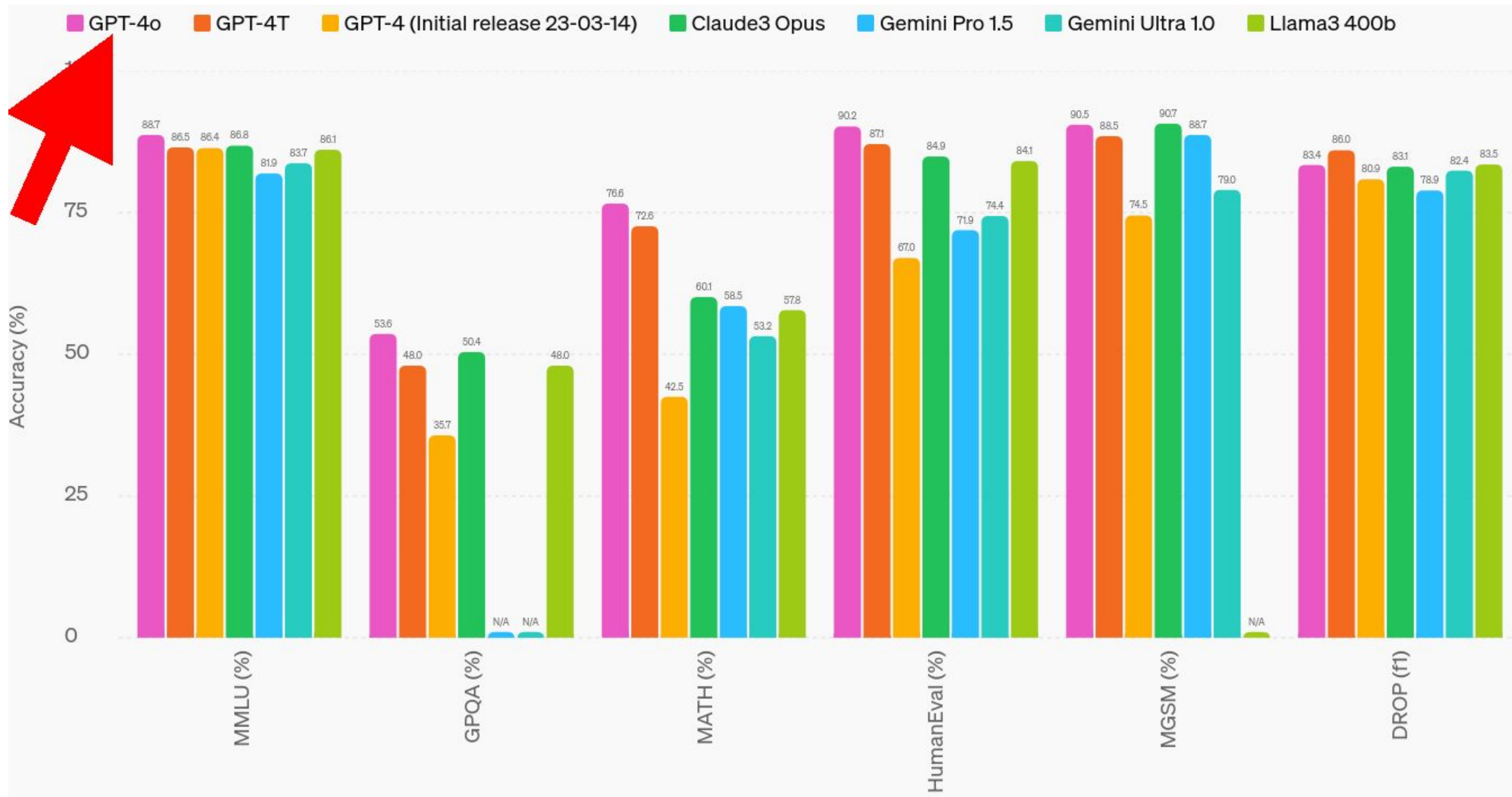
GPT-4o

- “o” stands for “omni”, as the model supports various modalities end-to-end: text, images, audio, and video.
- GPT-4o is faster and has better text and vision understanding than GPT-4 Turbo (released 11/06/2023).
- ChatSafetyAI gets faster and better with any new version of GPT!

GPT-4o vs Other LLMs

Text understanding performance


(Claude: Anthropic AI; Gemini: Google; Llama: Meta)



GPT-4o vs Other LLMs

Vision understanding performance

(Gemini: Google; Claude: Anthropic AI)



Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.5	63.2	

Regulations

- Recall that the regulation mode relies on a separate API of ours, that ChatSafetyAI calls to get the pieces of regulation relevant to the user query
- Search engine development steps:
 - Scrape the text from the regulation documents (HTML pages and PDFs)
 - Split that text into meaningful chunks
 - Get the representation of each chunk (embedding)
 - Build an index

Regulations

- I put a LOT of work into scraping and chunking to fix issues related to tables, equations, and very long standalone parts.
- Manually inspected and fixed the text scraped from 21 documents, totaling 3732 pages.
- Took me almost two months!
- Also: added OSHA 1910 (general industry).
ChatSafetyAI now knows about 1926 & 1910.

Regulations

New indexing and retrieval strategy:

- Documents are now split into both small and large chunks
- Small chunks: relevant, but may miss context
- Large chunks: context-rich, but potentially include unrelated parts
- Combining the two approaches balances precision and recall
- Means more results to process for the LLM (greater risk of errors), but GPT-4o is intelligent and robust enough.

Regulations

More flexible:

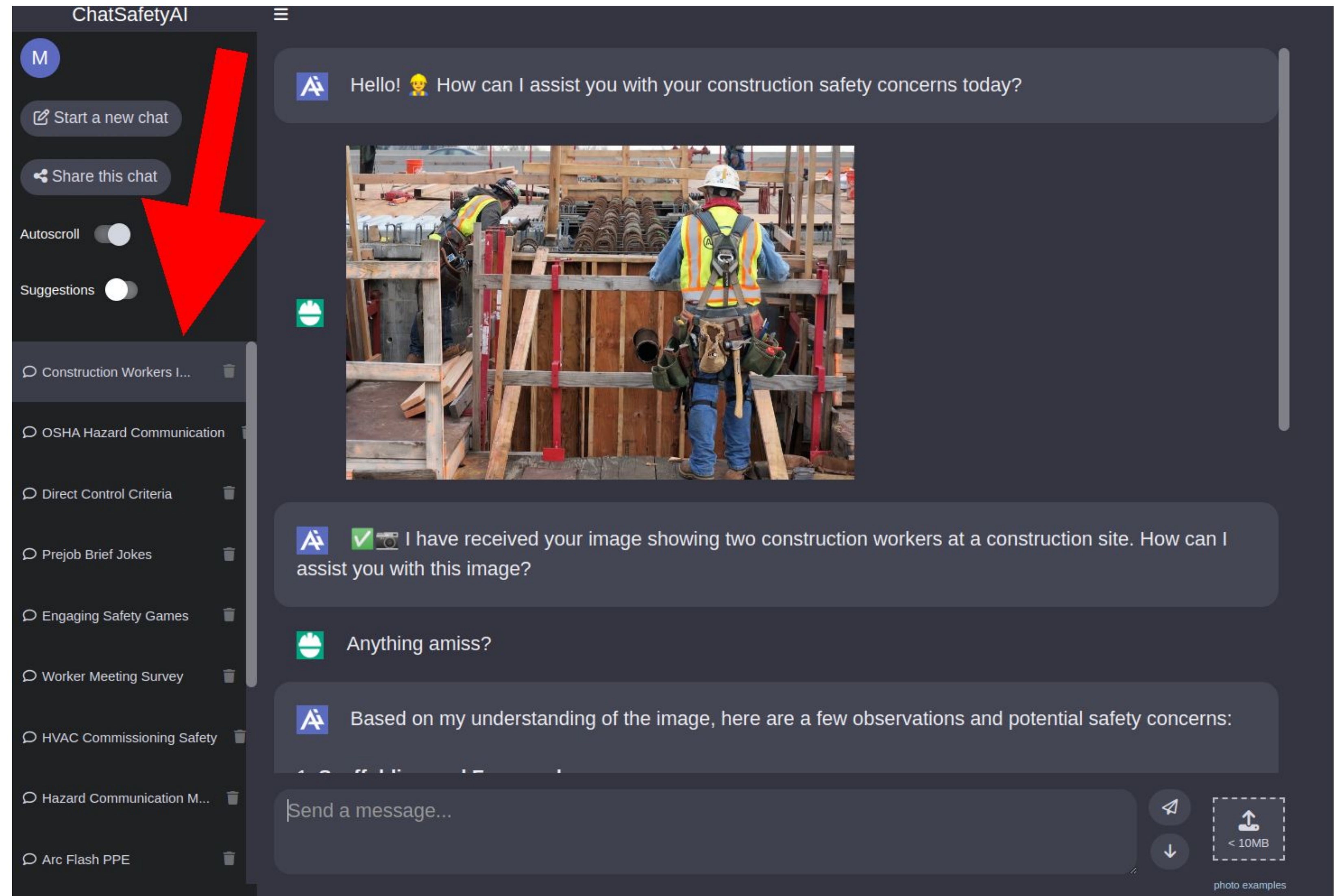
- New instructions with unconstrained output format, to better suit generation scenarios and meet all user requests.
- (Before, the regulation mode was limited to reorganizing and wrapping search engine results with fixed steps and output format).

Chat History

- Conversations and associated data are now saved and reloaded across sessions
- Saved in a non human-readable format (compressed binary files) to a secure database, using data encryption in transit and at rest
- Data not visible by anyone but me: random checks for continuous improvement purposes. Email me to opt out.

Chat History

- Every chat you have with ChatSafetyAI will be in the sidebar.
- When deleting a chat, the data and files associated with it are immediately removed from the database.



Off-topic Classifier

- Used for text, file, and photo inputs
- Crucial to prevent adversarial / unintended uses
- But annoying “Let’s focus on construction safety” false alarms
- I refactored the classifier
- It now takes the conversation into account to better understand queries in context
- I also gave it less stringent instructions
- Not perfect, but tests show it’s better now

Other Improvements

- More budget for file uploads (330k chars, which is about 120 single-spaced pages, compared to 75 in the previous version)
- Shorter greetings to save time and budget
- Many other small bug fixes and improvements...

Next Steps

- 1) Ability to call company's custom predictive models.
- 2) Knowledge of the company's own data: incident and observation datasets, internal documentation, wiki...
- 3) Learning preferences of each user (language, role, ideal output format and length...)
- 4) Incognito mode (no data persisted on session end)
- 5) Any other wish! Ideas for new modes?

=> I'll send a quick survey to prioritize/add items

Next Steps

- Call for data updates for the predictive models, SafetyAI Assistant, and SafetyAI Dashboard
 - Need the data from everyone to start updating the community models (Super Models, aka generic models)
 - This year, as predictors, I will include weather and climate variables in addition to attributes!

Usage Stats

- As of today, ChatSafetyAI is being used by 230 users, and the demo version by 404 users (users are unique and verified)
- 5.9k conversations and 228M tokens consumed so far
- I send new invitations every day.

Discussion & Feedback

Thank you!

Get access to the full app:

https://safetyapp.shinyapps.io/chatsafetyai_access_request/

Get access to the demo:

https://safetyapp.shinyapps.io/chatsafetyai_demo_access_request/

Full app lives at: <https://safetyapp.shinyapps.io/chatsafetyai/>

Demo lives at: https://safetyapp.shinyapps.io/chatsafetyai_demo/

Thank you!

