Course introduction

CMSC 848O, Spring 2025

Advanced Natural Language Processing

https://cs.umd.edu/~miyyer/cmsc848o

Mohit Iyyer

University of Maryland, College Park

About me

- I did my PhD at UMD CS from 2012-2017
  - *in AV Williams, not Iribe :(*
- I was at AI2 from 2017-2018, doing research on *small* language models
- I was then a prof at UMass Amherst until earlier this month
- Recently, my lab's research focus has been long-context LLMs!

Course logistics

- Three weeks of me lecturing, so that we’re all roughly on the same page
- Rest of semester: student presentations and discussions of assigned papers

other logistics

TA:

Chau Pham

email both of us at
longcontextseminar@gmail.com

course website:
https://cs.umd.edu/~miyyer/cmsc848o

office hours:
Thursdays 2-3pm, IRB 4142

can you get in off the waitlist?

- we don’t control the waitlist, so it’s out of our hands!
- enrollment will not increase beyond 50 students
- everyone is welcome to sit in on the class even if you’re unable to officially enroll!

what background is expected?

- basic ML/probability/stats/linear algebra/programming will help a lot
    - we won’t have any coding assignments, but we will be reading implementation-heavy papers and sometimes also their codebases
- prior knowledge of how LLMs are trained and deployed will certainly help
    - that said, don’t worry if you’re totally new to the field… we will be covering the basics over the next 3 weeks

If you want to review basic LLM/ NLP concepts on your own time

- I taught a class focusing on LLMs in Spring 2024: https://cs.umd.edu/~miyyer/cs685
  - Feel free to use these materials / videos to study!
  - It will be especially good to supplement the lectures in the next three weeks

Grading breakdown

- 20% writing assignments (hw1, hw2)
- 25% discussion question submission + in-class participation
- 30% exam (~early April, **in-class exam**)
- 25% presentations of assigned papers

Readings

- No need to buy any textbooks!
- Readings will be provided as PDFs on website
    - Usually NLP research papers / notes

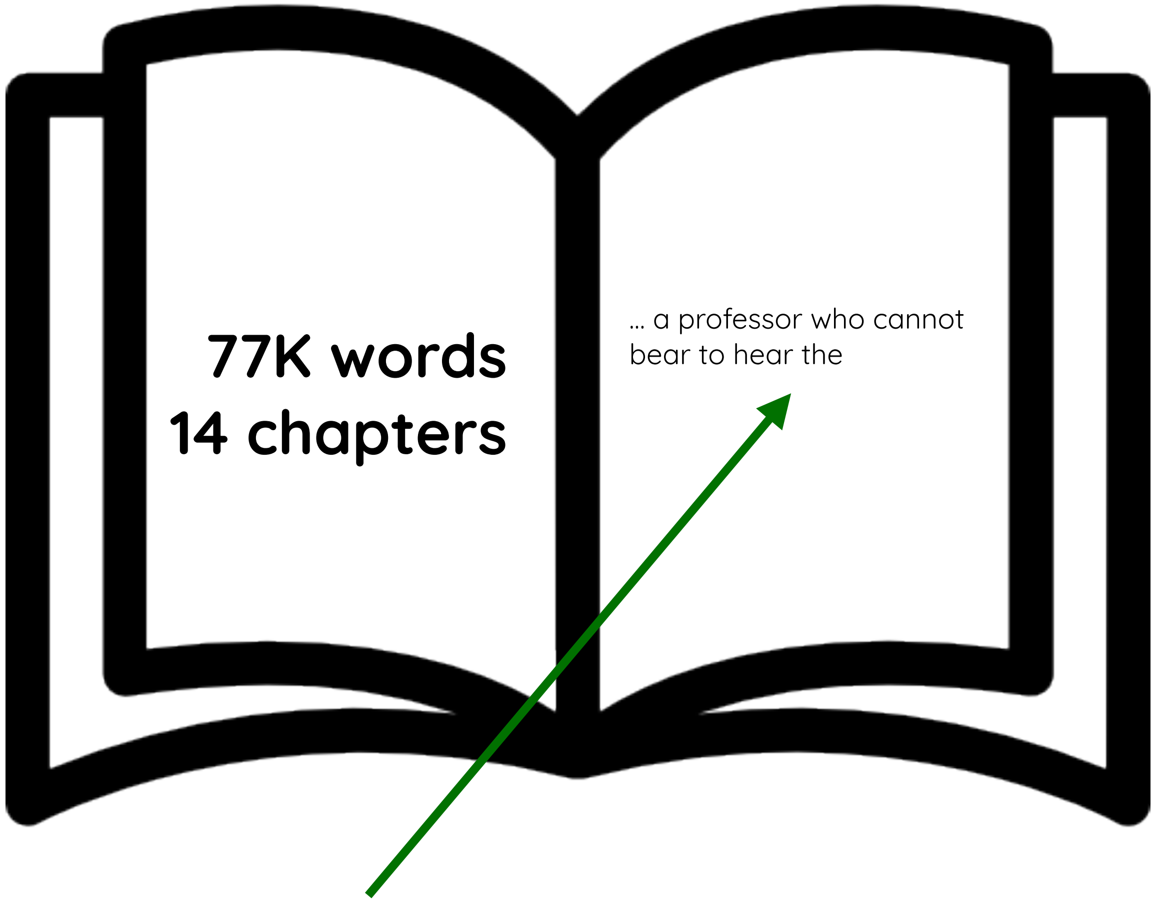
About you

Prompt: I’m preparing to teach my first class at UMD, which is a seminar on long-context language models for 50 graduate students. The students come from many different math and research backgrounds. What's a good set of non-lame questions for me to ask them on the very first day to get an idea of their background and interests? All questions should require just a show of hands in response.

Language models perform next-word prediction.

… a $\underbrace{\text{professor who cannot bear to hear the}}_{\text{prefix}}$ ____

Predicting the next word can require reasoning over long contexts.



77K words
14 chapters

… a professor who cannot bear to hear the

… a professor who cannot bear to hear the ______

Italo Calvino. “If on a winter’s night a traveler” (1979)

Sometimes, successfully predicting the next word requires reasoning over long contexts.

📖 **23,953 words prior:**

Every morning before my classes begin I do an hour of jogging... as many students do and also many of my colleagues.

... a professor who cannot bear to hear the ______

Italo Calvino. “If on a winter’s night a traveler” (1979)

Sometimes, successfully predicting the next word requires reasoning over long contexts.

📖 **23,953 words prior:**

Every morning before **my classes** begin I do an hour of jogging... as many **students** do and also many of **my colleagues**.

... a **professor** who cannot bear to hear the______

Italo Calvino. “If on a winter’s night a traveler” (1979)

Sometimes, successfully predicting the next word requires reasoning over long contexts.

📖 **23,953 words prior:**

Every morning before my classes begin I do an hour of jogging… as many students do and also many of my colleagues.

📖 **22,501 words prior:**

I am the **prisoner of… the telephone** ringing inside that house… There is a **telephone chasing me**…

… a professor who **cannot bear to hear** the ______

Italo Calvino. “If on a winter’s night a traveler” (1979)

Sometimes, successfully predicting the next word requires reasoning over long contexts.

📖 **23,953 words prior:**

Every morning before my classes begin I do an hour of jogging… as many students do and also many of my colleagues.

📖 **22,501 words prior:**

I am the prisoner of… the telephone ringing inside that house… There is a telephone chasing me…

… a professor who cannot bear to hear the **telephone**

Italo Calvino. “If on a winter’s night a traveler” (1979)

What is a “long” context?

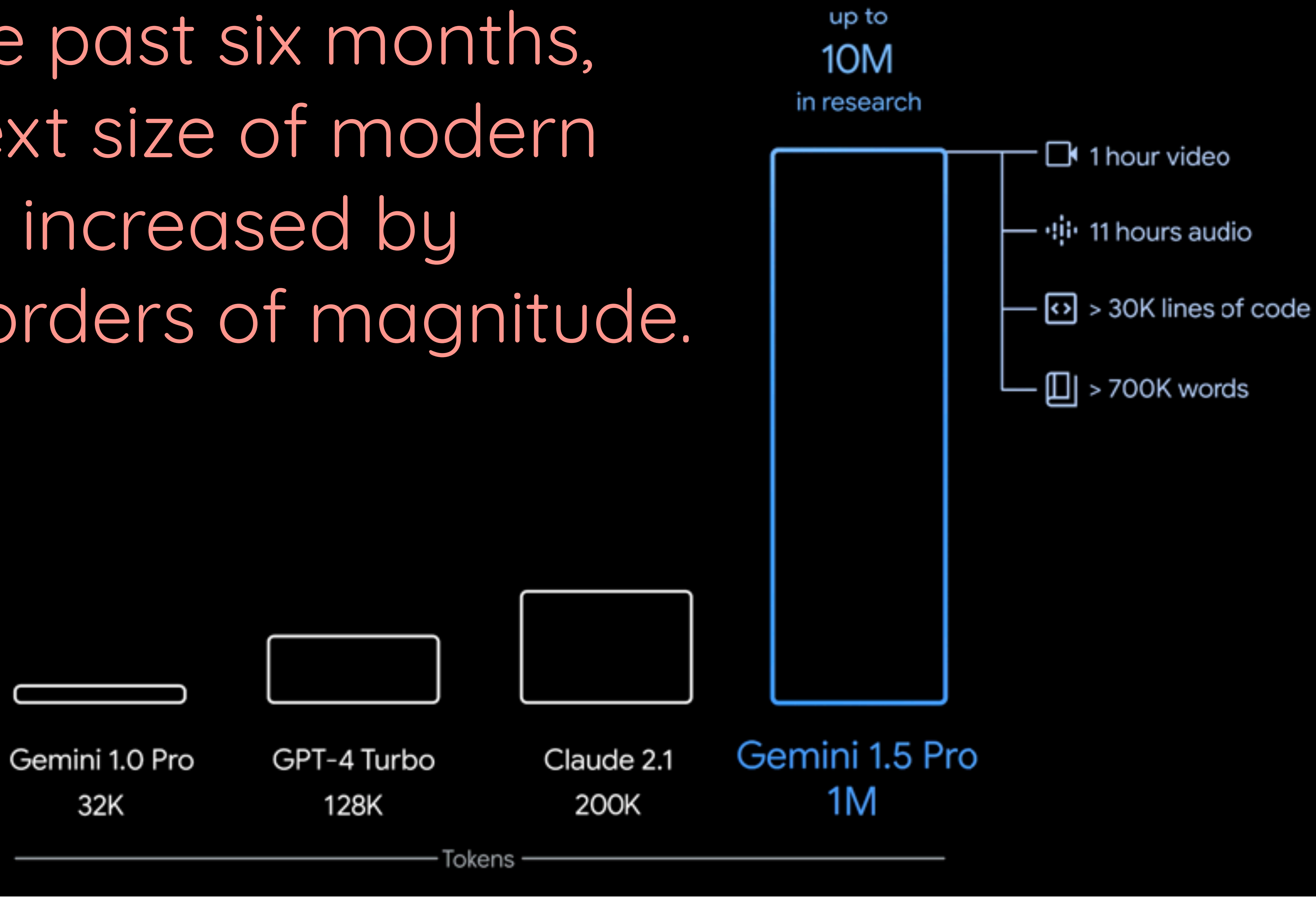
BERT / GPT (2018): 512 tokens

GPT-2 (2019): 1K tokens

ChatGPT (2022): 4K tokens

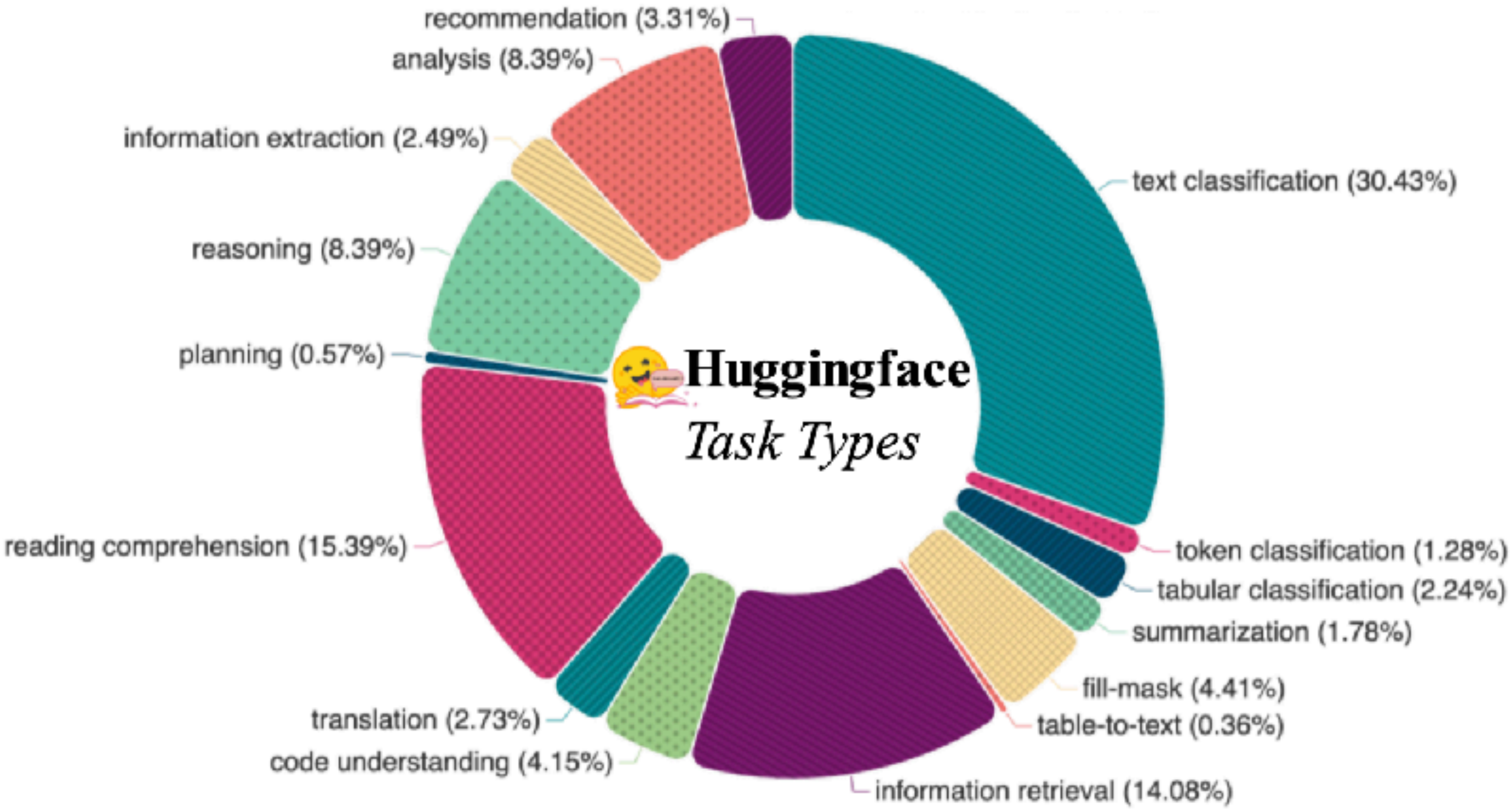
GPT-4 (2023): 8K tokens

Within the past six months, the context size of modern LLMs has increased by multiple orders of magnitude.



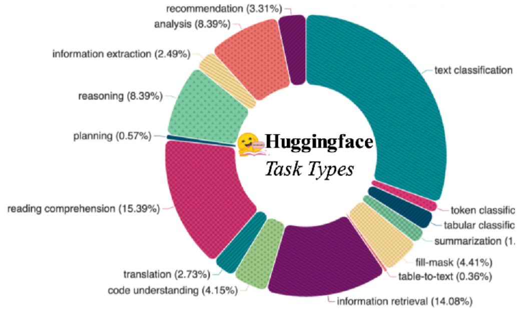
https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/

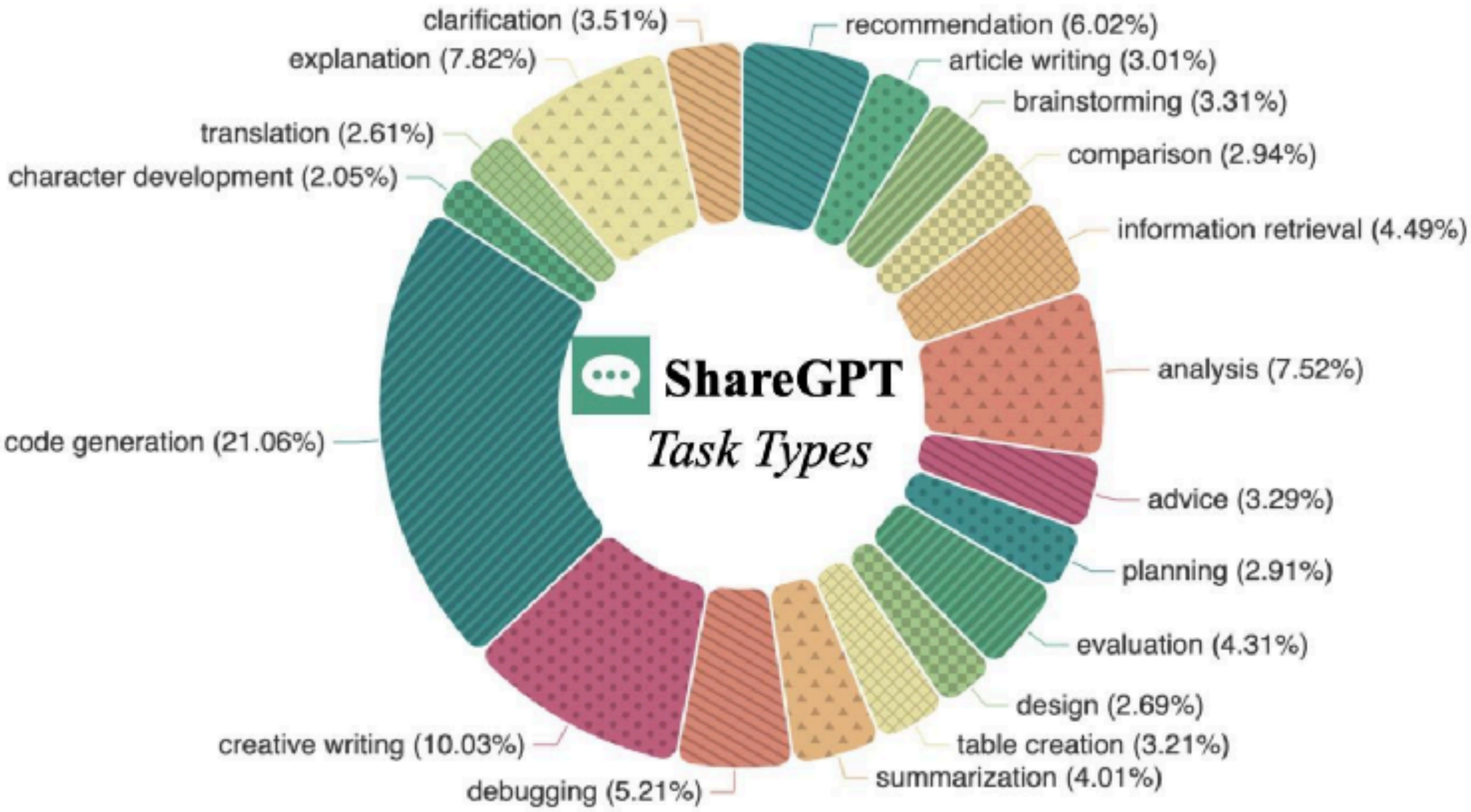
We are witnessing a shift away from short-form tasks...



Ouyang et al., 2023. "The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions"

… to **long-form** generation tasks like creative writing & coding.





Ouyang et al., 2023. "The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions"

Rough list of topics

- **Background**: language models, Transformers, LLM training cycle, evaluations
- **Student-led paper topics**:
    - Extending LLMs from short context to long context: continual pretraining, mid-training, post-training
    - Efficient attention mechanisms: pros and cons
    - Architectural modifications: state space models (e.g., Mamba) and hybrid models (e.g., Jamba)
    - Efficient implementations of vanilla attention: flash attention, ring attention
    - Evaluation of long-context language models: perplexity, point-wise retrieval, summarization, QA, etc.
    - Synthetic data generation for long-context instruction following and reasoning
    - Generating long outputs from long inputs
    - Long context vs. RAG

Timeline for student presentations

- Presentation groups should be formed by **2/7**
    - Groups of 3, either form them yourselves and tell us, or we will randomly assign you on 2/7
- Presentation format: 15-20 mins followed by 15-20 min discussion
    - Papers assigned by instructor
    - Each group member takes a different role:
        - Presenter, reviewer, archaeologist, etc.
- Everyone must submit discussion questions on assigned papers prior to the start of every class
    - I will call on ppl at random to ask their questions if discussion starts to stagnate!

Be on the lookout for

- **HW1:** released this week, due 2/14
- Readings on language models for Thursday
- Group assignment logistics on Piazza