
A
A > B
OK
B
B > A
© dall-e

Why Does ChatGPT "Delve" So Much?

Exploring the Sources of Lexical Overrepresentation in Large Language Models

21 Jan 2025 @ COLING25

T.S. Juzek & Z. B. Ward

Joint work with Zina Ward

04
OCTOBER
12:00 - 13:00
DSC-499

ZINA WARD

WHY 'DELVE'?
ON LEXICAL OVER-REPRESENTATION IN LARGE LANGUAGE MODELS

ABSTRACT: Large language models like ChatGPT frequently use words like "delve" and "intricate." Our research poses "the puzzle of lexical overrrepresentation": *why* are certain words overused by ChatGPT? This talk will explore several potential explanations.

More Info:
www.sc-ai.net

Throughout the project great input by Gordon Erlebacher

Links

Paper: https://arxiv.org/pdf/2412.11385

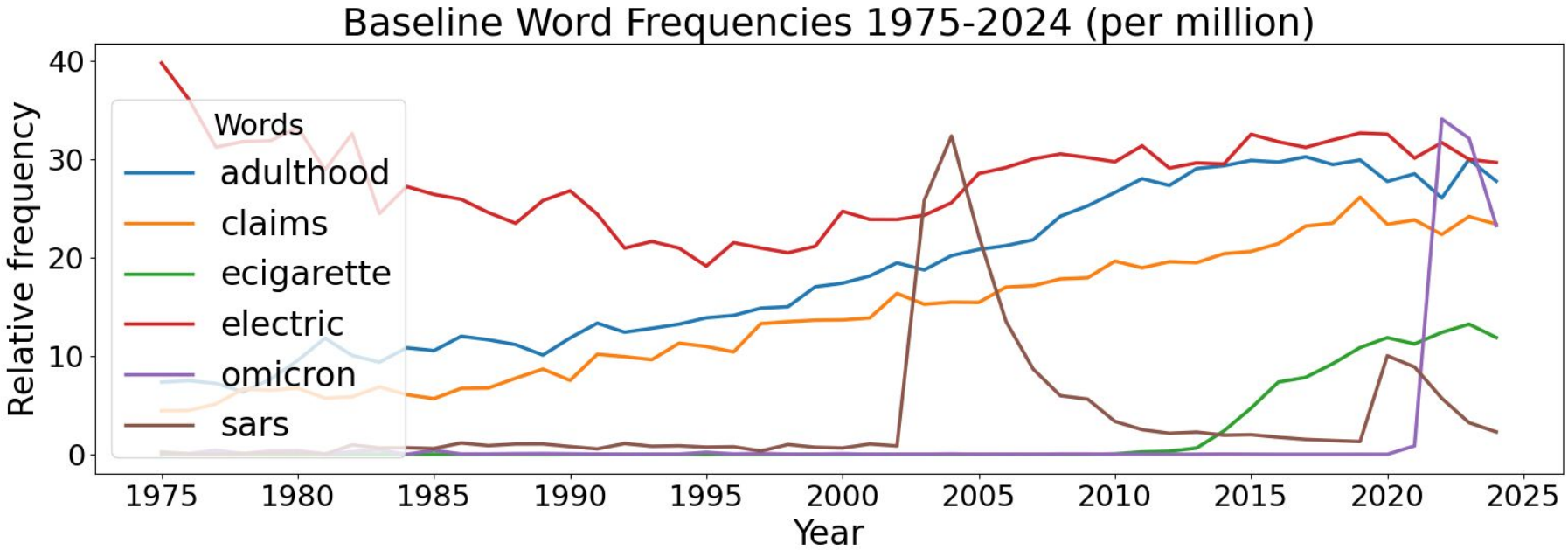
Repo: https://github.com/tjuzek/delve

Background

Language changes over time

Scientific English changes over time (→ Elke Teich's Team at Saarland University)

Examples:


Baseline Word Frequencies 1975-2024 (per million)
Relative frequency
Year
Words
adulthood
claims
ecigarette
electric
omicron
sars
0
10
20
30
40
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025

Background

There have been rapid changes recently

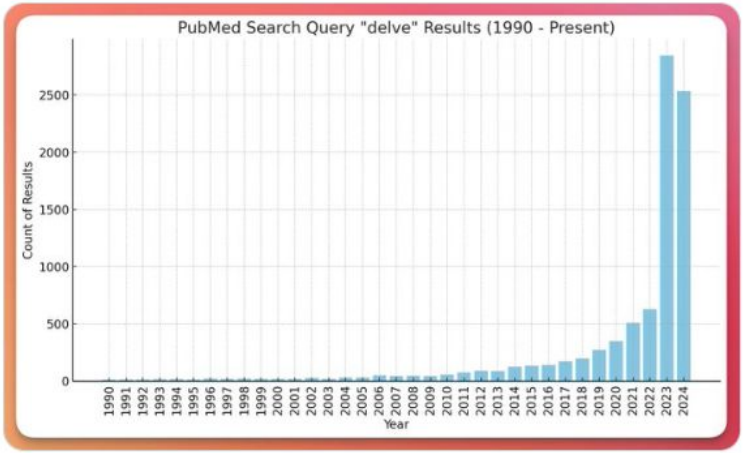
These changes are hard to explain ‘naturally’

Jeremy Nguyen ✍️ 🏯
@JeremyNguyenPhD

Are medical studies being written with ChatGPT?

Well, we all know ChatGPT overuses the word "delve".

Look below at how often the word 'delve' is used in papers on PubMed (2023 was the first full year of ChatGPT).



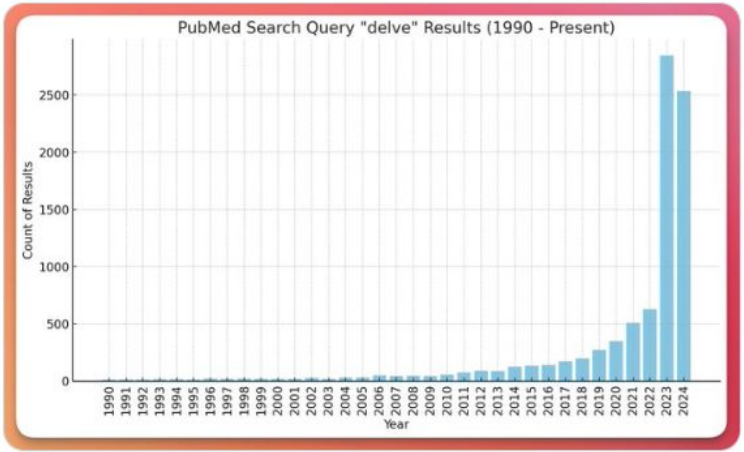
6:31 AM · Mar 30, 2024 · **2.3M** Views

Jeremy Nguyen ✍️ 🏯
@JeremyNguyenPhD

Are medical studies being written with ChatGPT?

Well, we all know ChatGPT overuses the word "delve".

Look below at how often the word 'delve' is used in papers on PubMed (2023 was the first full year of ChatGPT).



6:31 AM · Mar 30, 2024 · **2.3M** Views

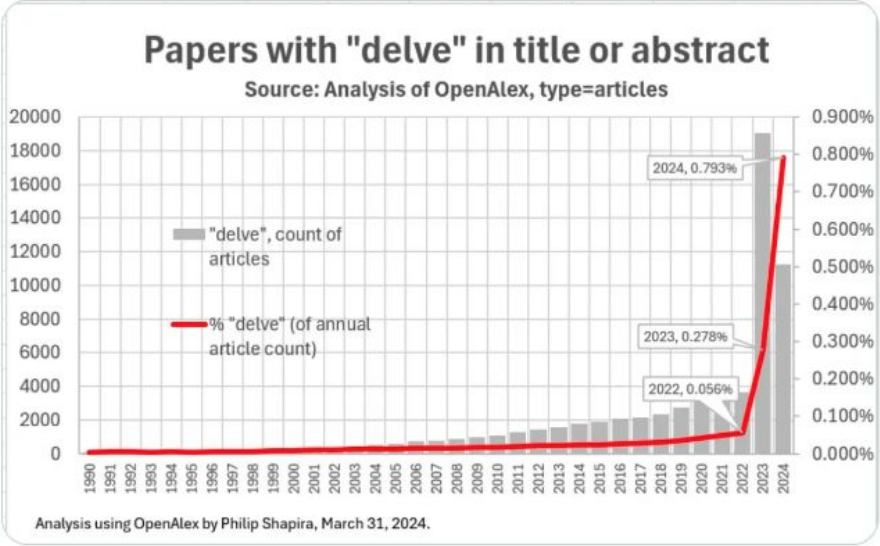
Paul Graham @paulg · Apr 7

Someone sent me a cold email proposing a novel project. Then I noticed it used the word "delve."

2.4K 7.3K 9.5K 17M

Paul Graham @paulg · Apr 7

My point here is not that I dislike "delve," though I do, but that it's a sign that text was written by ChatGPT.



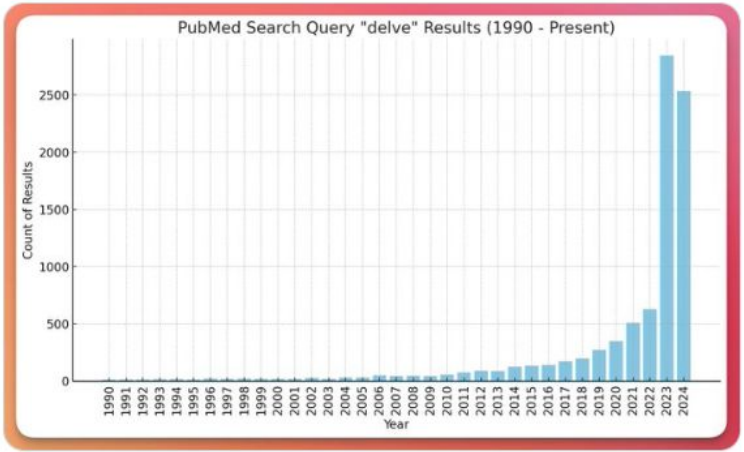
731 1.7K 5.7K 2.3M

Jeremy Nguyen ✍️ 🍨
@JeremyNguyenPhD

Are medical studies being written with ChatGPT?

Well, we all know ChatGPT overuses the word "delve".

Look below at how often the word 'delve' is used in papers on PubMed (2023 was the first full year of ChatGPT).



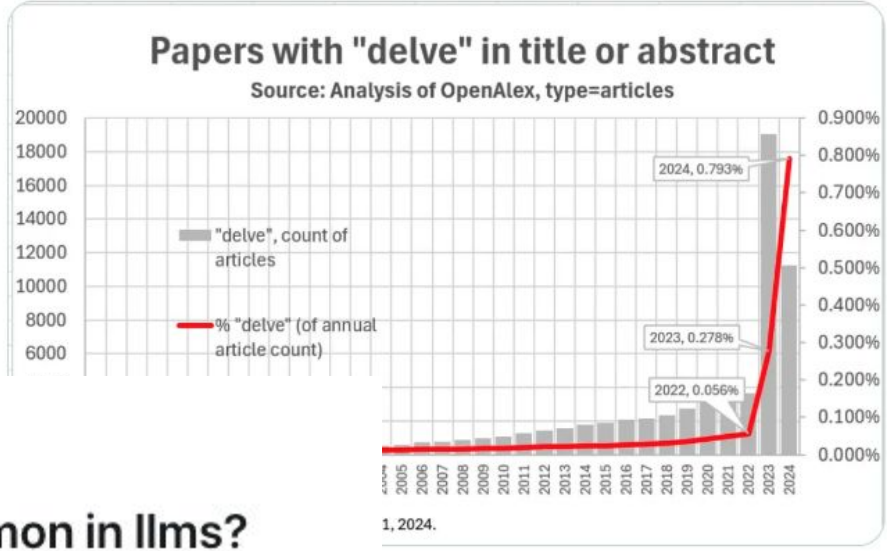
Paul Graham @paulg · Apr 7

Someone sent me a cold email proposing a novel project. Then I noticed it used the word "delve."

2.4K 7.3K 9.5K 17M

Paul Graham @paulg · Apr 7

My point here is not that I dislike "delve," though I do, but that it's a sign that text was written by ChatGPT.



5.7K 2.3M

r/OpenAI · 10 mo. ago
[deleted]

The word "Delve" - why is it so common in llms?

Question

Background

- *That* this is happening, is well established

Koppenburg, 2024; Nguyen, 2024; Shapira, 2024; Gray, 2024; Kobak et al., 2024; Liang et al., 2024; Liu and Bu, 2024; Matsui, 2024; Juzek and Ward 2025

Background

And early on, these changes were attributed to the influence of Large Language Models (LLMs) like ChatGPT

Background

However:

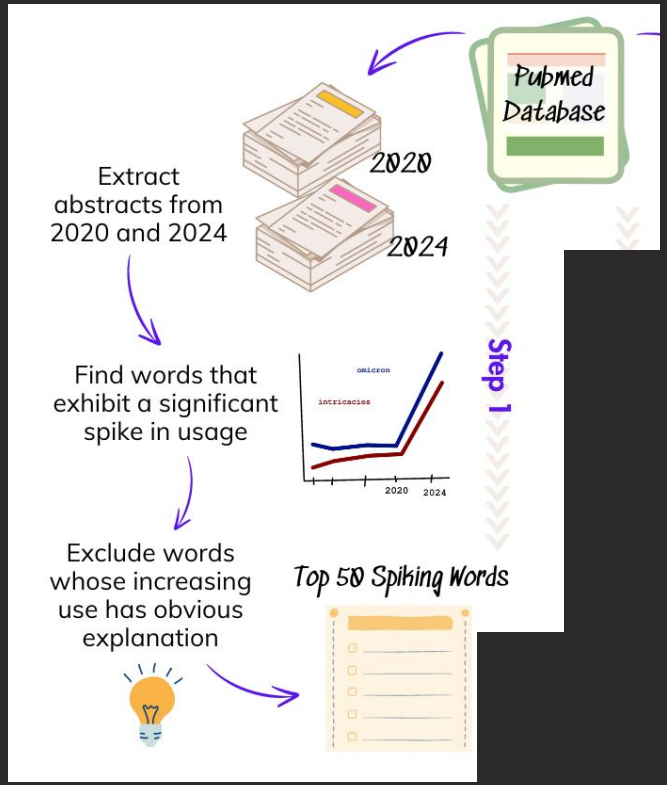
- Handpicked items
- Strengthen the link to LLMs
- Critically: not clear WHY LLMs do this
  - Informed speculation
    → RLHF: Hern, 2024; Sheikh, 2024

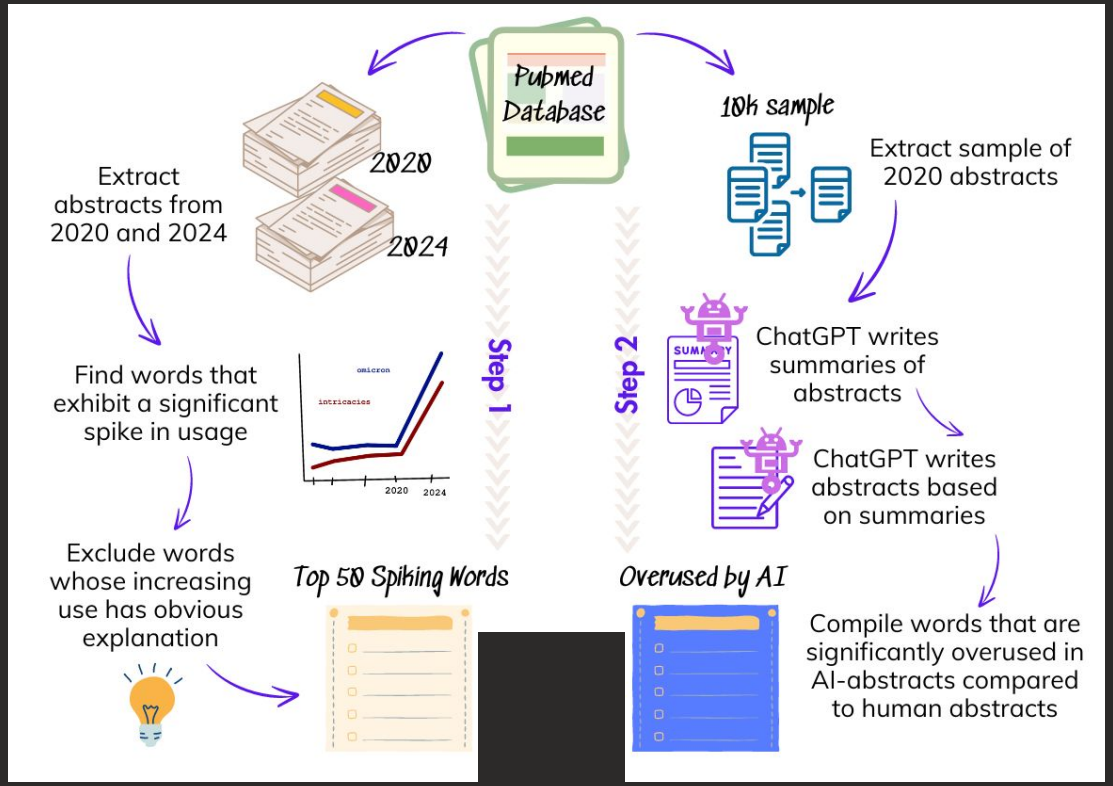
Background

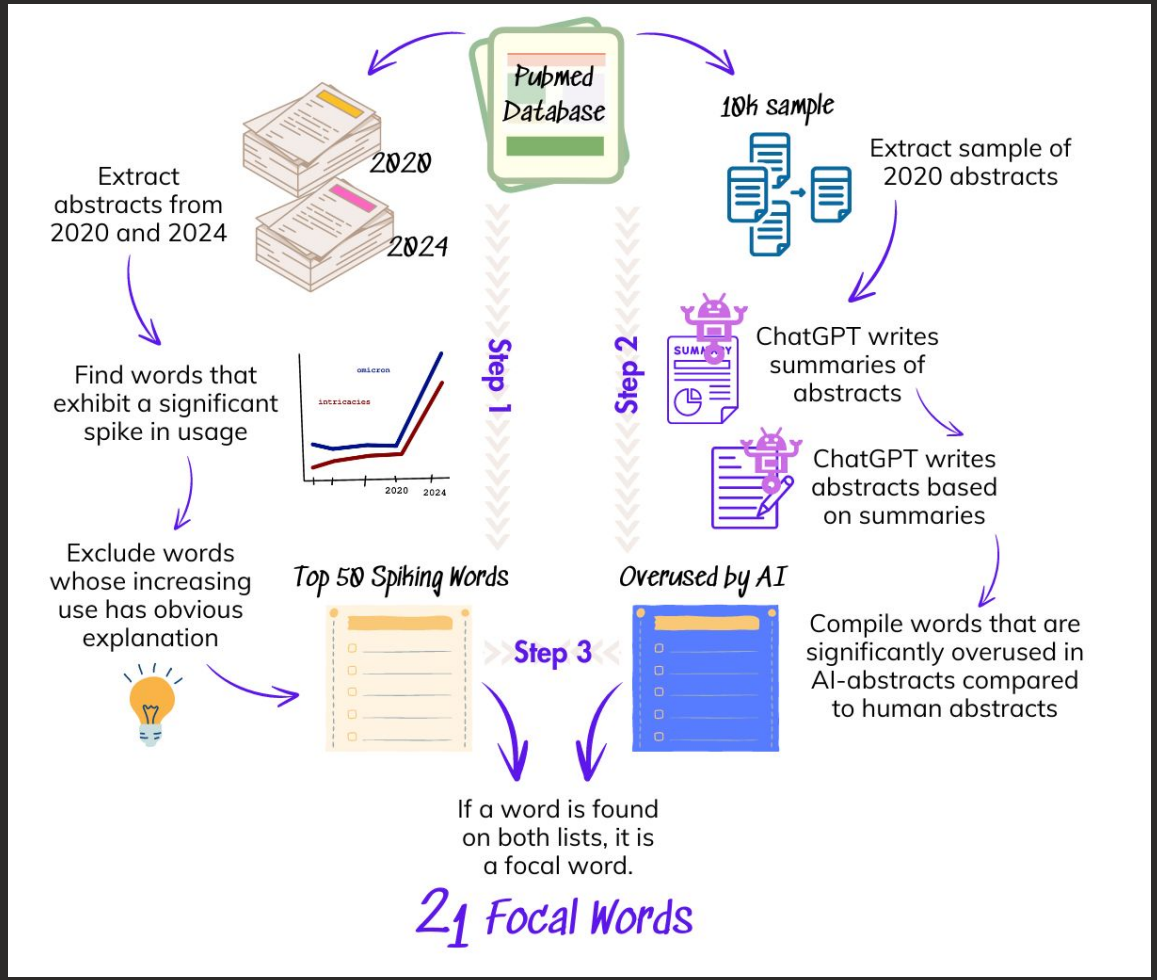
Our work:

- Procedure to systematically identify overused items
- *Why* are LLMs overusing certain words

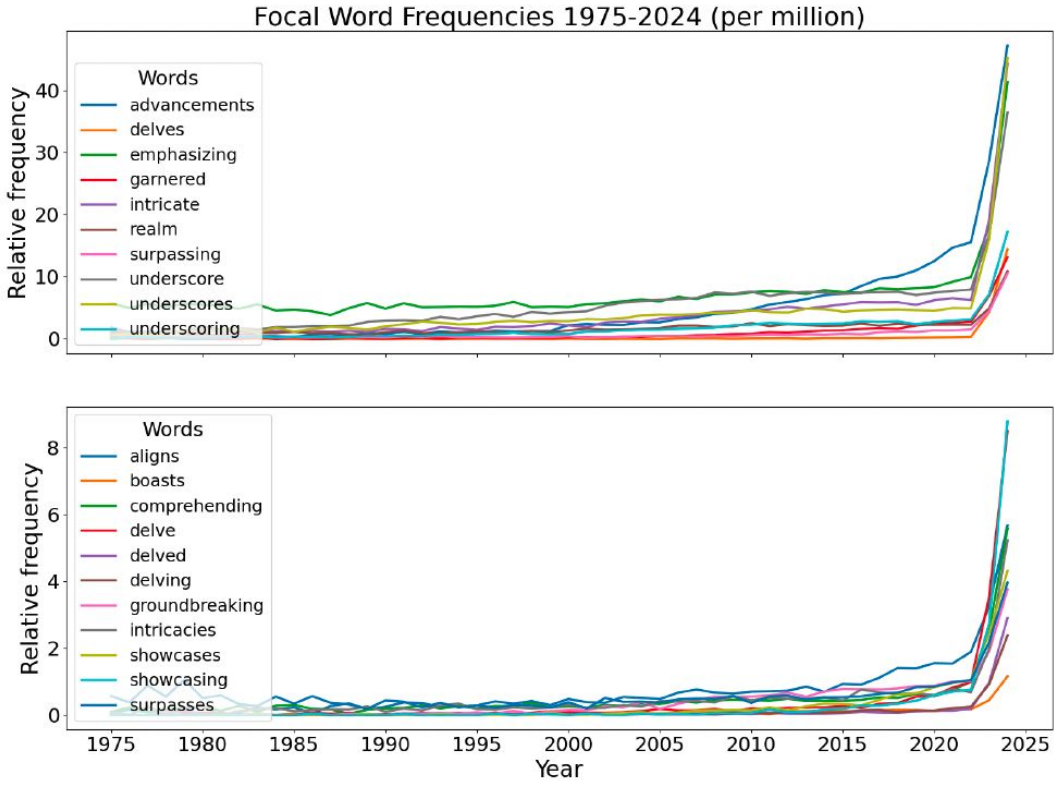
The procedure:


Pubmed Database
2020
2024
Extract abstracts from 2020 and 2024
Step 1
Find words that exhibit a significant spike in usage
omicron
intricacies
2020
2024
Exclude words whose increasing use has obvious explanation
Top 50 Spiking Words


Pubmed Database
Step 1
Extract abstracts from 2020 and 2024
2020
2024
Find words that exhibit a significant spike in usage
omicron
intricacies
2020
2024
Exclude words whose increasing use has obvious explanation
Top 50 Spiking Words
Step 2
10k sample
Extract sample of 2020 abstracts
ChatGPT writes summaries of abstracts
ChatGPT writes abstracts based on summaries
Compile words that are significantly overused in AI-abstracts compared to human abstracts
Overused by AI


Pubmed Database
10k sample
Extract sample of 2020 abstracts
Extract abstracts from 2020 and 2024
2020
2024
Step 1
Step 2
ChatGPT writes summaries of abstracts
Find words that exhibit a significant spike in usage
omicron
intricacies
2020
2024
ChatGPT writes abstracts based on summaries
Exclude words whose increasing use has obvious explanation
Top 50 Spiking Words
Overused by AI
Step 3
Compile words that are significantly overused in AI-abstracts compared to human abstracts
If a word is found on both lists, it is a focal word.
21 Focal Words

Word	opm 2020	opm 2024	Incr. %
delves	0.21	14.38	6697.14
delved	0.12	2.90	2240.47
delving	0.12	2.38	1816.83
showcasing	0.59	8.79	1396.03
delve	0.58	8.50	1374.92
boasts	0.11	1.15	918.18
underscores	4.50	45.19	903.61
comprehending	0.56	5.58	898.95
intricacies	0.60	5.22	772.85
surpassing	1.37	10.50	667.48
intricate	6.22	44.22	611.24
underscoring	2.70	17.17	536.94
garnered	2.44	13.13	437.19
showcases	0.82	4.31	422.45
emphasizing	8.30	41.27	397.12
underscore	7.42	36.40	390.65
realm	2.25	10.85	381.10
surpasses	0.85	3.96	367.55
groundbreaking	0.87	3.75	330.42
advancements	12.49	47.17	277.59
aligns	1.55	5.68	266.97



List of factors

- Initial training data
- Fine-tuning
- Model architecture
- Choice of algorithms
- Context priming
- Learning from Human Feedback
- Other factors (parameter settings, etc.)

List of factors

- Initial training data
- Fine-tuning
- Model architecture
- Choice of algorithms
- Context priming
- Learning from Human Feedback
- Other factors (parameter settings, etc.)

possible, but no strong starting points

List of factors

- Initial training data
- Fine-tuning
- Model architecture
- Choice of algorithms
- Context priming
- Learning from Human Feedback
- Other factors (parameter settings, etc.)

List of factors

- Initial training data
- Fine-tuning
- Model architecture
- Choice of algorithms
- Context priming
- Learning from Human Feedback
- Other factors (parameter settings, etc.)

language output by Llama Base (-LHF) vs Llama Instruct (+LHF) → indicator

List of factors

language output by

	Llama 2-Base	Llama 2-Chat
Human	1.616	1.051
AI	1.633	0.886

Table 1: Per-word entropy for human abstracts compared to ChatGPT-generated abstracts. Higher values of entropy mean that the model is more "surprised."

N.B.: RLHF, DPO, and LHF

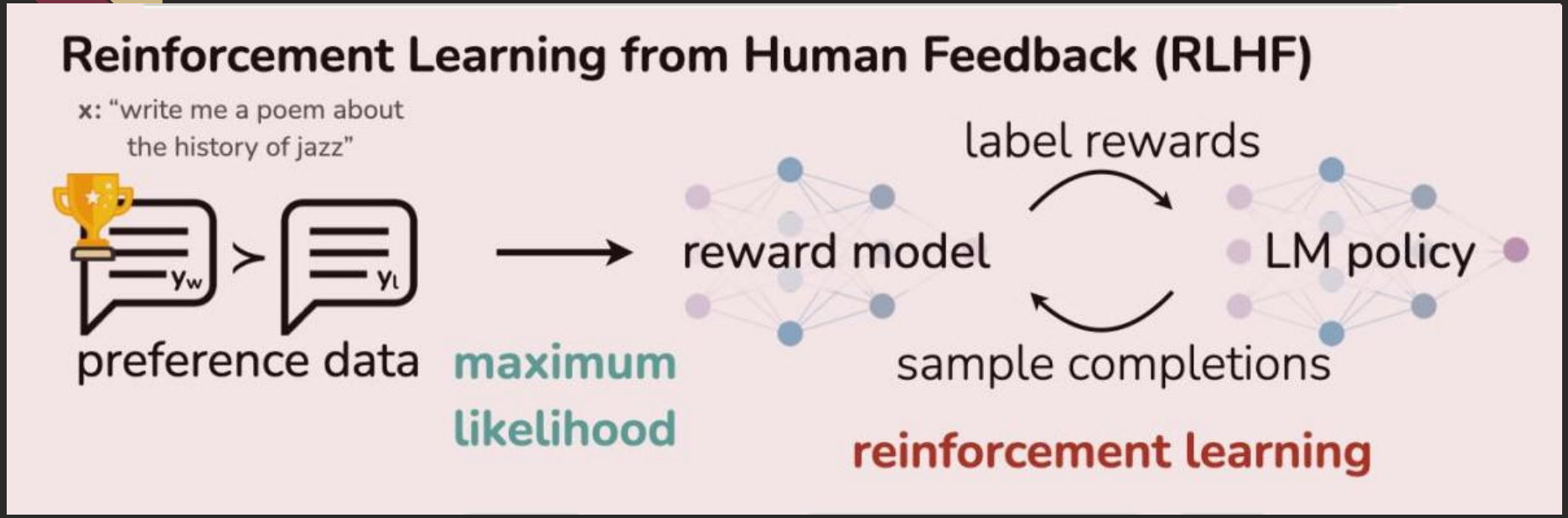
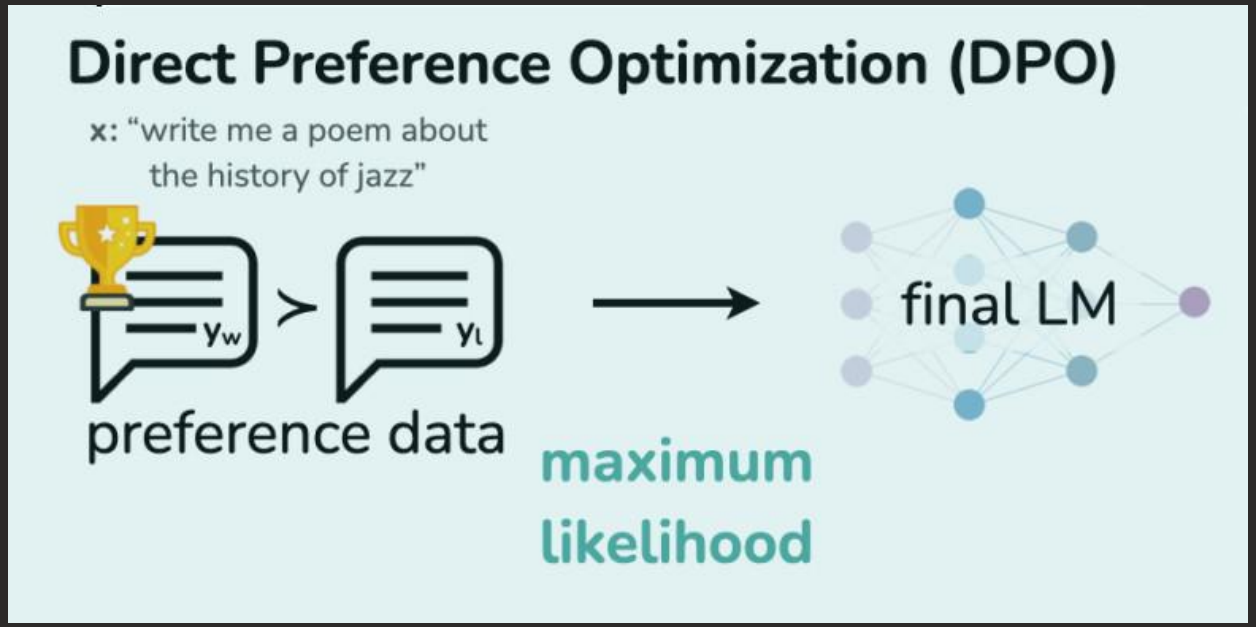


illustration from Rafailov et al. 2024

Direct Preference Learning



Rafailov et al. 2024

Roof term

Learning from Human Feedback (LHF)

Because of Direct Preference Learning → Llama 3

LHF

- LHF is a very plausible candidate,
- Others have pointed to it
- Experimental validation is needed

LHF

- typically:
  - done in global south
  - precarious conditions
  - Toxtli et al., 2021; Roberts, 2022; Novick, 2023
- lack of transparency

Experiment: Emulate LHF

- IRB
- LAMP stack for rating website
- Decision log
  - virtually everything pre-planned
  - pre-designed: coefficient, 2.5 vs 2
  - “preliminary” results
- Recruitment Prolific
- Emulate procedure: Demographics
  - Global South
- Highest standards, incl. NCP
- Random item order, random item position, etc. etc.

A novel approach has been devised for blocking c-di-GMP signaling pathways, a crucial mechanism in bacterial cell functioning. The technique employs a c-di-GMP-sequestering peptide (CSP) that exhibits strong affinity for c-di-GMP and effectively inhibits its signaling. Through targeted mutations, a potent, shortened variant of CSP has been developed, demonstrating efficient inhibition of biofilm formation in Pseudomonas aeruginosa. This innovative method provides a highly promising strategy for targeting c-di-GMP and holds potential for combating various bacterial infections. Further studies could focus on developing more potent and specific CSP variants to fully comprehend and utilize the role of c-di-GMP in regulating bacterial functions.

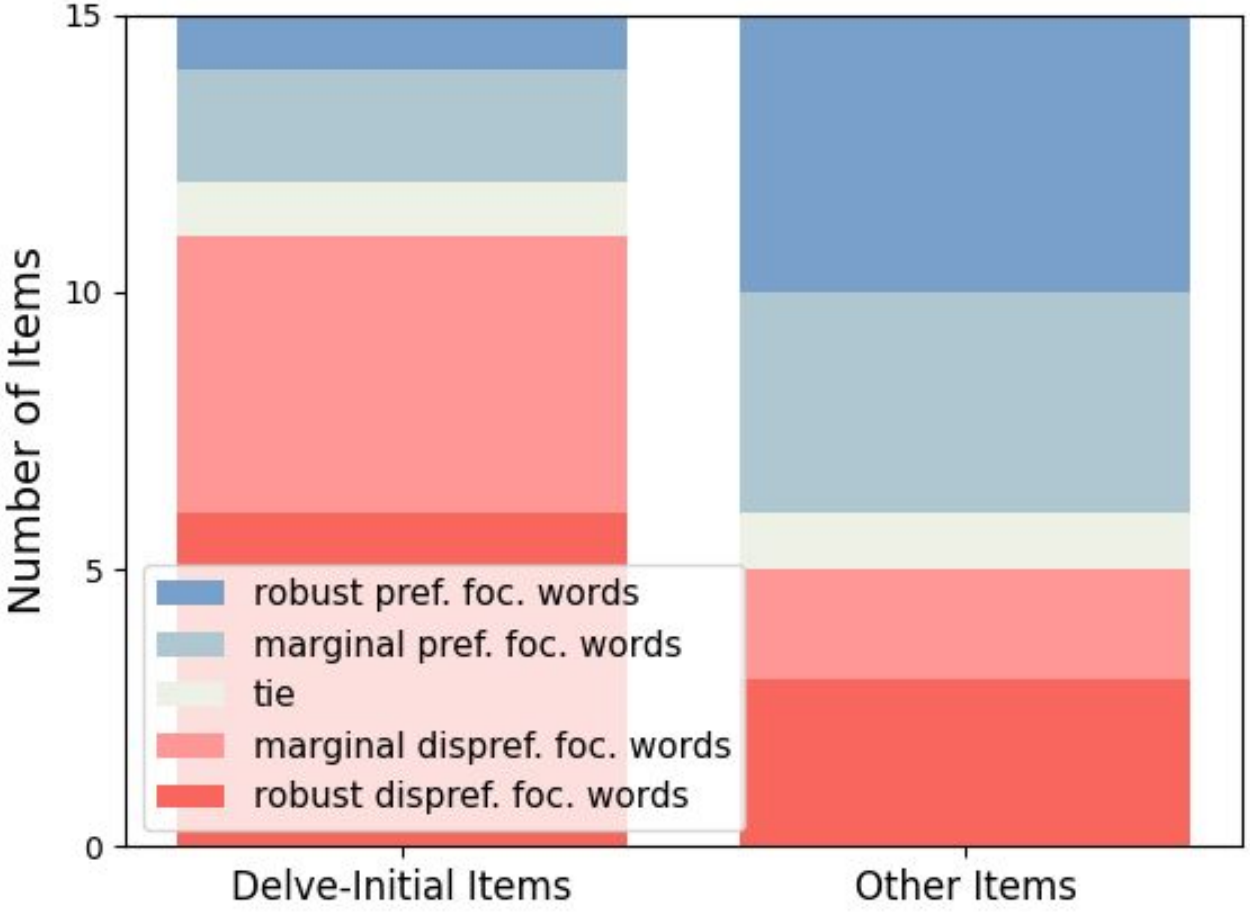
This paper showcases a novel approach for targeting and disrupting c-di-GMP signaling pathways in bacteria. By utilizing a c-di-GMP-sequestering peptide (CSP), the researchers have developed a method to bind and inhibit c-di-GMP, a key bacterial second messenger. Through structure-based mutations, a more powerful and compact variant of the CSP has been created, effectively preventing biofilm formation in Pseudomonas aeruginosa. This advancement holds promise for controlling bacterial behaviors mediated by c-di-GMP and could have implications for the development of new antibacterial strategies. The results of this study highlight the potential of CSP as a tool for delving into the intricate mechanisms of c-di-GMP signaling.

left is better

right is better

Analysis

- chi-square
- explorative multifactorial regression
- $\longrightarrow$ paper


15
10
5
0
Number of Items
robust pref. foc. words
marginal pref. foc. words
tie
marginal dispref. foc. words
robust dispref. foc. words
Delve-Initial Items
Other Items

Results

- Exclusion rate
- “Delve” pushback
  - We will come back to this

→virtually no chance to get conclusive experimental results

→conjecture → follow-up

Limitations

- Issues with experiment
- For ethical reasons: cannot truly emulate procedure
- Need to explore other factors

Intellectual merit

- procedure to identify overused words, “focal words”
- factors contributing to lexical bias
  - stronger, but not fully conclusive evidence → Learning from Human Feedback

Broader impacts

- Technology is strongly affecting language usage
- It was not clear: What do we make of the recent changes
- What do we make of the causes

Broader impacts

- The big unknown:
  - Variety
  - vs
  - Demographics: Age
- *It could be just 'normal' language change!*
- Or just the Task!
- →follow up

Broader impacts

- Critically:
  - *Insights can be gained*
    - It is tough, though, partly because:
  - Lack of procedure and data transparency slow down progress

Thank you