LLMs for Research: Content Generation

Ethan Tran

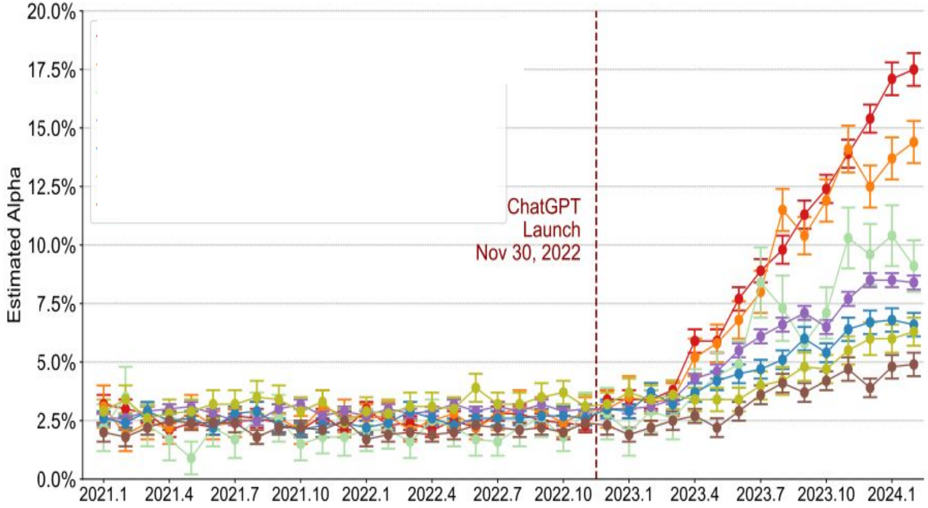
April 3rd, 2025

Agenda

- Mapping the Increasing Use of LLMs in Scientific Papers
- Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews
- LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts

Problem Overview and Implications

- Rapid adoption of LLM assistance in academic writing and review.
- Growing need to identify and quantify LLM modified content across different disciplines.



Estimated Fraction of LLM-Modified Sentences in Abstracts across Academic Writing Venues over Time

Image Credit: Liang et al. (2024). "Mapping the Increasing Use of LLMs in Scientific Papers." arXiv:2404.01268.

Mapping the Increasing Use of LLMs in Scientific Papers: Background

- Previous proposed methods for GPT Detectors
  - Zero-shot approach
  - Binary Classification
  - Watermarking*

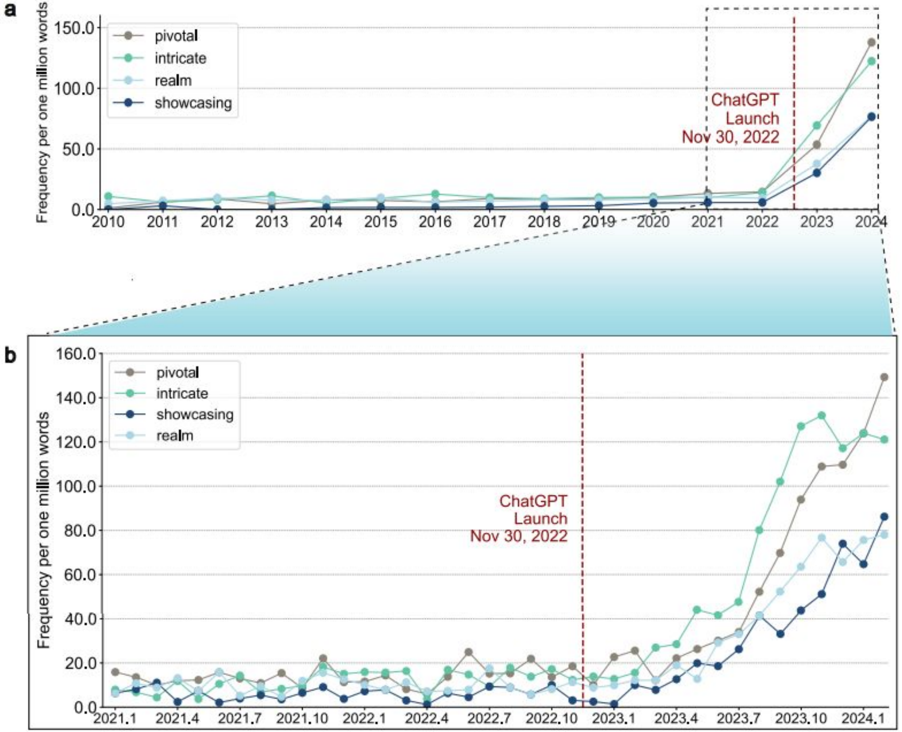
Mapping the Increasing Use of LLMs in Scientific Papers: Background

- Questions of both robustness and reliability still remained
  - Overfitting to a specific LLM
  - Possible circumventions
  - Bias against non-native English Speakers

- Introduction to a probabilistic method to classify LLM-modified content.

Distributional LLM Quantification: Idea

- Proposed statistical method to estimate the proportion of LLM modified content.
- Maximum likelihood estimation based on **token** occurrences.



Word Frequency Shift in arXiv Computer Science abstracts over 14 years

Distributional LLM Quantification: Generating Training Data

- Data collected from arXiv, bioRxiv, and Nature portfolio.
    - 2,000 samples from each from Jan 2020 - Feb 2024.
- A two stage approach to generate LLM produced data
    - Step 1: Given a paper known not to have LLM modifications, use an LLM to generate a bulleted list summary.
    - Step 2: Use an LLM (gpt-3.5) to generate a paragraph based on that outline.

LLM prompt

```
The aim here is to reverse-engineer the author's writing process by
    taking a piece of text from a paper and compressing it into a more
    concise form. This process simulates how an author might distill
    their thoughts and key points into a structured, yet not overly
    condensed form.

Now as a first step, first summarize the goal of the text, e.g., is it
    introduction, or method, results? and then given a complete piece of
    text from a paper, reverse-engineer it into a list of bullet points.
```

```
Following the initial step of reverse-engineering the author's writing
    process by compressing a text segment from a paper, you now enter the
     second phase. Here, your objective is to expand upon the concise
    version previously crafted. This stage simulates how an author
    elaborates on the distilled thoughts and key points, enriching them
    into a detailed, structured narrative.

Given the concise output from the previous step, your task is to develop
    it into a fully fleshed-out text.
```

```
Your task is to proofread the provided sentence for grammatical accuracy.
     Ensure that the corrections introduce minimal distortion to the
    original content.
```

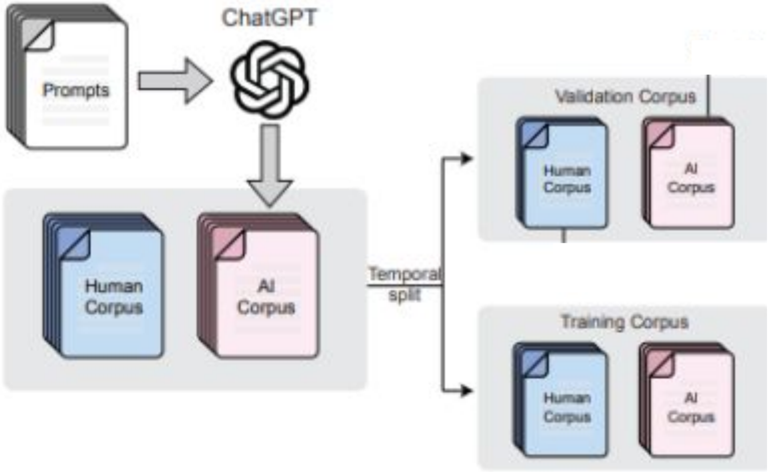
Distributional LLM Quantification: Mathematical Formulation

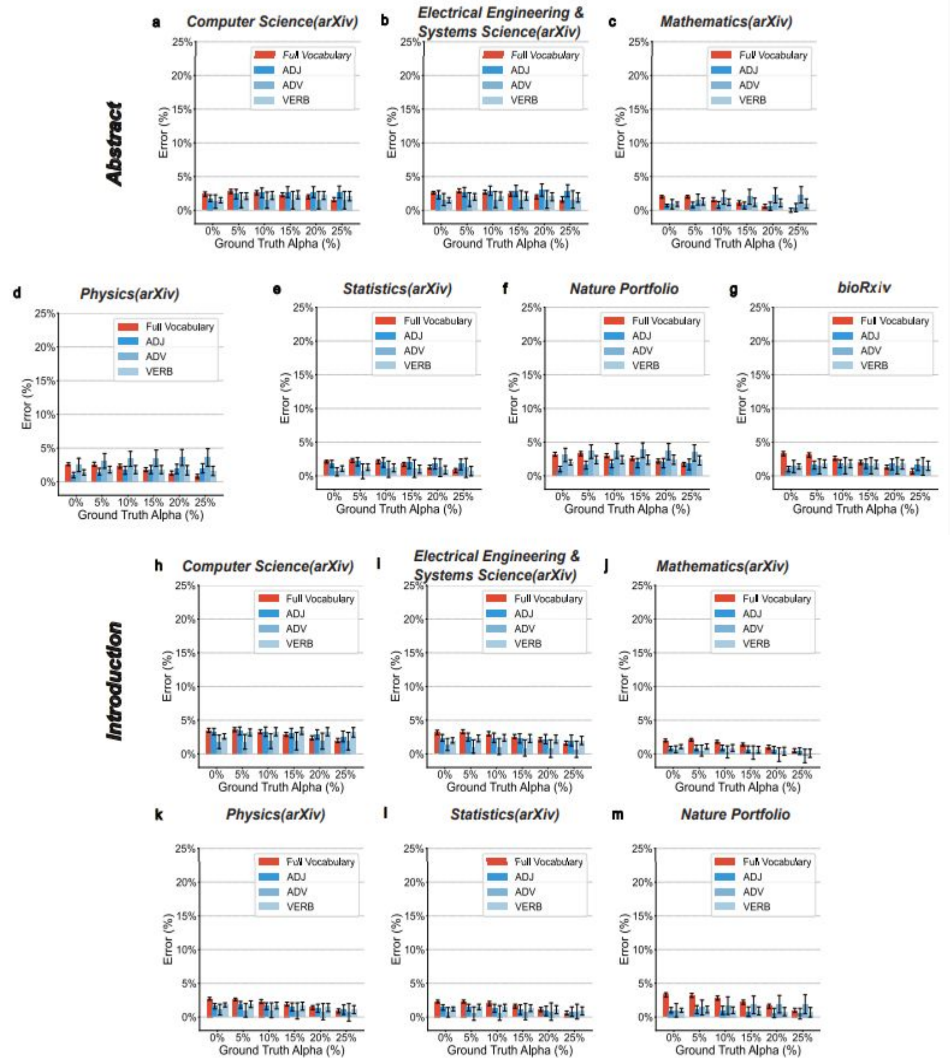
- Let $\mathcal{P}$ and $\mathcal{Q}$ be the probability distributions of human-written and LLM modified documents respectively and let X be the corpus of documents.
- The mixture distribution is given by $\mathcal{D}_a(X) = (1-\alpha)\mathcal{P}(x) + \alpha\mathcal{Q}(x)$, where α is the desired estimate of AI-modified document based on observed documents.
- Parameterizing the distribution of P and Q based on estimating the probability of token occurrence.
- Let $A = \{0, 0.05, 0.1, 0.15, 0.2, 0.25\}$.The optimal fraction α is given by maximizing the log-likelihood of the observed documents under the parameterized mixture distribution $\hat{\mathcal{D}}_{a,T} = (1-\alpha)\hat{\mathcal{P}}_T(x) + \alpha\hat{\mathcal{Q}}_T(x)$

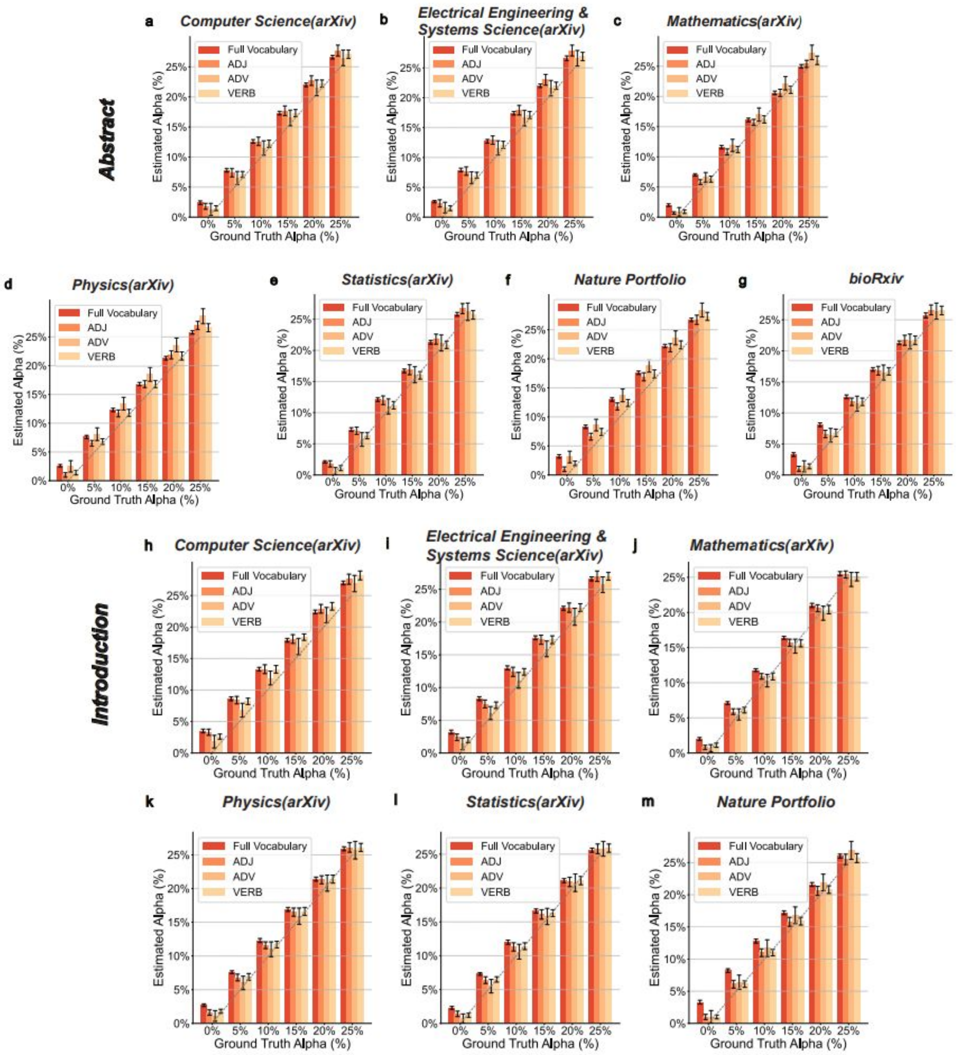
$$\hat{\alpha}_T^{\mathrm{MLE}} = \underset{\alpha \in A}{\operatorname{argmax}} = \sum_{i=1}^{N} \log((1-\alpha)\hat{\mathcal{P}}_T(x) + \alpha\hat{\mathcal{Q}}_T(x))$$

Data Split, Model Fit and Evaluation

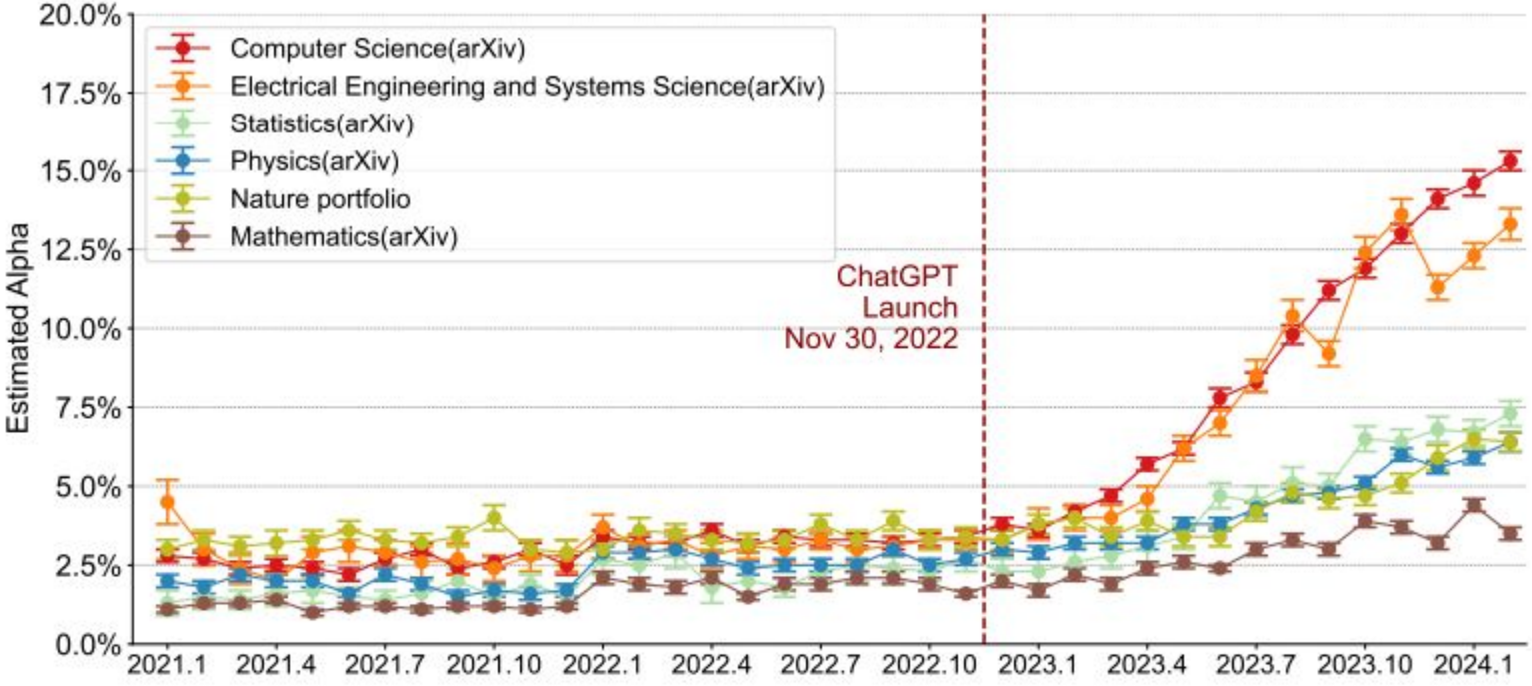
- Separate models for abstracts and introductions.
- Model fitted on data from 2020 and validated with data from 2021 on.
    - Grouped between pre and post ChatGPT era
- Ability to construct a validation set for a ground truth α.




Abstract
a Computer Science(arXiv)
b Electrical Engineering & Systems Science(arXiv)
c Mathematics(arXiv)
d Physics(arXiv)
e Statistics(arXiv)
f Nature Portfolio
g bioRxiv
Introduction
h Computer Science(arXiv)
i Electrical Engineering & Systems Science(arXiv)
j Mathematics(arXiv)
k Physics(arXiv)
l Statistics(arXiv)
m Nature Portfolio
Full Vocabulary
ADJ
ADV
VERB
Error (%)
0%
5%
10%
15%
20%
25%
Ground Truth Alpha (%)


Abstract
a Computer Science(arXiv)
b Electrical Engineering & Systems Science(arXiv)
c Mathematics(arXiv)
d Physics(arXiv)
e Statistics(arXiv)
f Nature Portfolio
g bioRxiv
Introduction
h Computer Science(arXiv)
i Electrical Engineering & Systems Science(arXiv)
j Mathematics(arXiv)
k Physics(arXiv)
l Statistics(arXiv)
m Nature Portfolio
Full Vocabulary
ADJ
ADV
VERB
Estimated Alpha (%)
0%
5%
10%
15%
20%
25%
Ground Truth Alpha (%)

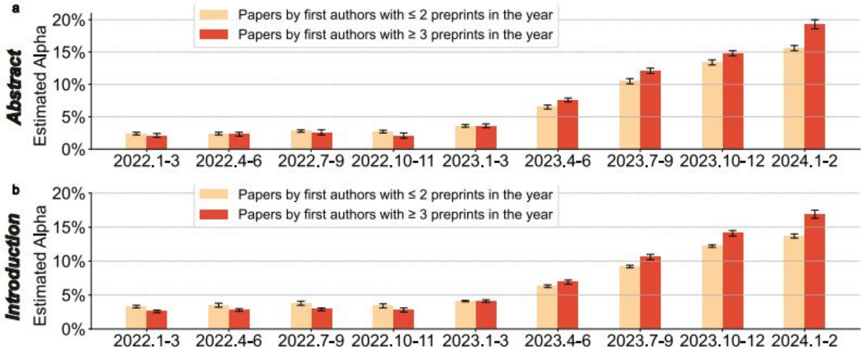
Results

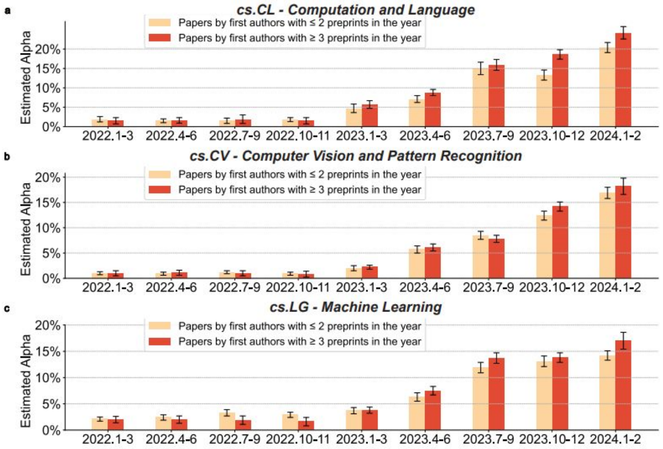


Computer Science(arXiv)
Electrical Engineering and Systems Science(arXiv)
Statistics(arXiv)
Physics(arXiv)
Nature portfolio
Mathematics(arXiv)
ChatGPT
Launch
Nov 30, 2022
Estimated Alpha
20.0%
17.5%
15.0%
12.5%
10.0%
7.5%
5.0%
2.5%
0.0%
2021.1
2021.4
2021.7
2021.10
2022.1
2022.4
2022.7
2022.10
2023.1
2023.4
2023.7
2023.10
2024.1

Relationships: Posting Frequency and LLM Usage

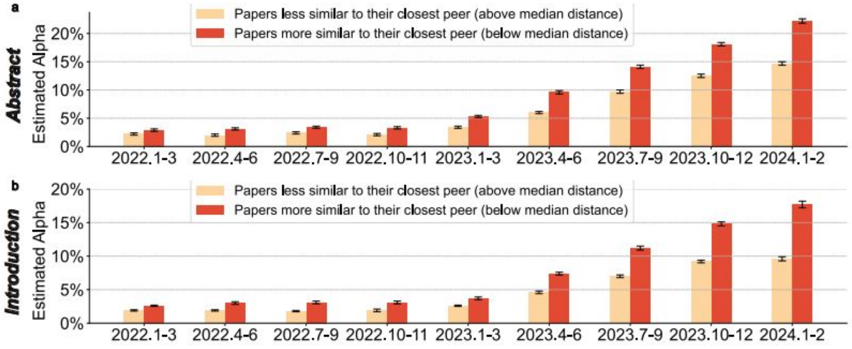
- Papers by authors with ≥3 preprints exhibit higher LLM-modified content.
  - Shown that 19.4% of sentences were modified by AI, compared to 15.6% for authors with <3 preprints.
  - Results are constant across sub-categories of CV, ML, and Computation.

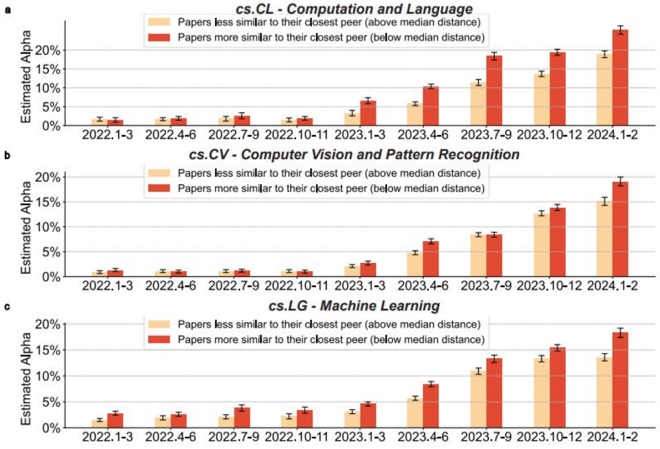




Relationships: Paper Similarity and LLM Usage

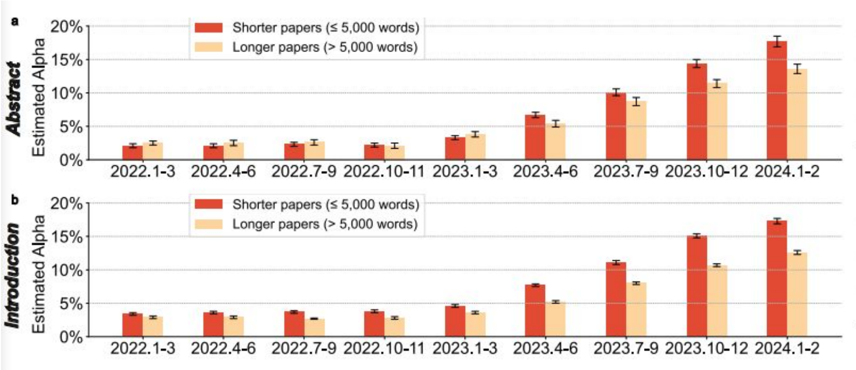
- Papers embedded as vectors and grouped by similarity score.
  - “More similar” papers had ~22.2% LLM modification vs. ~14.7% for “less similar.”
  - Consistent trends observed across sub-categories

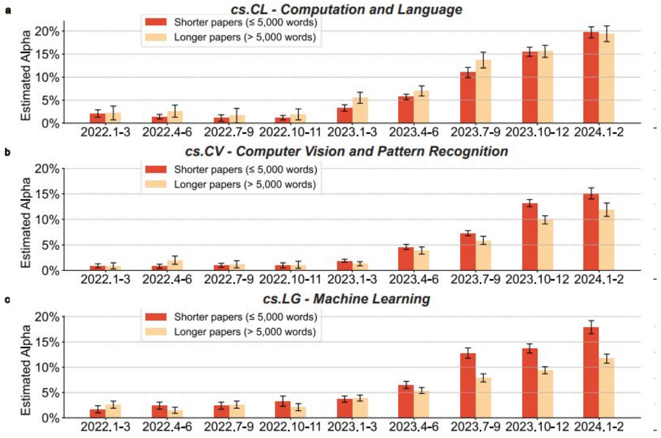




Relationships: Paper Length and LLM Usage

- Shorter papers (≤5,000 words) exhibit higher LLM-modified content compared to longer papers
  - Abstracts of shorter papers showed ~17.7% modification vs. ~13.6% for longer papers.
  - Trend holds for sub-categories such as cs.CV and cs.LG, but not for cs.CL.





Limitations

- Solely focuses on ChatGPT.
  - The framework might not fully capture nuances from other LLMs
- Increased LLM research
  - Research on LLMs after ChatGPT's release might inadvertently affect the method's accuracy.
- Language Shifts
  - Prior work shows some detection methods might mistakenly flag texts by non-native speakers

Agenda

- Mapping the Increasing Use of LLMs in Scientific Papers
- Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews
- LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts

A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews

- Context and Motivation
  - Study focuses on the use of LLM assistance in the peer review process.
  - Examines how AI influences the quality of evaluation
- Uses the distributional GPT quantification framework.
- Paper explores further relationships between various circumstances and AI use.

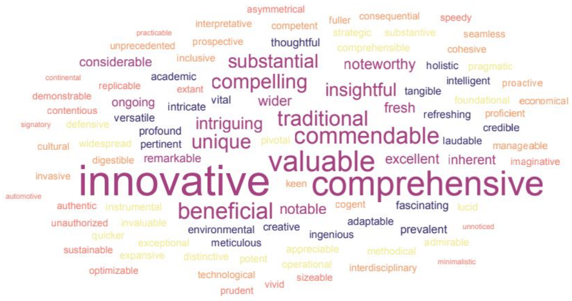
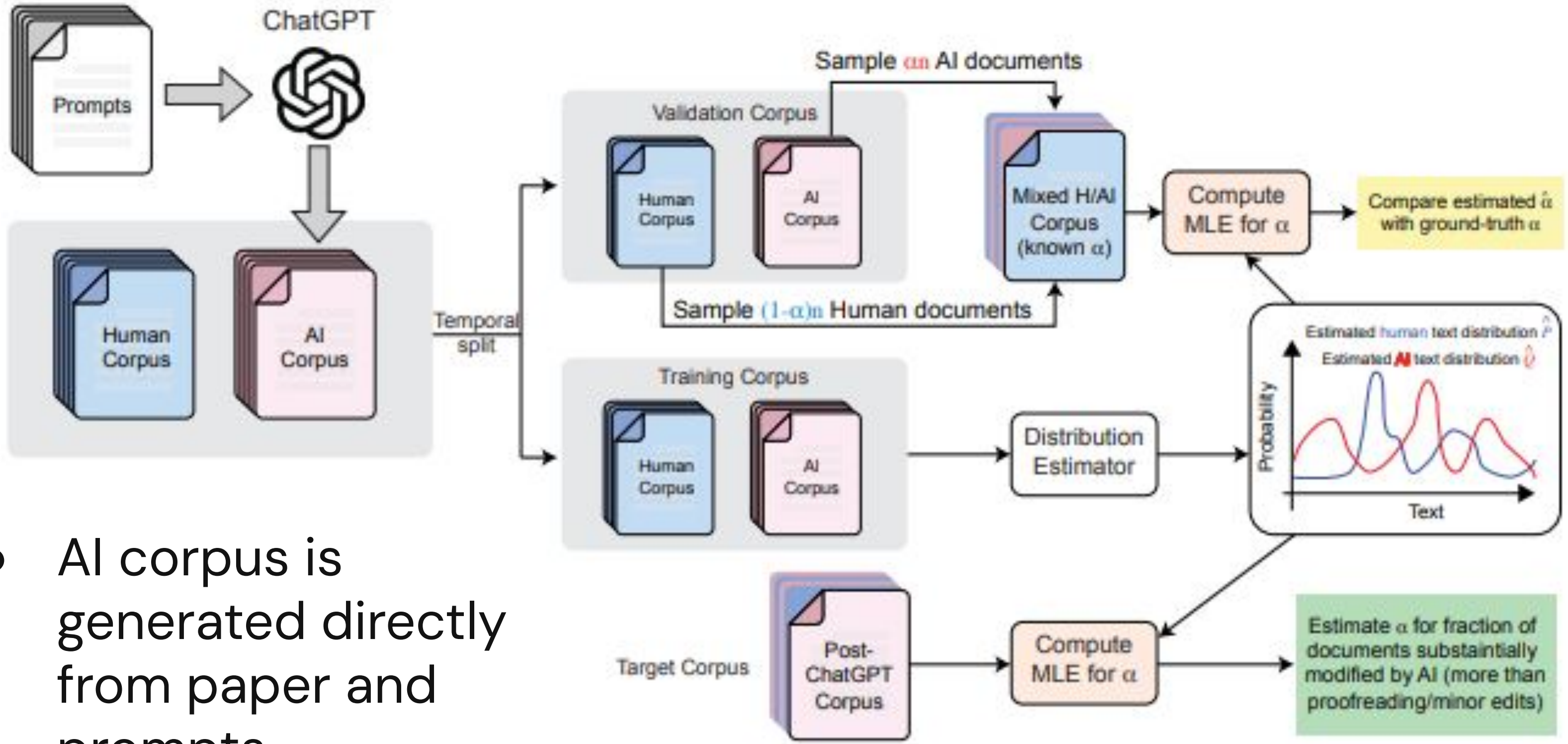


Figure 12: Word cloud of top 100 adjectives in LLM feedback, with font size indicating frequency.

Method Overview and Validation



- AI corpus is generated directly from paper and prompts

Data Collection and Evaluation on Validation Set

Conference	Post ChatGPT	Data Split	# of Official Reviews
ICLR 2018	Before	Training	2,930
ICLR 2019	Before	Training	4,764
ICLR 2020	Before	Training	7,772
ICLR 2021	Before	Training	11,505
ICLR 2022	Before	Training	13,161
ICLR 2023	Before	Validation	18,564
ICLR 2024	After	Inference	27,992
NeurIPS 2017	Before	OOD Validation	1,976
NeurIPS 2018	Before	OOD Validation	3,096
NeurIPS 2019	Before	OOD Validation	4,396
NeurIPS 2020	Before	OOD Validation	7,271
NeurIPS 2021	Before	OOD Validation	10,217
NeurIPS 2022	Before	OOD Validation	9,780
NeurIPS 2023	After	Inference	14,389
CoRL 2021	Before	OOD Validation	558
CoRL 2022	Before	OOD Validation	756
CoRL 2023	After	Inference	759
EMNLP 2023	After	Inference	6,419

No.	Validation Data Source	Ground Truth α	Estimated		Prediction Error
			α	CI ($\pm$)	
(1)	ICLR 2023	0.0%	1.6%	0.1%	1.6%
(2)	ICLR 2023	2.5%	4.0%	0.5%	1.5%
(3)	ICLR 2023	5.0%	6.2%	0.6%	1.2%
(4)	ICLR 2023	7.5%	8.3%	0.6%	0.8%
(5)	ICLR 2023	10.0%	10.5%	0.6%	0.5%
(6)	ICLR 2023	12.5%	12.6%	0.7%	0.1%
(7)	ICLR 2023	15.0%	14.7%	0.7%	0.3%
(8)	ICLR 2023	17.5%	16.9%	0.7%	0.6%
(9)	ICLR 2023	20.0%	19.0%	0.8%	1.0%
(10)	ICLR 2023	22.5%	21.1%	0.9%	1.4%
(11)	ICLR 2023	25.0%	23.3%	0.8%	1.7%
(12)	NeurIPS 2022	0.0%	1.8%	0.2%	1.8%
(13)	NeurIPS 2022	2.5%	4.4%	0.5%	1.9%
(14)	NeurIPS 2022	5.0%	6.6%	0.6%	1.6%
(15)	NeurIPS 2022	7.5%	8.8%	0.7%	1.3%
(16)	NeurIPS 2022	10.0%	11.0%	0.7%	1.0%
(17)	NeurIPS 2022	12.5%	13.2%	0.7%	0.7%
(18)	NeurIPS 2022	15.0%	15.4%	0.8%	0.4%
(19)	NeurIPS 2022	17.5%	17.6%	0.7%	0.1%
(20)	NeurIPS 2022	20.0%	19.8%	0.8%	0.2%
(21)	NeurIPS 2022	22.5%	21.9%	0.8%	0.6%
(22)	NeurIPS 2022	25.0%	24.1%	0.8%	0.9%
(23)	CoRL 2022	0.0%	2.4%	0.6%	2.4%
(24)	CoRL 2022	2.5%	4.6%	0.6%	2.1%
(25)	CoRL 2022	5.0%	6.8%	0.6%	1.8%
(26)	CoRL 2022	7.5%	8.8%	0.7%	1.3%
(27)	CoRL 2022	10.0%	10.9%	0.7%	0.9%
(28)	CoRL 2022	12.5%	13.0%	0.7%	0.5%
(29)	CoRL 2022	15.0%	15.0%	0.8%	0.0%
(30)	CoRL 2022	17.5%	17.0%	0.8%	0.5%
(31)	CoRL 2022	20.0%	19.1%	0.8%	0.9%
(32)	CoRL 2022	22.5%	21.1%	0.8%	1.4%
(33)	CoRL 2022	25.0%	23.2%	0.8%	1.8%

Further Evaluation

No.	Validation Data Source	Ground Truth α	Estimated α	Estimated CI ($\pm$)	Prediction Error
(1)	ICLR 2023	0.0%	1.6%	0.1%	1.6%
(2)	ICLR 2023	2.5%	4.0%	0.5%	1.5%
(3)	ICLR 2023	5.0%	6.2%	0.6%	1.2%
(4)	ICLR 2023	7.5%	8.3%	0.6%	0.8%
(5)	ICLR 2023	10.0%	10.5%	0.6%	0.5%
(6)	ICLR 2023	12.5%	12.6%	0.7%	0.1%
(7)	ICLR 2023	15.0%	14.7%	0.7%	0.3%
(8)	ICLR 2023	17.5%	16.9%	0.7%	0.6%
(9)	ICLR 2023	20.0%	19.0%	0.8%	1.0%
(10)	ICLR 2023	22.5%	21.1%	0.9%	1.4%
(11)	ICLR 2023	25.0%	23.3%	0.8%	1.7%
(12)	NeurIPS 2022	0.0%	1.8%	0.2%	1.8%
(13)	NeurIPS 2022	2.5%	4.4%	0.5%	1.9%
(14)	NeurIPS 2022	5.0%	6.6%	0.6%	1.6%
(15)	NeurIPS 2022	7.5%	8.8%	0.7%	1.3%
(16)	NeurIPS 2022	10.0%	11.0%	0.7%	1.0%
(17)	NeurIPS 2022	12.5%	13.2%	0.7%	0.7%
(18)	NeurIPS 2022	15.0%	15.4%	0.8%	0.4%
(19)	NeurIPS 2022	17.5%	17.6%	0.7%	0.1%
(20)	NeurIPS 2022	20.0%	19.8%	0.8%	0.2%
(21)	NeurIPS 2022	22.5%	21.9%	0.8%	0.6%
(22)	NeurIPS 2022	25.0%	24.1%	0.8%	0.9%
(23)	CoRL 2022	0.0%	2.4%	0.6%	2.4%
(24)	CoRL 2022	2.5%	4.6%	0.6%	2.1%
(25)	CoRL 2022	5.0%	6.8%	0.6%	1.8%
(26)	CoRL 2022	7.5%	8.8%	0.7%	1.3%
(27)	CoRL 2022	10.0%	10.9%	0.7%	0.9%
(28)	CoRL 2022	12.5%	13.0%	0.7%	0.5%
(29)	CoRL 2022	15.0%	15.0%	0.8%	0.0%
(30)	CoRL 2022	17.5%	17.0%	0.8%	0.5%
(31)	CoRL 2022	20.0%	19.1%	0.8%	0.9%
(32)	CoRL 2022	22.5%	21.1%	0.8%	1.4%
(33)	CoRL 2022	25.0%	23.2%	0.8%	1.8%

Comparison to Instance-Based Detection Methods

- Benchmarking against recently published AI detection methods
  - GPT quantification shown to be optimal in reducing prediction error.
  - 10 million times more computationally efficient during inference time
  - Robustness to Proofreading

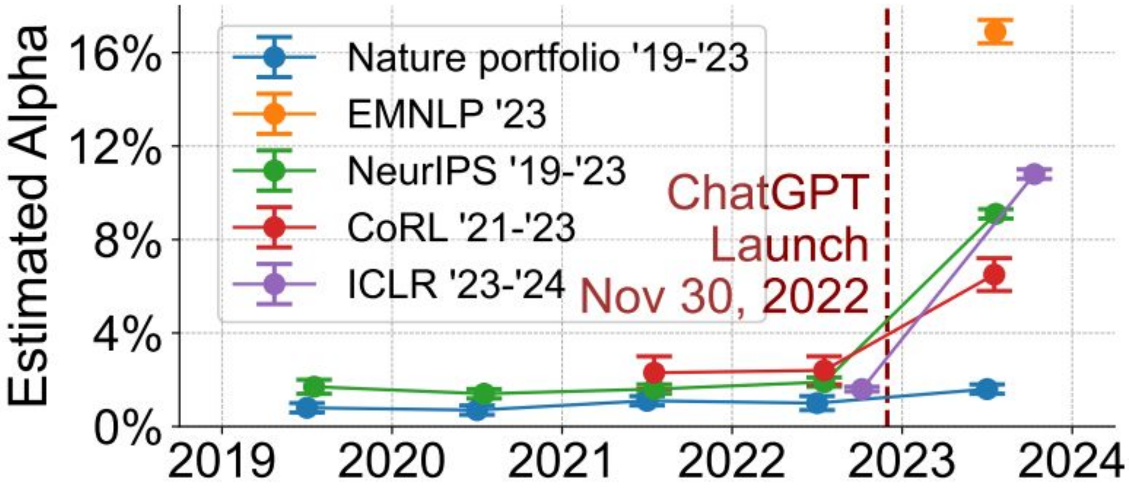
No.	Validation Data Source	Ground Truth α	RADAR Estimated α	Deepfake Estimated α	Fast-DetectGPT Estimated α	BERT Estimated α	BERT Predictor Error
(1)	ICLR 2023	0.0%	99.3%	0.2%	11.3%	1.1%	1.1%
(2)	ICLR 2023	2.5%	99.4%	0.2%	11.2%	2.9%	0.4%
(3)	ICLR 2023	5.0%	99.4%	0.3%	11.2%	4.7%	0.3%
(4)	ICLR 2023	7.5%	99.4%	0.2%	11.4%	6.4%	1.1%
(5)	ICLR 2023	10.0%	99.4%	0.2%	11.6%	8.0%	2.0%
(6)	ICLR 2023	12.5%	99.4%	0.3%	11.6%	9.9%	2.6%
(7)	ICLR 2023	15.0%	99.4%	0.3%	11.8%	11.6%	3.4%
(8)	ICLR 2023	17.5%	99.4%	0.2%	11.9%	13.4%	4.1%
(9)	ICLR 2023	20.0%	99.4%	0.3%	12.2%	15.3%	4.7%
(10)	ICLR 2023	22.5%	99.4%	0.2%	12.0%	17.0%	5.5%
(11)	ICLR 2023	25.0%	99.4%	0.3%	12.1%	18.8%	6.2%
(12)	NeurIPS 2022	0.0%	99.2%	0.2%	10.5%	1.1%	1.1%
(13)	NeurIPS 2022	2.5%	99.2%	0.2%	10.5%	2.3%	0.2%
(14)	NeurIPS 2022	5.0%	99.2%	0.3%	10.7%	3.6%	1.4%
(15)	NeurIPS 2022	7.5%	99.2%	0.2%	10.9%	5.0%	2.5%
(16)	NeurIPS 2022	10.0%	99.2%	0.2%	10.9%	6.1%	3.9%
(17)	NeurIPS 2022	12.5%	99.2%	0.3%	11.1%	7.2%	5.3%
(18)	NeurIPS 2022	15.0%	99.2%	0.3%	11.0%	8.6%	6.4%
(19)	NeurIPS 2022	17.5%	99.3%	0.2%	11.0%	9.9%	7.6%
(20)	NeurIPS 2022	20.0%	99.2%	0.3%	11.3%	11.3%	8.7%
(21)	NeurIPS 2022	22.5%	99.3%	0.2%	11.4%	12.5%	10.0%
(22)	NeurIPS 2022	25.0%	99.2%	0.3%	11.5%	13.8%	11.2%
(23)	CoRL 2022	0.0%	99.5%	0.2%	10.2%	1.5%	1.5%
(24)	CoRL 2022	2.5%	99.5%	0.2%	10.4%	3.3%	0.8%
(25)	CoRL 2022	5.0%	99.5%	0.2%	10.4%	5.0%	0.0%
(26)	CoRL 2022	7.5%	99.5%	0.3%	10.8%	6.8%	0.7%
(27)	CoRL 2022	10.0%	99.5%	0.3%	11.0%	8.4%	1.6%
(28)	CoRL 2022	12.5%	99.5%	0.3%	10.9%	10.2%	2.3%
(29)	CoRL 2022	15.0%	99.5%	0.3%	11.1%	11.8%	3.2%
(30)	CoRL 2022	17.5%	99.5%	0.3%	11.1%	13.8%	3.7%
(31)	CoRL 2022	20.0%	99.5%	0.3%	11.4%	15.5%	4.5%
(32)	CoRL 2022	22.5%	99.5%	0.2%	11.6%	17.4%	5.1%
(33)	CoRL 2022	25.0%	99.5%	0.3%	11.7%	18.9%	6.1%

Ours	RADAR(RoBERTa)	Deepfake(Longformer)	Fast-DetectGPT(Zero-shot)	BERT
6.809×10^{-8}	9.671	50.781	84.669	2.721

Amortized inference computation cost per 32-token sentence in GFLOPs

Results of Real Reviews

- Addresses main question of case study

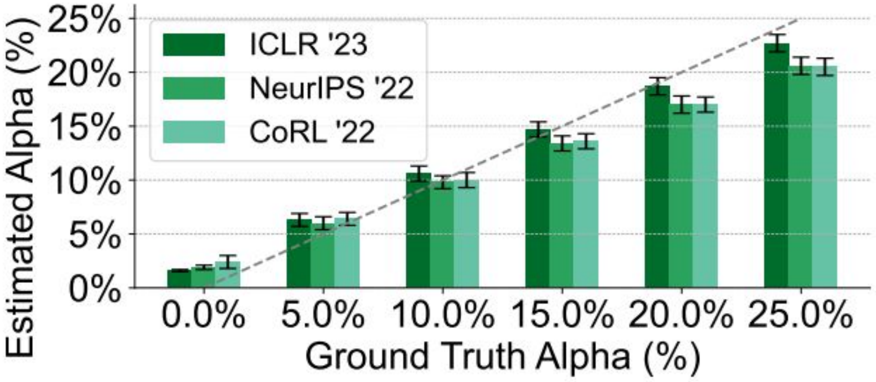


No.	Validation Data Source	Estimated α	Estimated CI (±)
(1)	NeurIPS 2019	1.7%	0.3%
(2)	NeurIPS 2020	1.4%	0.1%
(3)	NeurIPS 2021	1.6%	0.2%
(4)	NeurIPS 2022	1.9%	0.2%
(5)	NeurIPS 2023	9.1%	0.2%
(6)	ICLR 2023	1.6%	0.1%
(7)	ICLR 2024	10.6%	0.2%
(8)	CoRL 2021	2.4%	0.7%
(9)	CoRL 2022	2.4%	0.6%
(10)	CoRL 2023	6.5%	0.7%
(11)	EMNLP 2023	16.9%	0.5%

(1)	Nature portfolio 2019	0.8%	0.2%
(2)	Nature portfolio 2020	0.7%	0.2%
(3)	Nature portfolio 2021	1.1%	0.2%
(4)	Nature portfolio 2022	1.0%	0.3%
(5)	Nature portfolio 2023	1.6%	0.2%

Detection of Outputs based off Outline

- Uses the same two-stage approach as discussed in first paper:
  - Create review outline while reading paper
  - Feed outline into LLM
- Uses same distributional framework to predict α.

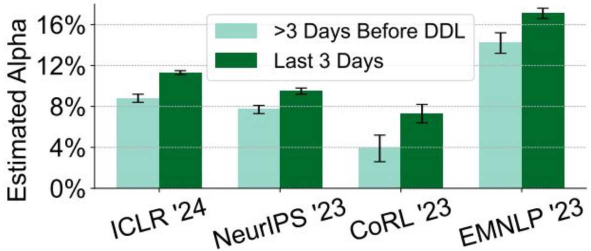


Skeleton generated LLM Reviews

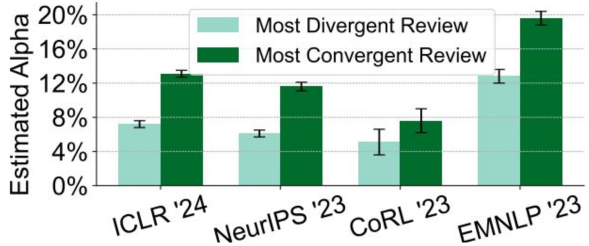
Relationships and Trends

1. Deadline Effect
2. Reference Effect
3. Lower Reply Rate Effect
4. Homogenization Effect
5. Low Confidence effect

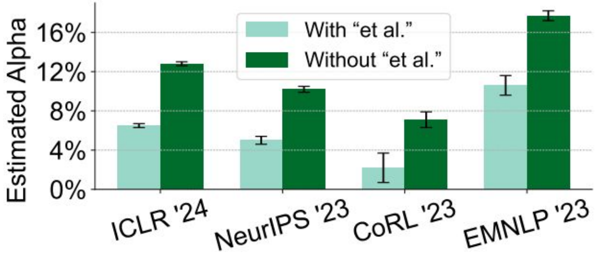
1

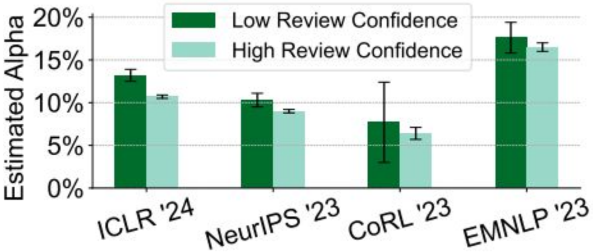


4

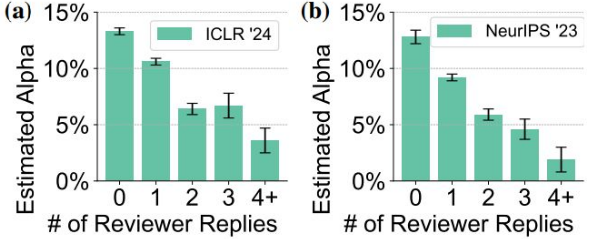


2





3



5

Key Limitations

- Potential Confounding Factors:
  - Variations in review guidelines, changes in reviewer demographics, etc.
- Synthetic Validation Limitations
  - Synthetic data may still not fully capture complexity of real world limitations

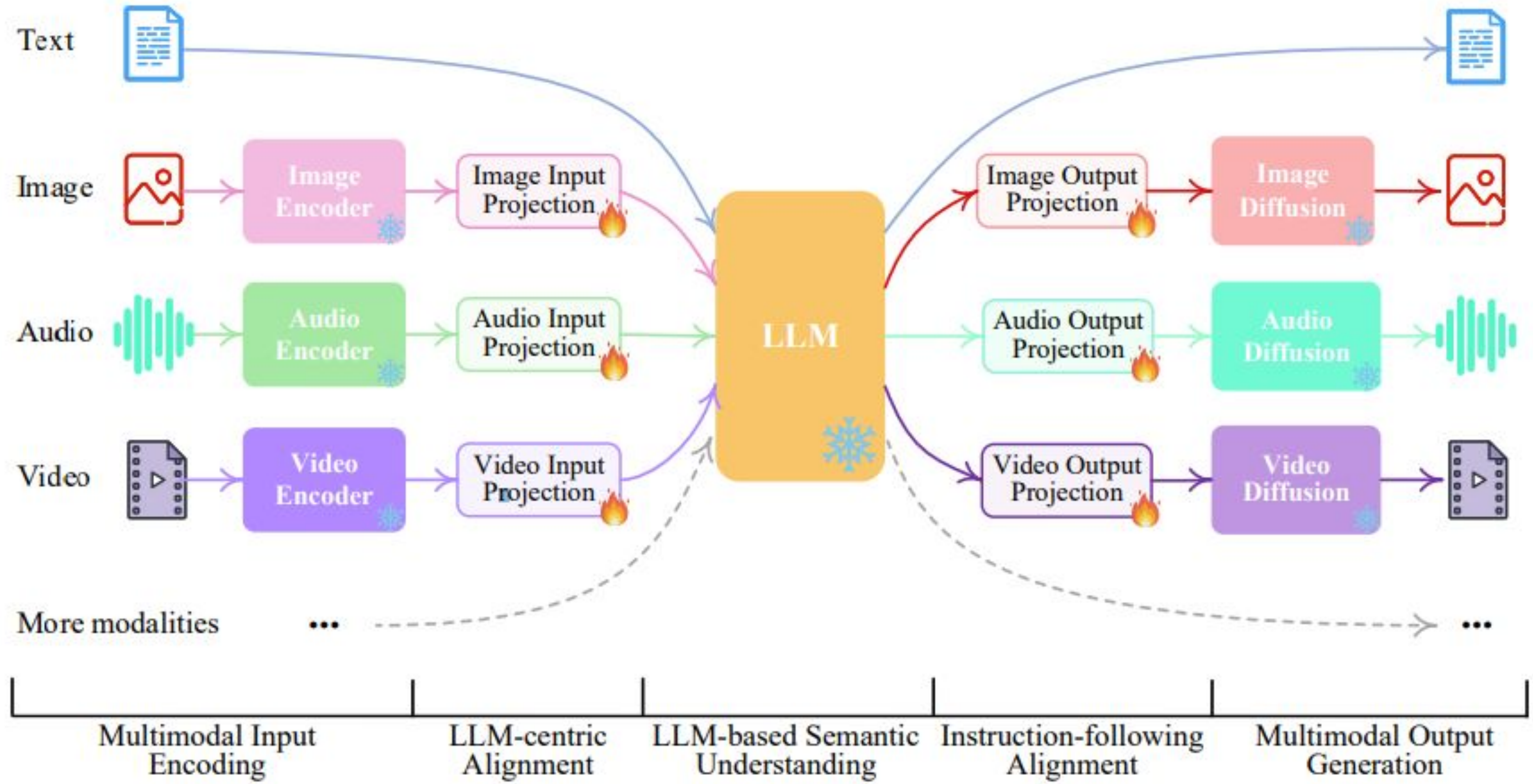
Agenda

- Mapping the Increasing Use of LLMs in Scientific Papers
- Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews
- LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts

Key Motivation

- Develop a general-purpose multimodal LLM that can accept inputs and deliver outputs in any combination of modalities
    - Overall goal is to mimic human-like any-to-any modalities

NExT-GPT: An Overall Look



NExT-GPT: An Overall Look

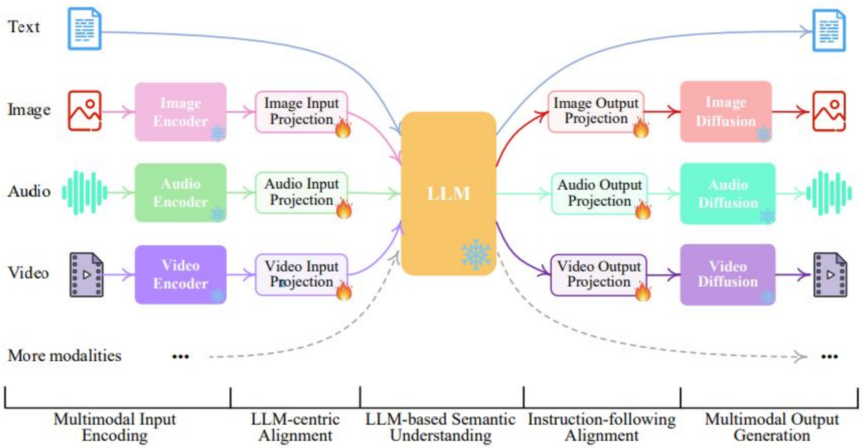
- Multimodal Encoding Stage
  - Leverages well-established models to encode inputs (ImageBind)
- LLM Understanding and Reasoning Stage
  - Outputs textual response directly and instructions to encoding layer (Vicuna 7B-v0)
- Generation Stage
  - Routes instruction signal from LLM stage and synthesizes content for output (Stable Diffusion, Zeroscope, AudioLDM)

Table 1. Summary of NExT-GPT system configuration. Only 1% of parameters need updating during fine-tuning.

	Encoder		Input Projection		LLM		Output Projection		Diffusion	
	Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—			—	—	—	—
Image					Vicuna	7B❄	Transformer	31M🔥	SD	1.3B❄
Audio	ImageBind	1.2B❄	Grouping	28M🔥	(LoRA	33M🔥)	Transformer	31M🔥	AudioLDM	975M❄
Video							Transformer	32M🔥	Zeroscope	1.8B❄

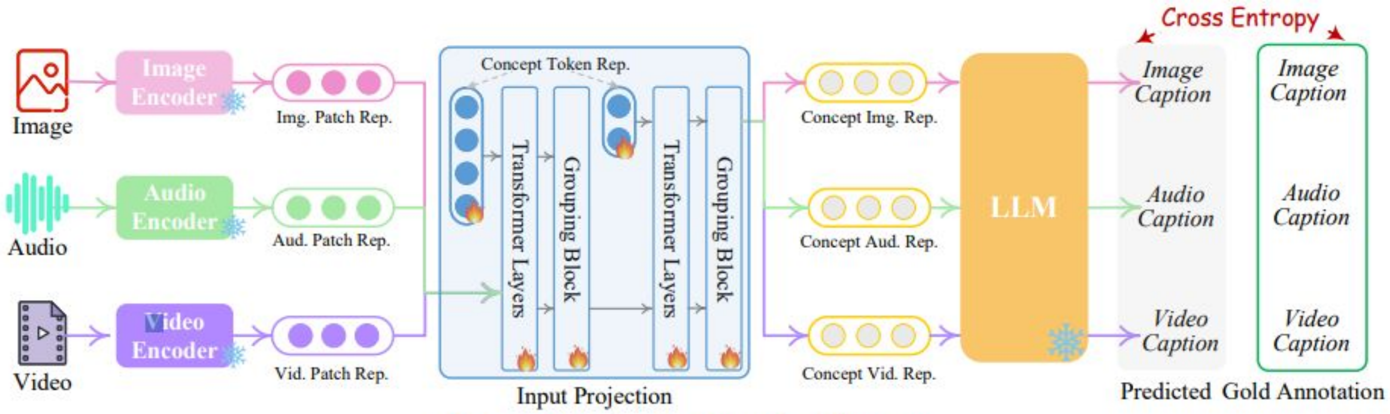
Multimodal Alignment Learning

- Each input has its own feature space
    - Must all be mapped into a single compatible space
    - Signals produced by the LLM must be accurately communicated to encoders
    - Must be able to do so quickly and efficiently



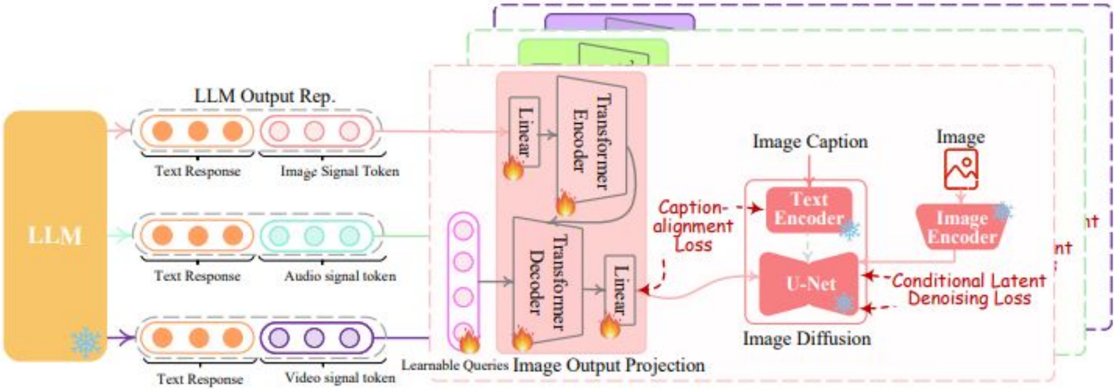
Encoding-side Alignment

- Different modalities are typically represented using patch-based or grid-based features.
- Introduce learnable concept tokens
  - Act as intermediary representations and grouped into concepts
  - Representations are fed into ‘frozen’ LLM to train conversions



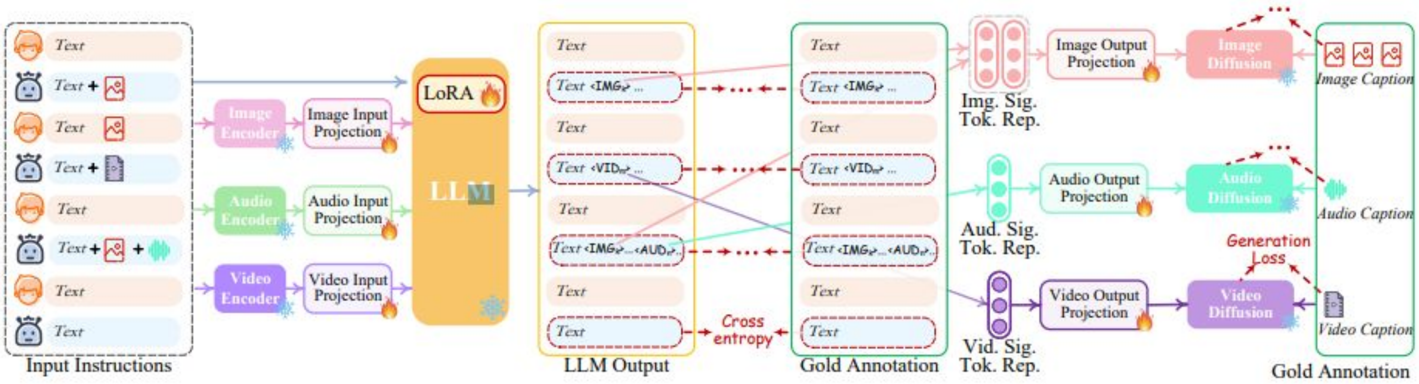
Decoding-side Alignment

- Intermediary LLM output produces signal (instruction) token.
- Utilizes a special set of tokens to guide output projection layers
  - Only projection layers are fine tuned.
  - Loss function penalizes the model for deviating from correct caption.



Experiment Setup

- Utilizes specially curated modality switching instruction tuning "Test+X" dataset (MosIT)
  - Constructed as an (INPUT, OUTPUT) pair,
  - Quantitative metrics for text outputs, and image/video quality used.



Results

Table 2. Zero-shot evaluation of image captioning with CIDEr (↑) score on NoCaps (Agrawal et al., 2019), Flickr 30K (Young et al., 2014) and COCO (Karpathy & Fei-Fei, 2017), and image question answering on VQA^{v2} (Goyal et al., 2017), VizWiz (Gurari et al., 2018) and OKVQA (Marino et al., 2019), and two evaluation-only benchmarks, MMB (Liu et al., 2023c) and SEED (Li et al., 2023a). The best results are marked in bold, and the second ones are underlined.

Model	Version	Image Captioning			Image Question Answering			Comprehensive	
		NoCaps	Flickr 30K	COCO	VQA^{v2}	VizWiz	OKVQA	MMB	SEED
InstructBLIP (Dai et al., 2023)	Vicuna-7B	<u>123.1</u>	82.4	102.2	-	33.4	33.9	36.0	-
LLaVA (Liu et al., 2023b)	LLaMA-2-7B-Chat	120.7	<u>82.7</u>	-	-	-	-	36.2	-
mPLUG-Owl (Ye et al., 2023b)	LLaMA-7B	117.0	80.3	<u>119.3</u>	-	39.0	-	46.6	<u>34.0</u>
Emu (Sun et al., 2023)	LLaMA-7B	-	-	117.7	40.0	35.4	34.7	-	-
DREAMLLM (Dong et al., 2023)	Vicuna-7B	-	-	115.4	56.6	45.8	44.3	49.9	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	-	-	**74.7**	<u>48.1</u>	-	**60.9**	-
NExT-GPT	Vicuna-7B	**123.7**	**84.5**	**124.9**	<u>66.7</u>	**48.4**	**52.1**	<u>58.0</u>	**57.5**

Table 3. Comparison of video reasoning tasks on MSRVTT (Xu et al., 2016), MSVD-QA and MSRVTT-QA (Xu et al., 2017) and NExTQA (Xiao et al., 2021), and the audio captioning task on AudioCaps (Kim et al., 2019). Scores with ∗ means being fine-tuned on the training dataset.

Model	Version	Video Captioning	Video Question Answering			Audio Captioning
		MSR-VTT	MSVD-QA	MSRVTT-QA	NExTQA	AudioCaps
Codi (Tang et al., 2023)	-	<u>74.4</u>*	-	-	-	<u>78.9</u>*
UIO-2XXL (Lu et al., 2023)	6.8B	48.8*	41.5	52.1	-	48.9*
Video-LLaMA (Zhang et al., 2023c)	LLaMA-7B	-	51.6	-	<u>29.6</u>	-
Video-LLaVA (Lin et al., 2023)	Vicuna-7B	-	**70.7**	<u>59.2</u>	-	-
Emu (Sun et al., 2023)	LLaMA-7B	-	32.4	14.0	6.8	-
NExT-GPT	Vicuna-7B	**76.2***	<u>64.5</u>	**61.4**	**50.7**	**81.3***

Impact of Signal Tokens and Grouping Mechanisms

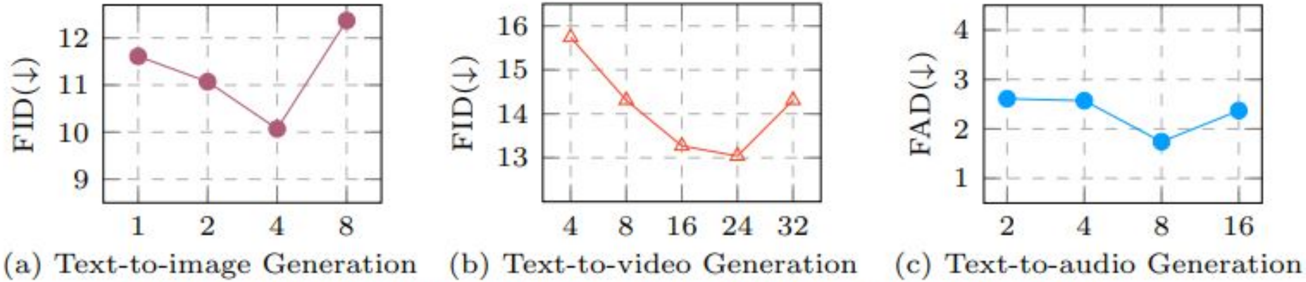


Figure 6. The generation quality under different numbers of modality signal tokens.

Table 4. Results on text-to-image/audio/video generation (MS COCO (Lin et al., 2014), AudioCaps (Kim et al., 2019), and MSRVTT (Xu et al., 2016)). †: zero-shot results.

Model	Image FID (↓)	Audio FAD (↓)	Video CLIPSIM (↑)
SD-1.5 (Wang et al., 2022c)	11.21	-	-
Codi (Huang et al., 2023a)	11.26	1.80	28.90
AudioLDM-L (Liu et al., 2023a)	-	1.96	-
GILL-8B† (Koh et al., 2023)	12.20	-	-
Emu-13B† (Sun et al., 2023)	11.66	-	-
UIO-2XXL (Lu et al., 2023)	13.39	2.64	-
NExT-GPT	**10.07**	**1.68**	**31.97**
NExT-GPT†	11.18	1.74	30.96

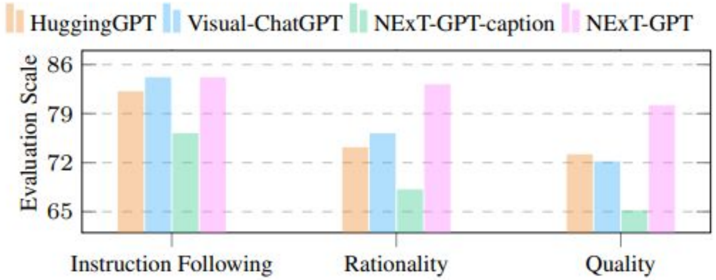


Figure 5. Human Evaluation (1-100 scale, results are on average) of NExT-GPT in comparison with pipeline baselines.

Figure 4. Qualitative examples showcasing the interpretative and generative capabilities of NExT-GPT across diverse modalities or their combinations.

Limitations

- GPT is constrained by the quantity of fine-tuned data as well as the quality of the off the shelf models used.
    - The possibility of low quality responses and hallucinations still imminent.

Conclusion

- In this presentation, we quantified the usage Large Language Models in the academic sphere at a macro level.
  - Quantified the usage in academic papers and peer reviews in various fields.
  - Analyzed the relationship of various circumstances and the rate of usage.
- We studied the development of a human-level AI Agent
  - Studied the end-to-end architecture of a MM-LLM and its utilization of mainstream encoders and decoders.
  - Further examined an efficient fine tuning processes to ensure proper alignment of transformers.

Questions?