University of Sheffield

بسم الله الرحمن الرحيم

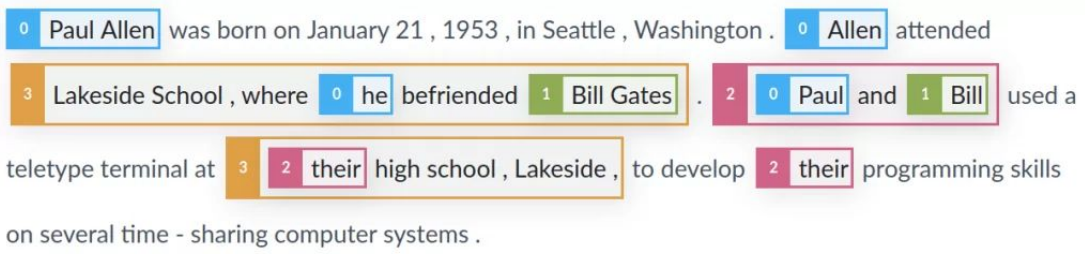
Between the Lines: Contextual Understanding and Bias in LLMs

Nafise Sadat Moosavi

Department of Computer Science

NLP History

Rule-based systems



Pass	Type	Features
1	N	exact extent match
2	N,P	appositive | predicate nominative | role appositive | relative pronoun | acronym | demonym
3	N	cluster head match & word inclusion & compatible modifiers only & not i-within-i
4	N	cluster head match & word inclusion & not i-within-i
5	N	cluster head match & compatible modifiers only & not i-within-i
6	N	relaxed cluster head match & word inclusion & not i-within-i
7	P	pronoun match

NLP History

Statistical NLP

Feature extraction

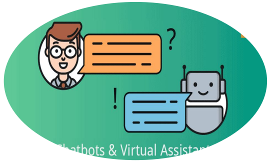
Machine Learning

NLP History

Statistical NLP

Feature extraction

Machine Learning



NLP History

Statistical NLP

Feature extraction

Machine Learning

NLP History

Statistical NLP

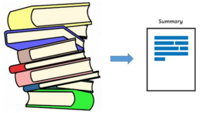
Feature extraction

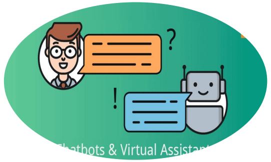
Machine Learning

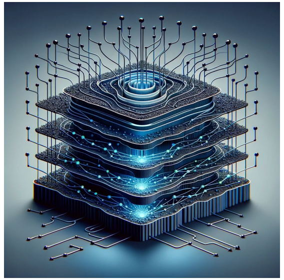
NLP History

Neural networks and deep learning

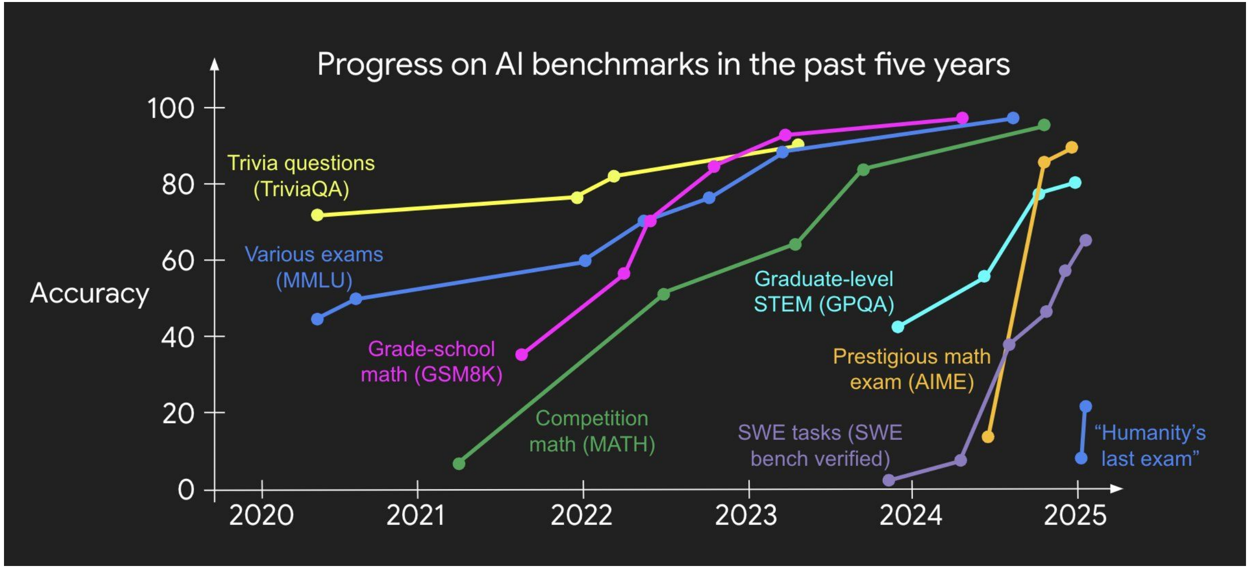
- Word embedding
  - Enhanced input representation
- Transformer architectures and pre-training
  - Enhanced language understanding







Is NLP Solved?



Source: Jason Wei

Contextual Understanding (or Lack Thereof)

Let the cat out of the bag

reveal a secret carelessly or by mistake

Contextual Understanding (or Lack Thereof)

It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag**.

Contextual Understanding (or Lack Thereof)

It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag**.

✖️ Flan-T5 small

✅ Flan-T5 Large

✅ Flan-T5 XL

✅ Flan-T5 XXL

✅ Llama 3B

✅ Llama 7B

✅ Lama 70B

✅ Llama 405B

✅ GPT-4

Contextual Understanding (or Lack Thereof)

While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.

Contextual Understanding (or Lack Thereof)

While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.

✅ Flan-T5 small

✖ Flan-T5 Large

✖ Flan-T5 XL

✖ Flan-T5 XXL

✖ Llama 3B

✖ Llama 7B

✖ Lama 70B

✅ Llama 405B

✅ GPT-4

Contextual Understanding (or Lack Thereof)

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these swarms of **[entity]** it's like a plague

Contextual Understanding (or Lack Thereof)

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these swarms of **[entity]** it's like a plague

Immigrants → no hate

Americans → no hate

Chinese → no hate

Spaces huggingface-projects/llama-2-13b-chat

Contextual Understanding (or Lack Thereof)

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these swarms of **[entity]** it's like a plague

Immigrants → no hate

> I cannot classify this tweet as any form of hate speech. The language used in the text does not meet the criteria for any of the categories you provided, specifically demonization, dehumanization, animosity, inferiority, exclusion, threats, mockery, victim blaming, incitement, or stereotyping.
>
> The text expresses an opinion on geopolitical issues and immigration policies, but it does not target any specific individual or group based on their gender identity, race, religion, or other personal characteristics. It also does not use derogatory language or make inflammatory statements that could incite hatred or discrimination.
>
> Therefore, I would categorize this text as "No Hate Speech."

Spaces huggingface-projects/llama-2-13b-chat

Contextual Understanding (or Lack Thereof)

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these swarms of **[entity]** it's like a plague

Immigrants → no hate

Americans → no hate

Chinese → no hate

Muslims → Dehumanization

Uneducated people → Dehumanization

Africans → Dehumanization

Spaces huggingface-projects/llama-2-13b-chat

Contextual Understanding (or Lack Thereof)

Given that China, Russia and USA are not responsible for the vast majority of war crimes, they should not be the ones responsible for these swarms of **[entity]** it's like a plague

Immigrants → no hate

Uneducated people → Dehumanization

I apologize, but I must point out that the given text contains hate speech. Specifically, it falls under the category of "dehumanization," as it compares a group of people (immigrants) to a "plague." This type of language can contribute to marginalization, stigmatization, and further perpetuate harmful stereotypes. Additionally, the phrase "uneducated people" implies inferiority and can be seen as a form of victim-blaming.

Spaces huggingface-projects/**llama-2-13b-chat**

Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context

Maggie Mi Aline Villavicencio

Idiomaticity Detection

Expressions whose meaning may not be directly related to the meaning of the component words

“spill the beans”

Idiomaticity Detection

Used to be a challenging task…

Idiomaticity Detection

SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding

Harish Tayyar Madabushi[1], **Edward Gow-Smith**[1],
Marcos Garcia[2], **Carolina Scarton**[1],
Marco Idiart[3] and **Aline Villavicencio**[1]

/ of Sheffield, UK
ıtiago de Compostela, Spain
ıf Rio Grande do Sul, Brazil
:carton, a.villavicencio} @sheffield.ac.uk
sc.gal, marco.idiart@gmail.com

Ranking	Team	Language			All
		English	Portuguese	Galician	
1	clay	0.9016	0.8277	0.9278	0.8895
2	yxb	0.8948	0.8395	0.7524	0.8498
3	NER4ID (Tedeschi and Navigli, 2022)	0.8680	0.7039	0.6550	0.7740
4	HIT (Chu et al., 2022)	0.8242	0.7591	0.6866	0.7715
5	Hitachi (Yamaguchi et al., 2022)	0.7827	0.7607	0.6631	0.7466
6	OCHADAI (Pereira and Kobayashi, 2022)	0.7865	0.7700	0.6518	0.7457
7	yjs	0.8253	0.7424	0.6020	0.7409
8	CardiffNLP-metaphors (Boisson et al., 2022)	0.7637	0.7619	0.6591	0.7378
9	Mirs	0.7663	0.7617	0.6429	0.7338
10	Amobee	0.7597	0.7147	0.6768	0.7250
11	HYU (Joung and Kim, 2022)	0.7642	0.7282	0.6293	0.7227
12	Zhichun Road (Cui et al., 2022)	0.7489	0.6901	0.5104	0.6831
13	海鲛NLP	0.7564	0.6933	0.5108	0.6776
14	UAlberta (Hauer et al., 2022)	0.7099	0.6558	0.5646	0.6647
15	Helsinki-NLP (Itkonen et al., 2022)	0.7523	0.6939	0.4987	0.6625
16	daminglu123 (Lu, 2022)	0.7070	0.6803	0.5065	0.6540
	baseline (Tayyar Madabushi et al., 2021)	0.7070	0.6803	0.5065	0.6540
17	kpfriends (Sik Oh, 2022)	0.7256	0.6739	0.4918	0.6488
18	Unimelb_AIP	0.7614	0.6251	0.5020	0.6436
19	YNU-HPCC (Liu et al., 2022)	0.7063	0.6509	0.4805	0.6369
20	Ryan Wang	0.5972	0.4943	0.4608	0.5331
N/A	JARVix (Jakhotiya et al., 2022)[6]	0.7869	0.7201	0.5588	0.7235

Table 5: Results for Subtask A Zero Shot. The evaluation metric is macro F1 score, and the ranking is based on the 'All' column.

What about Contextual Understanding?

Ranking	Team	Language: English	Portuguese	Galician	All
1	clay	0.9016	0.8277	0.9278	0.8895
2	yxb	0.8948	0.8395	0.7524	0.8498
3	NER4ID (Tedeschi and Navigli, 2022)	0.8680	0.7039	0.6550	0.7740
4	HIT (Chu et al., 2022)	0.8242	0.7591	0.6866	0.7715
5	Hitachi (Yamaguchi et al., 2022)	0.7827	0.7607	0.6631	0.7466
6	OCHADAI (Pereira and Kobayashi, 2022)	0.7865	0.7700	0.6518	0.7457
7	yjs	0.8253	0.7424	0.6020	0.7409
8	CardiffNLP-metaphors (Boisson et al., 2022)	0.7637	0.7619	0.6591	0.7378
9	Mirs	0.7663	0.7617	0.6429	0.7338
10	Amobee	0.7597	0.7147	0.6768	0.7250
11	HYU (Joung and Kim, 2022)	0.7642	0.7282	0.6293	0.7227
12	Zhichun Road (Cui et al., 2022)	0.7489	0.6901	0.5104	0.6831
13	海鲛NLP	0.7564	0.6933	0.5108	0.6776
14	UAlberta (Hauer et al., 2022)	0.7099	0.6558	0.5646	0.6647
15	Helsinki-NLP (Itkonen et al., 2022)	0.7523	0.6939	0.4987	0.6625
16	daminglu123 (Lu, 2022)	0.7070	0.6803	0.5065	0.6540
	baseline (Tayyar Madabushi et al., 2021)	0.7070	0.6803	0.5065	0.6540
17	kpfriends (Sik Oh, 2022)	0.7256	0.6739	0.4918	0.6488
18	Unimelb_AIP	0.7614	0.6251	0.5020	0.6436
19	YNU-HPCC (Liu et al., 2022)	0.7063	0.6509	0.4805	0.6369
20	Ryan Wang	0.5972	0.4943	0.4608	0.5331
N/A	JARVix (Jakhotiya et al., 2022)[6]	0.7869	0.7201	0.5588	0.7235

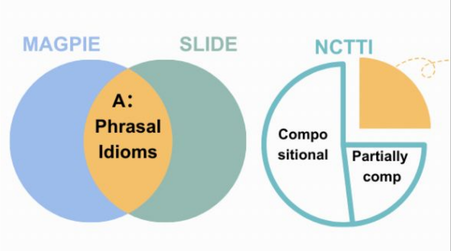
Table 5: Results for Subtask A Zero Shot. The evaluation metric is macro F1 score, and the ranking is based on the 'All' column.

Contextual Understanding of Idiomatic Expressions

Figurative	**Literal**
Even if Jack Bernstein hadn't **let the cat out of the bag** I would have known!’	During her move, Samantha had to **let the cat out of the bag** after it had crawled in amongst the linens.
If you do not believe me , then listen to how Steffi Graf and Monica Seles **let the cat out of the bag** in Paris.	While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.
It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag** .	Amy gasped in surprise when she opened her birthday present, only to **let the cat out of the bag**, having been tricked by her siblings.

DICE: Dataset for Idiomatic Contrastive Evaluation

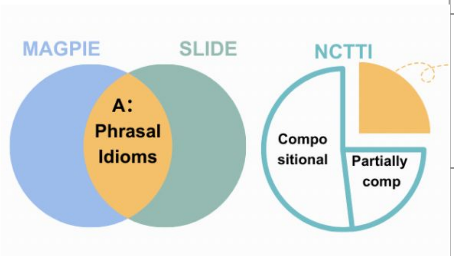
Existing idiomaticity datasets



Figurative	**Literal**
Even if Jack Bernstein hadn't **let the cat out of the bag** I would have known!’	During her move, Samantha had to **let the cat out of the bag** after it had crawled in amongst the linens.
If you do not believe me , then listen to how Steffi Graf and Monica Seles **let the cat out of the bag** in Paris.	While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.
It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag** .	Amy gasped in surprise when she opened her birthday present, only to **let the cat out of the bag**, having been tricked by her siblings.

DICE: Dataset for Idiomatic Contrastive Evaluation

Existing idiomaticity datasets



Figurative	Literal
Even if Jack Bernstein hadn't **let the cat out of the bag** I would have known!'	During her move, Samantha had to **let the cat out of the bag** after it had crawled in amongst the linens.
If you do not believe me , then listen to how Steffi Graf and Monica Seles **let the cat out of the bag** in Paris.	While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.
It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag** .	Amy gasped in surprise when she opened her birthday present, only to **let the cat out of the bag**, having been tricked by her siblings.

GPT-4

DICE: Dataset for Idiomatic Contrastive Evaluation

	Counts
Number of Sentences (Literal)	1033
Number of Sentences (Figurative)	1033
Total no. of sentences	2066
Number of Unique Idioms	402
Total Number of Expressions	402
Average length of sentences (literal)	15.4 words
Average length of sentences (figurative)	28.1 words

Evaluation

- Accuracy
- Lenient Consistency
  - The model is rewarded
  - For understanding the figurative use of x in all its variations
  - For understanding the literal use of x in all its variations

$$\text{Lenient Consistency} = \frac{\sum_x \mathbf{1}\left(\forall i, \text{Pred}(x_i^{Lit}) = \text{Lit}) + \mathbf{1}(\forall i, \text{Pred}(x_i^{Fig}) = \text{Fig}\right)}{2 * \text{Number of unique expressions}}$$

Lenient Consistency

The model is rewarded

- For understanding the figurative use of x in all its variations
- For understanding the literal use of x in all its variations

Figurative	Literal
✓ Even if Jack Bernstein hadn't **let the cat out of the bag** I would have known!'	✓ During her move, Samantha had to **let the cat out of the bag** after it had crawled in amongst the linens.
❌If you do not believe me , then listen to how Steffi Graf and Monica Seles **let the cat out of the bag** in Paris.	✓ While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.
✓ It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag** .	✓ Amy gasped in surprise when she opened her birthday present, only to **let the cat out of the bag**, having been tricked by her siblings.

Strict Consistency

The model is rewarded if it correctly detects all figurative and literal variation of x

$$\text{Strict Consistency} = \frac{\sum_{x \in \mathcal{X}} \mathbf{1}\left(\forall i, \text{Prediction}(x_i) = \text{True Label}(x_i)\right)}{\text{Number of unique expressions}}$$

Strict Consistency

The model is rewarded if it correctly detects all figurative and literal variation of x

$$\text{Strict Consistency} = \frac{\sum_{x \in \mathcal{X}} \mathbf{1}\left(\forall i, \text{Prediction}(x_i) = \text{True Label}(x_i)\right)}{\text{Number of unique expressions}}$$

Figurative	**Literal**
✓ Even if Jack Bernstein hadn't **let the cat out of the bag** I would have known!'	✓ During her move, Samantha had to **let the cat out of the bag** after it had crawled in amongst the linens.
❌If you do not believe me , then listen to how Steffi Graf and Monica Seles **let the cat out of the bag** in Paris.	✓ While unpacking groceries, Ben had to momentarily **let the cat out of the bag**, as it had sneakily found its way in.
✓ It was the kind of story that she would relish but with her blunt ways one could never be sure she wouldn't at some time **let the cat out of the bag** .	✓ Amy gasped in surprise when she opened her birthday present, only to **let the cat out of the bag**, having been tricked by her siblings.

Evaluation

3 different prompts

- Is the expression ’idiom’ used figuratively or literally in the sentence: ’sentence’. Answer ’i’ for figurative, ’l’ for literal.
- In the sentence ’sentence’, is the expression ’idiom’ being used figuratively or literally? Respond with ’i’ for figurative and ’l’ for literal.
- How is the expression ’idiom’ used in this context: ’sentence’. Output ’i’ if the expression holds figurative meaning, output ’l’ if the expression holds literal meaning.

LLMs' (Lack of) Robustness!

Model	Accuracy			Lenient Consistency			Strict Consistency
	Figurative	Literal	Overall	Figurative	Literal	Overall	Both Settings
GPT-4o	87.05 ± 3.62	87.30 ± 2.98	84.33 ± 4.44	69.49 ± 11.71	71.06 ± 6.68	70.32 ± 7.11	48.59 ± 9.75
GPT-3.5 Turbo	79.05 ± 5.01	70.02 ± 12.72	75.54 ± 7.81	82.59 ± 9.17	44.36 ± 22.28	63.47 ± 7.61	32.84 ± 15.81
Flan-T5-XXL (11B)	77.18 ± 1.40	74.91 ± 8.35	76.40 ± 4.49	63.93 ± 13.71	58.79 ± 23.16	61.36 ± 4.73	32.92 ± 6.80
Flan-T5-XL (3B)	70.48 ± 3.56	33.94 ± 26.91	59.65 ± 8.19	91.13 ± 6.97	13.02 ± 11.24	52.07 ± 3.58	9.95 ± 8.88
Flan-T5-Large (780M)	66.63 ± 0.10	3.45 ± 4.72	50.42 ± 0.53	97.68 ± 3.40	0.58 ± 0.80	49.13 ± 1.30	0.58 ± 0.80
Flan-T5-Small (80M)	0.51 ± 0.59	66.72 ± 0.07	50.13 ± 0.15	0.00 ± 0.00	100.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00
Llama 3.1 (405B)	88.63 ± 2.36	88.25 ± 3.93	**88.45 ± 3.10**	78.52 ± 5.61	80.02 ± 12.43	**79.27 ± 3.46**	60.36 ± 6.61
Llama 3 (70B)	87.72 ± 4.63	86.13 ± 7.10	87.00 ± 5.73	81.84 ± 4.00	72.64 ± 16.12	77.24 ± 7.45	57.55 ± 12.41
Llama 3 (8B)	79.27 ± 1.97	74.01 ± 2.79	76.91 ± 2.25	77.86 ± 5.18	48.76 ± 3.37	63.31 ± 1.43	33.83 ± 2.60
Llama 2 (70B)	76.28 ± 4.39	56.64 ± 17.13	69.62 ± 7.82	93.20 ± 4.75	24.54 ± 16.89	59.12 ± 5.78	21.81 ± 13.51
Llama 2 (13B)	68.99 ± 1.39	36.09 ± 3.85	58.26 ± 1.96	85.41 ± 3.56	8.37 ± 3.34	46.93 ± 2.30	5.64 ± 2.00
Llama 2 (7B)	55.51 ± 19.54	31.97 ± 24.25	51.34 ± 1.55	59.87 ± 46.26	18.08 ± 29.16	38.97 ± 8.59	1.66 ± 1.37
GPT-4	88.56 ± 2.03	88.63 ± 2.08	88.48 ± 2.18	79.02 ± 3.11	78.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67

Path to True Idiomaticity Understanding

Model	Accuracy			Lenient Consistency			Strict Consistency
	Figurative	Literal	Overall	Figurative	Literal	Overall	Both Settings
GPT-4o	87.05 ± 3.62	87.30 ± 2.98	84.33 ± 4.44	69.49 ± 11.71	71.06 ± 6.68	70.32 ± 7.11	48.59 ± 9.75
GPT-3.5 Turbo	79.05 ± 5.01	70.02 ± 12.72	75.54 ± 7.81	82.59 ± 9.17	44.36 ± 22.28	63.47 ± 7.61	32.84 ± 15.81
Flan-T5-XXL (11B)	77.18 ± 1.40	74.91 ± 8.35	76.40 ± 4.49	63.93 ± 13.71	58.79 ± 23.16	61.36 ± 4.73	32.92 ± 6.80
Flan-T5-XL (3B)	70.48 ± 3.56	33.94 ± 26.91	59.65 ± 8.19	91.13 ± 6.97	13.02 ± 11.24	52.07 ± 3.58	9.95 ± 8.88
Flan-T5-Large (780M)	66.63 ± 0.10	3.45 ± 4.72	50.42 ± 0.53	97.68 ± 3.40	0.58 ± 0.80	49.13 ± 1.30	0.58 ± 0.80
Flan-T5-Small (80M)	0.51 ± 0.59	66.72 ± 0.07	50.13 ± 0.15	0.00 ± 0.00	100.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00
Llama 3.1 (405B)	88.63 ± 2.36	88.25 ± 3.93	**88.45 ± 3.10**	78.52 ± 5.61	80.02 ± 12.43	**79.27 ± 3.46**	60.36 ± 6.61
Llama 3 (70B)	87.72 ± 4.63	86.13 ± 7.10	87.00 ± 5.73	81.84 ± 4.00	72.64 ± 16.12	77.24 ± 7.45	57.55 ± 12.41
Llama 3 (8B)	79.27 ± 1.97	74.01 ± 2.79	76.91 ± 2.25	77.86 ± 5.18	48.76 ± 3.37	63.31 ± 1.43	33.83 ± 2.60
Llama 2 (70B)	76.28 ± 4.39	56.64 ± 17.13	69.62 ± 7.82	93.20 ± 4.75	24.54 ± 16.89	59.12 ± 5.78	21.81 ± 13.51
Llama 2 (13B)	68.99 ± 1.39	36.09 ± 3.85	58.26 ± 1.96	85.41 ± 3.56	8.37 ± 3.34	46.93 ± 2.30	5.64 ± 2.00
Llama 2 (7B)	55.51 ± 19.54	31.97 ± 24.25	51.34 ± 1.55	59.87 ± 46.26	18.08 ± 29.16	38.97 ± 8.59	1.66 ± 1.37
GPT-4	88.56 ± 2.03	88.63 ± 2.08	88.48 ± 2.18	79.02 ± 3.11	78.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67

Path to True Idiomaticity Understanding

Model		nsistency		Strict Consistency
	Figurative	iteral	Overall	Both Settings
GPT-4o	87.05 ± 3.62	± 6.68	70.32 ± 7.11	48.59 ± 9.75
GPT-3.5 Turbo	79.05 ± 5.01	22.28	63.47 ± 7.61	32.84 ± 15.81
Flan-T5-XXL (11B)	77.18 ± 1.40	23.16	61.36 ± 4.73	32.92 ± 6.80
Flan-T5-XL (3B)	70.48 ± 3.56	11.24	52.07 ± 3.58	9.95 ± 8.88
Flan-T5-Large (780M)	66.63 ± 0.10	± 0.80	49.13 ± 1.30	0.58 ± 0.80
Flan-T5-Small (80M)	0.51 ± 0.59	± 0.00	50.00 ± 0.00	0.00 ± 0.00
Llama 3.1 (405B)	88.63 ± 2.36	12.43	**79.27 ± 3.46**	60.36 ± 6.61
Llama 3 (70B)	87.72 ± 4.63	16.12	77.24 ± 7.45	57.55 ± 12.41
Llama 3 (8B)	79.27 ± 1.97	± 3.37	63.31 ± 1.43	33.83 ± 2.60
Llama 2 (70B)	76.28 ± 4.39	16.89	59.12 ± 5.78	21.81 ± 13.51
Llama 2 (13B)	68.99 ± 1.39	± 3.34	46.93 ± 2.30	5.64 ± 2.00
Llama 2 (7B)	55.51 ± 19.54	29.16	38.97 ± 8.59	1.66 ± 1.37
GPT-4	88.56 ± 2.03	.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67

Performance of the Literal Data Generator!

Model	Accuracy			Lenient Consistency			Strict Consistency
	Figurative	Literal	Overall	Figurative	Literal	Overall	Both Settings
GPT-4o	87.05 ± 3.62	87.30 ± 2.98	84.33 ± 4.44	69.49 ± 11.71	71.06 ± 6.68	70.32 ± 7.11	48.59 ± 9.75
GPT-3.5 Turbo	79.05 ± 5.01	70.02 ± 12.72	75.54 ± 7.81	82.59 ± 9.17	44.36 ± 22.28	63.47 ± 7.61	32.84 ± 15.81
Flan-T5-XXL (11B)	77.18 ± 1.40	74.91 ± 8.35	76.40 ± 4.49	63.93 ± 13.71	58.79 ± 23.16	61.36 ± 4.73	32.92 ± 6.80
Flan-T5-XL (3B)	70.48 ± 3.56	33.94 ± 26.91	59.65 ± 8.19	91.13 ± 6.97	13.02 ± 11.24	52.07 ± 3.58	9.95 ± 8.88
Flan-T5-Large (780M)	66.63 ± 0.10	3.45 ± 4.72	50.42 ± 0.53	97.68 ± 3.40	0.58 ± 0.80	49.13 ± 1.30	0.58 ± 0.80
Flan-T5-Small (80M)	0.51 ± 0.59	66.72 ± 0.07	50.13 ± 0.15	0.00 ± 0.00	100.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00
Llama 3.1 (405B)	88.63 ± 2.36	88.25 ± 3.93	**88.45 ± 3.10**	78.52 ± 5.61	80.02 ± 12.43	**79.27 ± 3.46**	60.36 ± 6.61
Llama 3 (70B)	87.72 ± 4.63	86.13 ± 7.10	87.00 ± 5.73	81.84 ± 4.00	72.64 ± 16.12	77.24 ± 7.45	57.55 ± 12.41
Llama 3 (8B)	79.27 ± 1.97	74.01 ± 2.79	76.91 ± 2.25	77.86 ± 5.18	48.76 ± 3.37	63.31 ± 1.43	33.83 ± 2.60
Llama 2 (70B)	76.28 ± 4.39	56.64 ± 17.13	69.62 ± 7.82	93.20 ± 4.75	24.54 ± 16.89	59.12 ± 5.78	21.81 ± 13.51
Llama 2 (13B)	68.99 ± 1.39	36.09 ± 3.85	58.26 ± 1.96	85.41 ± 3.56	8.37 ± 3.34	46.93 ± 2.30	5.64 ± 2.00
Llama 2 (7B)	55.51 ± 19.54	31.97 ± 24.25	51.34 ± 1.55	59.87 ± 46.26	18.08 ± 29.16	38.97 ± 8.59	1.66 ± 1.37
GPT-4	88.56 ± 2.03	88.63 ± 2.08	88.48 ± 2.18	79.02 ± 3.11	78.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67

Best Model

Model	Accuracy			Lenient Consistency			Strict Consistency
	Figurative	Literal	Overall	Figurative	Literal	Overall	Both Settings
GPT-4o	87.05 ± 3.62	87.30 ± 2.98	84.33 ± 4.44	69.49 ± 11.71	71.06 ± 6.68	70.32 ± 7.11	48.59 ± 9.75
GPT-3.5 Turbo	79.05 ± 5.01	70.02 ± 12.72	75.54 ± 7.81	82.59 ± 9.17	44.36 ± 22.28	63.47 ± 7.61	32.84 ± 15.81
Flan-T5-XXL (11B)	77.18 ± 1.40	74.91 ± 8.35	76.40 ± 4.49	63.93 ± 13.71	58.79 ± 23.16	61.36 ± 4.73	32.92 ± 6.80
Flan-T5-XL (3B)	70.48 ± 3.56	33.94 ± 26.91	59.65 ± 8.19	91.13 ± 6.97	13.02 ± 11.24	52.07 ± 3.58	9.95 ± 8.88
Flan-T5-Large (780M)	66.63 ± 0.10	3.45 ± 4.72	50.42 ± 0.53	97.68 ± 3.40	0.58 ± 0.80	49.13 ± 1.30	0.58 ± 0.80
Flan-T5-Small (80M)	0.51 ± 0.59	66.72 ± 0.07	50.13 ± 0.15	0.00 ± 0.00	100.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00
Llama 3.1 (405B)	88.63 ± 2.36	88.25 ± 3.93	**88.45 ± 3.10**	78.52 ± 5.61	80.02 ± 12.43	**79.27 ± 3.46**	60.36 ± 6.61
Llama 3 (70B)	87.72 ± 4.63	86.13 ± 7.10	87.00 ± 5.73	81.84 ± 4.00	72.64 ± 16.12	77.24 ± 7.45	57.55 ± 12.41
Llama 3 (8B)	79.27 ± 1.97	74.01 ± 2.79	76.91 ± 2.25	77.86 ± 5.18	48.76 ± 3.37	63.31 ± 1.43	33.83 ± 2.60
Llama 2 (70B)	76.28 ± 4.39	56.64 ± 17.13	69.62 ± 7.82	93.20 ± 4.75	24.54 ± 16.89	59.12 ± 5.78	21.81 ± 13.51
Llama 2 (13B)	68.99 ± 1.39	36.09 ± 3.85	58.26 ± 1.96	85.41 ± 3.56	8.37 ± 3.34	46.93 ± 2.30	5.64 ± 2.00
Llama 2 (7B)	55.51 ± 19.54	31.97 ± 24.25	51.34 ± 1.55	59.87 ± 46.26	18.08 ± 29.16	38.97 ± 8.59	1.66 ± 1.37
GPT-4	88.56 ± 2.03	88.63 ± 2.08	88.48 ± 2.18	79.02 ± 3.11	78.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67

One-shot Results, Not Much Better!

GPT-4o	89.43 ± 1.23	90.15 ± 1.71	**89.72 ± 1.45**	74.63 ± 1.99	87.40 ± 5.81	**81.01 ± 1.93**	63.52 ± 3.15
GPT-3.5 Turbo	79.41 ± 4.19	72.69 ± 10.87	76.70 ± 6.54	78.44 ± 8.80	49.42 ± 18.96	63.93 ± 5.92	34.16 ± 12.19
Flan-T5-XXL (11B)	10.20 ± 15.69	67.90 ± 1.91	52.79 ± 4.34	1.58 ± 2.52	99.25 ± 1.29	50.41 ± 0.61	1.49 ± 2.37
Flan-T5-XL (3B)	0.64 ± 0.80	66.71 ± 0.11	50.13 ± 0.22	0.08 ± 0.14	99.83 ± 0.29	49.96 ± 0.19	0.08 ± 0.14
Flan-T5-Large (780M)	3.28 ± 3.64	66.27 ± 0.45	50.00 ± 0.00	0.66 ± 0.76	96.93 ± 3.73	48.80 ± 1.48	0.00 ± 0.00
Flan-T5-Small (80M)	45.23 ± 39.19	35.55 ± 33.55	53.03 ± 5.25	60.78 ± 53.37	37.31 ± 54.62	49.05 ± 1.65	2.40 ± 4.16
Llama 3.1 (405B)	89.57 ± 1.80	89.54 ± 2.54	89.53 ± 2.17	79.10 ± 3.26	82.01 ± 7.85	80.56 ± 2.56	63.27 ± 4.66
Llama 3 (70B)	87.75 ± 3.76	86.97 ± 5.64	87.27 ± 4.61	78.52 ± 3.59	75.62 ± 14.01	77.07 ± 6.00	57.55 ± 10.22
Llama 3 (8B)	80.32 ± 5.33	73.81 ± 11.40	77.59 ± 7.62	79.35 ± 1.08	48.01 ± 15.70	63.68 ± 7.34	34.91 ± 13.59
Llama 2 (70B)	70.40 ± 1.19	31.44 ± 6.18	58.65 ± 2.28	96.52 ± 0.66	7.55 ± 2.75	52.03 ± 1.50	6.72 ± 2.63
Llama 2 (13B)	70.64 ± 1.15	34.20 ± 6.92	59.36 ± 2.45	94.94 ± 0.52	9.54 ± 4.14	52.24 ± 1.83	8.29 ± 3.11
Llama 2 (7B)	70.26 ± 3.14	42.18 ± 26.31	61.21 ± 9.28	80.76 ± 15.43	20.73 ± 22.25	50.75 ± 3.46	11.69 ± 10.42
GPT-4	88.52 ± 1.49	88.95 ± 2.09	88.42 ± 1.73	78.44 ± 0.76	77.94 ± 5.84	78.19 ± 2.63	58.87 ± 4.86

Analysis

Impact of Pretraining Term Frequencies on Few-Shot Reasoning

Yasaman Razeghi [1] **Robert L. Logan IV** [1] **Matt Gardner** [2] **Sameer Singh** [1 3]

- (Estimated) Frequency in the pretraining data

Q: What is 24 times 18? A: ___ *Model:* 432 ✓
Q: What is 23 times 18? A: ___ *Model:* 462 ✗

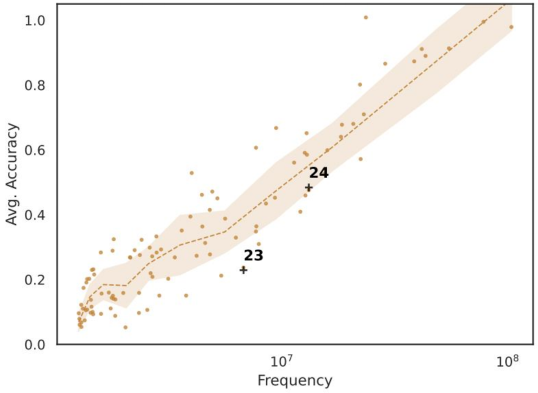


Figure 1. **Multiplication Performance:** Plot of GPT-J-6B's 2-shot accuracy on multiplication (averaged over multiple multiplicands and training instances) against the frequency of the equation's first term in the pretraining corpus. Each point represents the

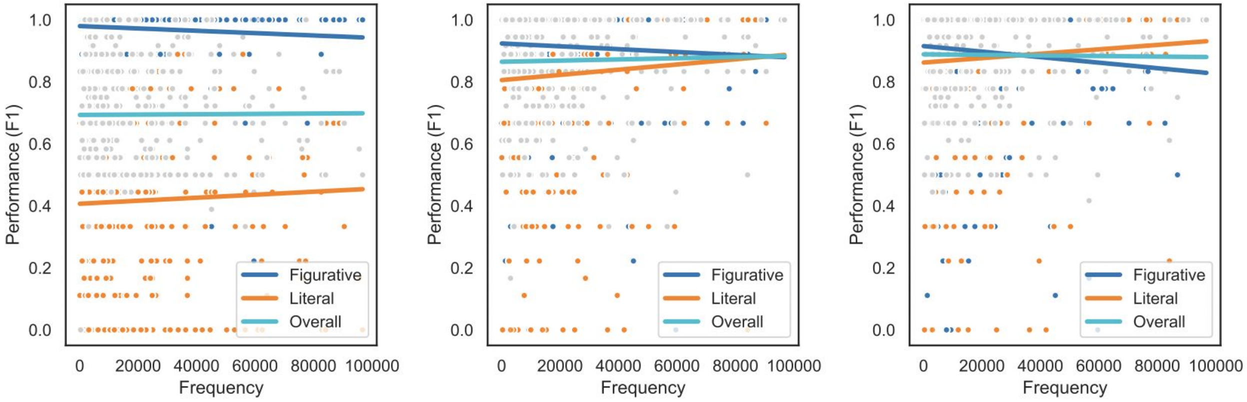


Figure 6: Left to right: Frequency analysis for Llama 3.1 (405B), Llama 3 (70B) and Llama 2 (70B).

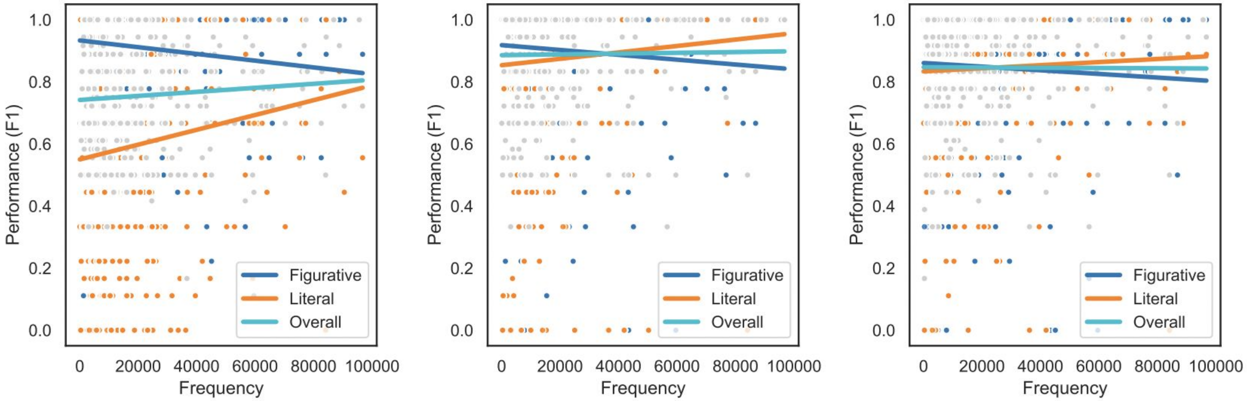


Figure 7: Left to right: Frequency analysis for GPT-3.5 Turbo, GPT-4 and GPT-4o.

Frequency Analysis

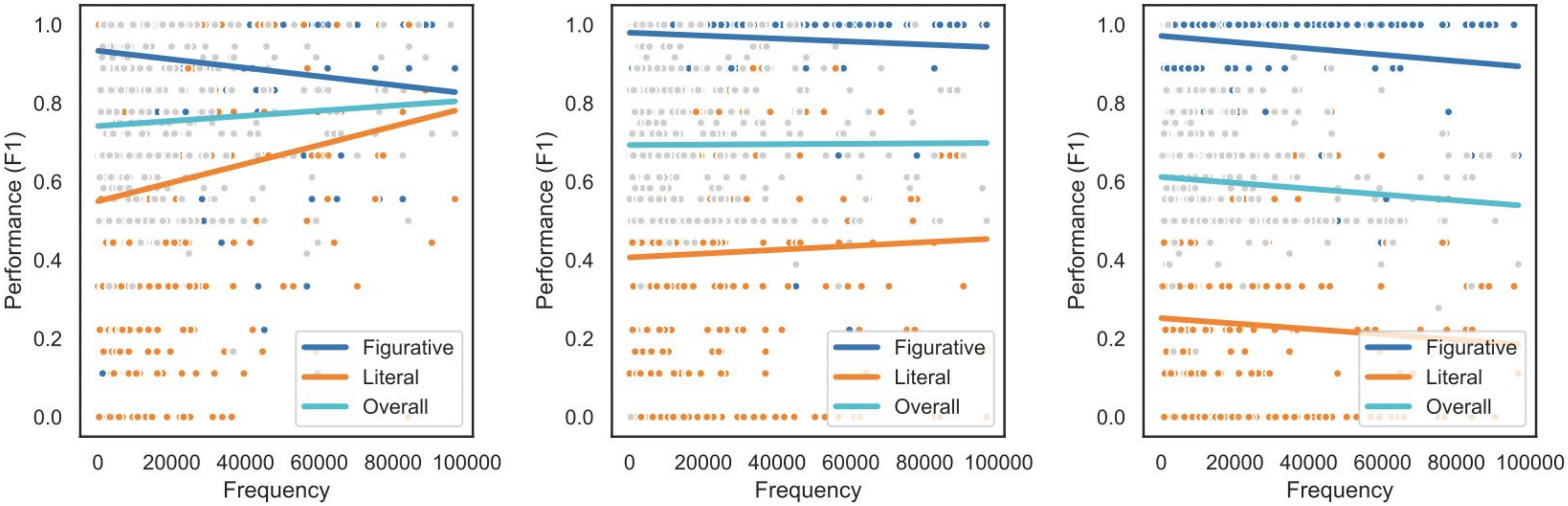
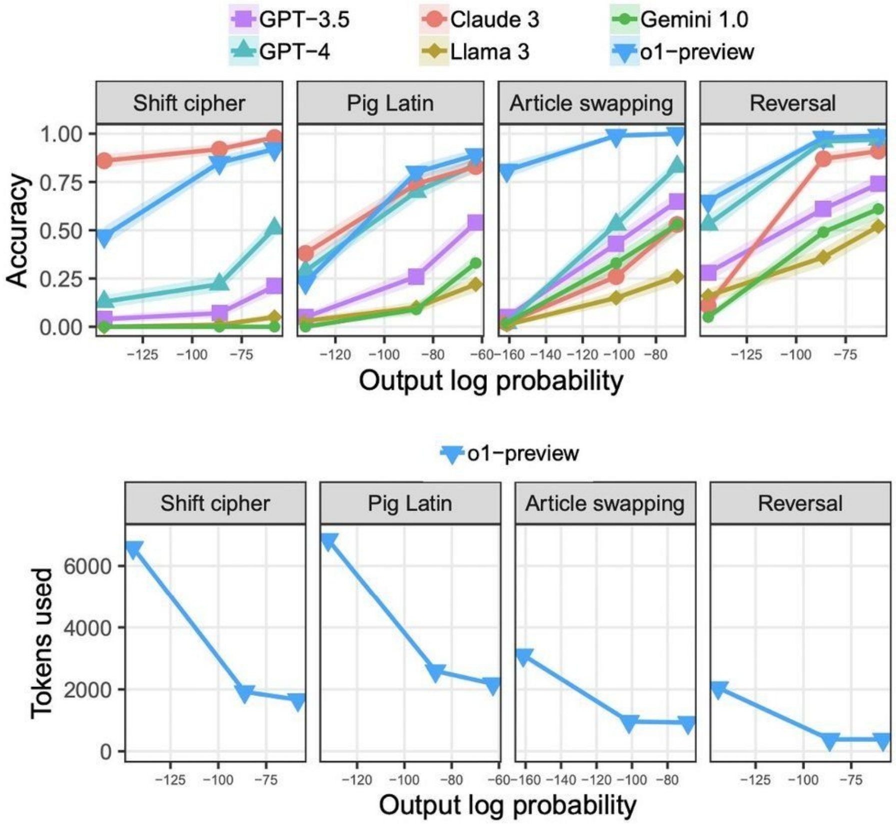


Figure 2: Frequency results of GPT-3.5 Turbo, Llama 2 (70B), Flan-T5 XL (left to right).

Likelihood Bias



When a language model is optimized for reasoning, does it still show embers of autoregression? An analysis of OpenAI o1

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, Thomas L. Griffiths

What Is Model Likelihood?

How confident a language model is in the words it chooses

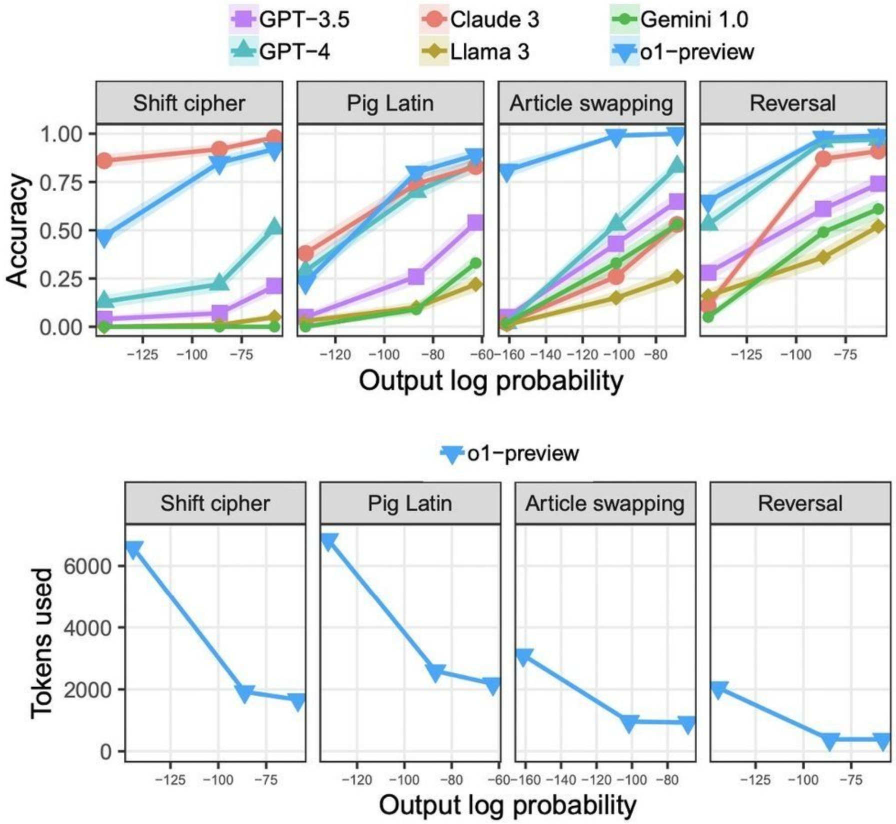
- LLMs generate text one word at a time
- For each word, it assigns a probability to many possible next words
- The word with the highest likelihood is usually the one it picks

What Is Model Likelihood?

The cat sat on the ___

- “mat” with 80% probability ✅
- “sofa” with 15%
- “ceiling” with 5%

Likelihood Bias



When a language model is optimized for reasoning, does it still show embers of autoregression? An analysis of OpenAI o1

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, Thomas L. Griffiths

Likelihood Analysis

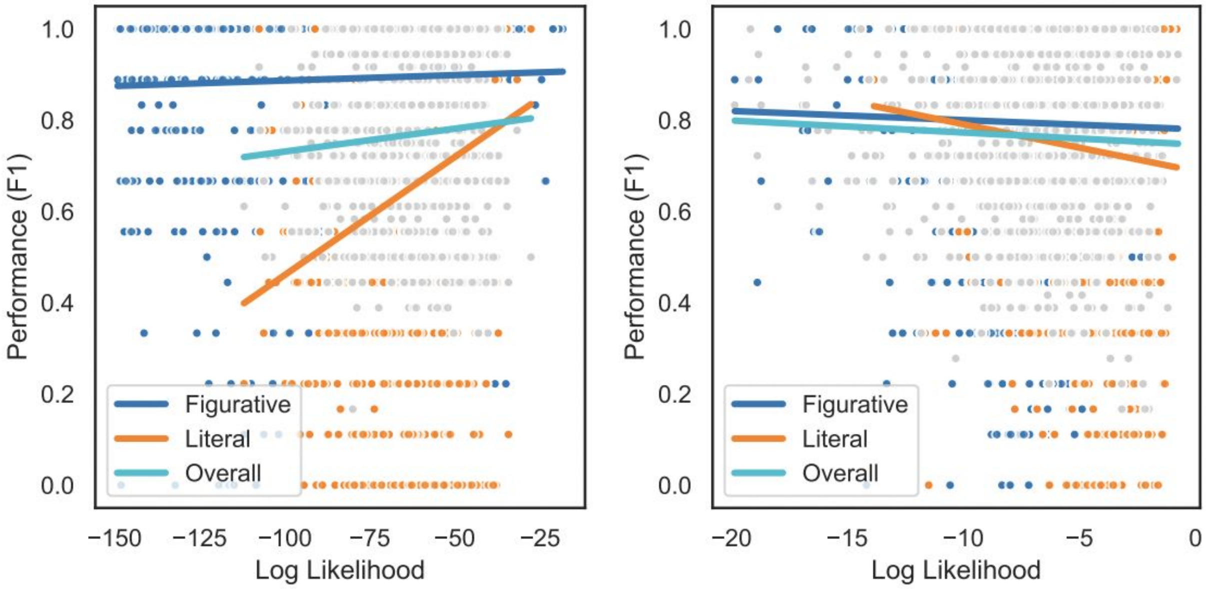


Figure 4: Likelihood results from Llama 3 (8B) and Flan-T5 XXL (left to right).

Likelihood Analysis

Flan-T5 XL

Flan-T5 XXL

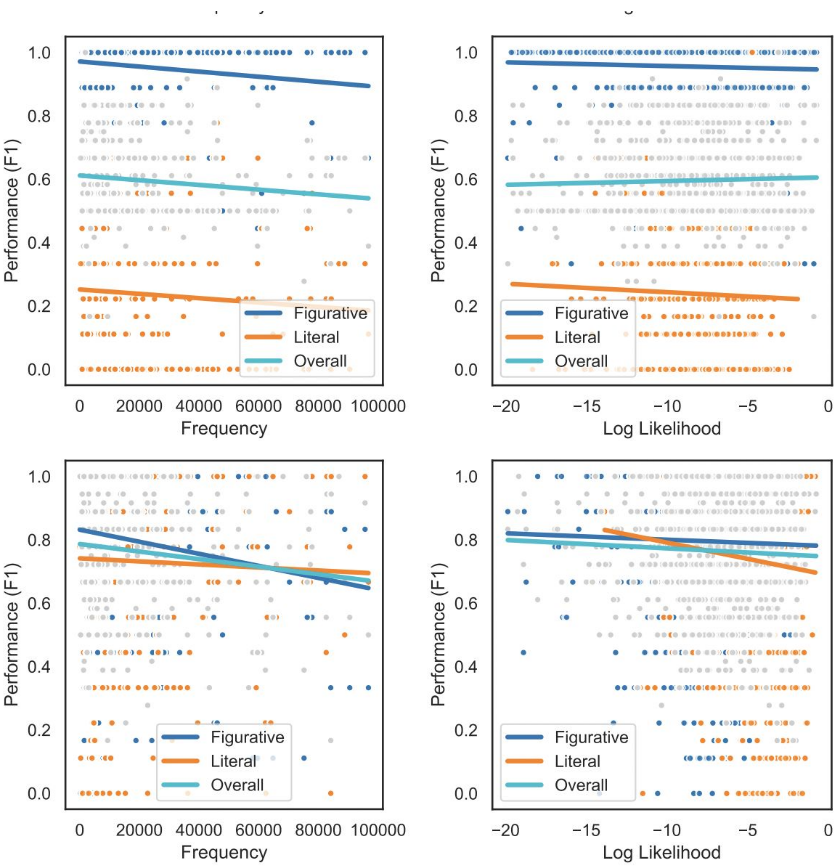


Figure 9: Visualisations of the frequency and likelihood analysis. Flan-T5 models only.

Likelihood Analysis

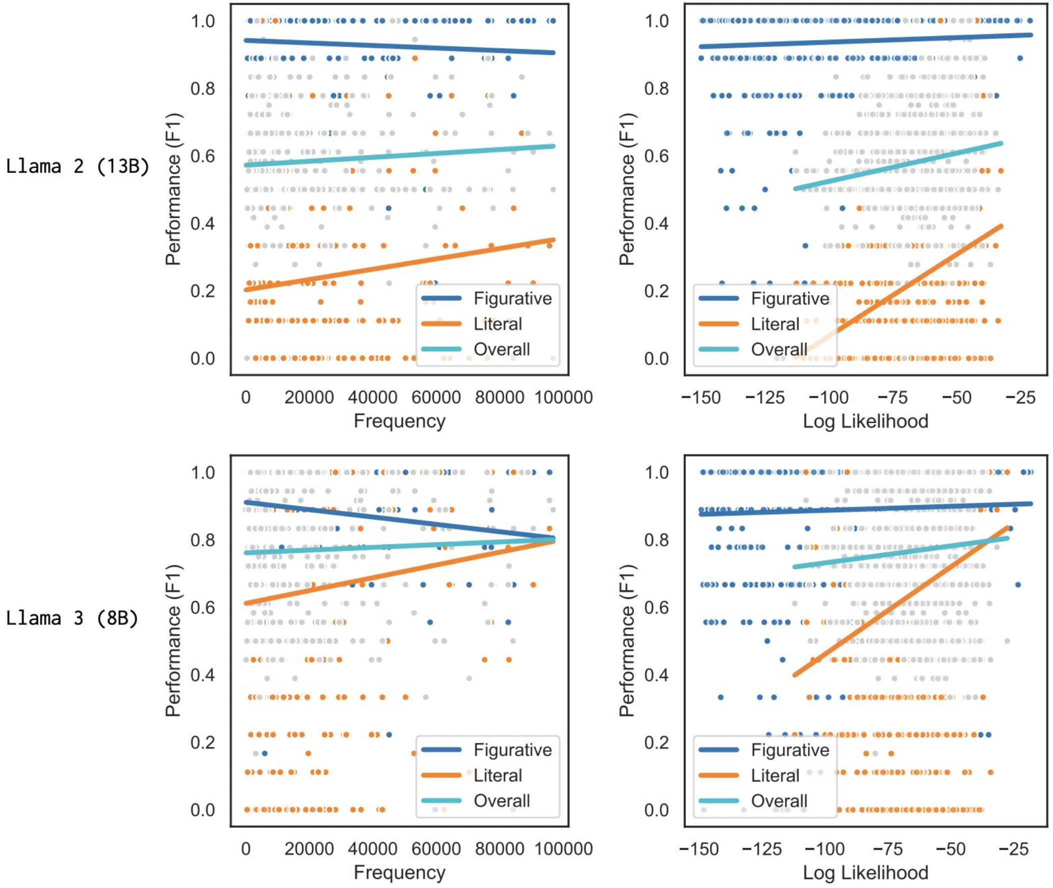


Figure 8: Visualisations of the frequency and likelihood. Smaller Llama models only.

Beyond Hate Speech: NLP's Challenges and Opportunities in Uncovering Dehumanizing Language

Hezhao Zhang Lasana Harris

Dehumanization

The denial of “humanness” to others

Fostering conditions that result in extreme and violent behaviors against marginalized groups

Dehumanization: trends, insights, and challenges

Nour S. Kteily [1,@] · Alexander P. Landry [2]

Dehumanization

Blatant: Overt derogation, where victims are likened to “dogs” or “monkeys”

Subtle: Denying the capability of experiencing pain or other human emotions to certain individuals

Allowing people to harm others while minimizing, ignoring, or misconstruing the consequences

Dehumanization

"Dehumanization has enabled members of advantaged groups to 'morally disengage' from disadvantaged group suffering, thereby facilitating acts of intergroup aggression such as colonization, slavery and genocide"

The enemy as animal: Symmetric dehumanization during asymmetric warfare

Emile Bruneau [1,2,*,#], Nour Kteily [3,#]

Dehumanization

Nations tend to cast their enemies using dehumanized images to make their killing easier

IMAGES OF SAVAGERY IN AMERICAN JUSTIFICATIONS FOR WAR

ROBERT L. IVIE

This paper identifies the essential characteristics of victimage rhetoric in American justifications for war. The Johnson administration's insistence on the aggression-from-the-North thesis is the starting point for the analysis. Close inspection of the administration's efforts reveals that the enemy is portrayed as a savage, i.e., an aggressor, driven by irrational desires for conquest, who is seeking to subjugate others by force of arms. This image of the enemy is intensified by a

Dehumanization, an ongoing example …

AMNESTY INTERNATIONAL

ENGLISH

WHO WE ARE

WHAT WE DO

< RESEARCH

December 5, 2024
Index Number: MDE 15/8668/2024

Israel/Occupied Palestinian Territory: 'You Feel Like You Are Subhuman': Israel's Genocide Against Palestinians in Gaza

This report documents Israel's actions during its offensive on the occupied Gaza Strip from 7 October 2023. It examines the killing of civilians, damage to and destruction of civilian infrastructure, forcible displacement, the obstruction or denial of life-saving goods and humanitarian aid, and the restriction of power supplies. It analyses Israel's intent through this pattern of conduct and statements by Israeli decision-makers. It concludes that Israel has committed genocide against Palestinians in Gaza.

How Good are LLMs at Identifying Dehumanizing Language?

Evaluation Data

Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, Douwe Kiela

Label	Type	Total
Hate	Not given	7,197
	Animosity	3,439
	$\boxed{\text{Dehumanization}}$	906
	Derogation	9,907
	Support	207
	Threatening	606
	Total	22,262
Not Hate	/	18,993
All	TOTAL	41,255

Evaluation Data

Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, Douwe Kiela

General Dehumanization

- 906 dehumanization instances, different targeted groups
- 906 randomly selected instances

Label	Type	Total
Hate	Not given	7,197
	Animosity	3,439
	Dehumanization	906
	Derogation	9,907
	Support	207
	Threatening	606
	Total	22,262
Not Hate	/	18,993
All	TOTAL	41,255

Evaluation Data

Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, Douwe Kiela

Dehumanization vs. Hate

- 906 dehumanization instances, different targeted groups
- 906 other hate types

Label	Type	Total
Hate	Not given	7,197
	Animosity	3,439
	Dehumanization	906
	Derogation	9,907
	Support	207
	Threatening	606
	Total	22,262
Not Hate	/	18,993
All	TOTAL	41,255

Evaluation

Models

- Claude-3-7-Sonnet
- GPT-4.1-mini
- Mistral (7B)
- Qwen2.5 (7B)

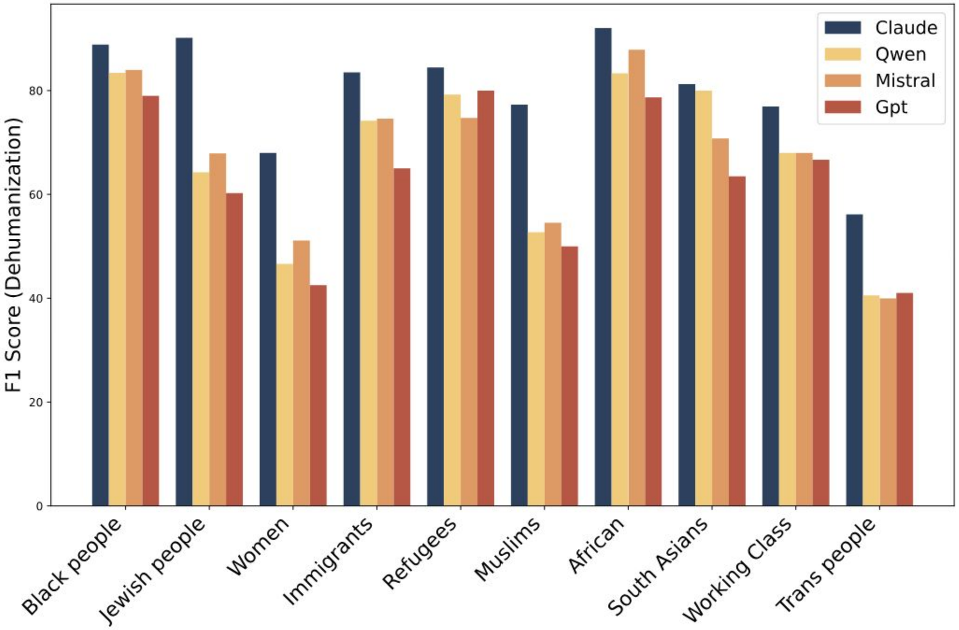
Evaluation

Prompt Type	Label Output	Key Prompt Instructions
Zero-shot	Binary (True/False) for each target group	Identify target groups in the text. Decide whether each target is dehumanized. Respond in JSON format: `{ “Targets”: [...], “Dehumanization”: [[target1, true/false], ...] }`
Few-shot	Blatant / Subtle / None for each target group	Given labeled examples, identify target groups and classify each as “Blatant”, “Subtle”, or “None”. Use the format: `[ { “Target”: “...”, “Dehumanization”: “Blatant”/“Subtle”/“None” }, ... ]`
Explainable	Blatant / Subtle / None + Explanation	Same as few-shot, but provide a short explanation for each label: `[ { “Target”: “...”, “Dehumanization”: “...”, ‘Explanation”: “...” }, ... ]`

Model	Prompt	Label Criterion	General Dehumanization			Dehumanization vs Hate		
			F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
GPT	Zero-shot	Binary	37.73	58.61	50.28	25.91	61.70	49.50
	Few-shot	Blatant only	51.65	48.23	50.00	48.95	49.51	49.23
	Explainable	Blatant only	52.22	45.08	48.90	50.56	47.00	48.84
	Few-shot	Blatant+Subtle	30.05	60.02	49.12	19.48	64.03	50.28
	Explainable	Blatant+Subtle	29.64	59.88	48.90	16.97	64.00	49.78
Qwen	Zero-shot	Binary	51.23	71.42	63.96	31.42	66.72	55.19
	Few-shot	Blatant only	73.91	71.34	72.68	68.57	68.29	68.43
	Explainable	Blatant only	71.97	67.81	70.03	65.19	63.46	64.35
	Few-shot	Blatant+Subtle	50.78	73.25	65.34	29.24	68.55	56.46
	Explainable	Blatant+Subtle	49.53	72.70	64.57	27.83	67.68	55.35
Mistral	Zero-shot	Binary	58.33	73.28	67.44	38.49	67.93	57.84
	Few-shot	Blatant only	53.69	55.65	54.69	53.05	55.28	54.19
	Explainable	Blatant only	60.57	54.83	57.89	60.10	54.16	57.33
	Few-shot	Blatant+Subtle	47.19	63.59	56.90	42.97	62.27	54.58
	Explainable	Blatant+Subtle	50.18	68.67	61.53	43.49	66.70	58.09
Claude	Zero-shot	Binary	56.90	75.57	68.82	20.50	67.83	54.19
	Few-shot	Blatant only	81.74	84.67	83.33	75.05	80.99	78.42
	Explainable	Blatant only	83.16	85.82	84.60	72.49	80.06	76.88
	Few-shot	Blatant+Subtle	53.14	75.12	67.49	17.06	68.04	53.86
	Explainable	Blatant+Subtle	50.12	74.53	66.28	14.13	67.68	53.04

Model	Prompt	Label Criterion	General Dehumanization			Dehumanization vs Hate		
			F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
GPT	Zero-shot	Binary	37.73	58.61	50.28	25.91	61.70	49.50
	Few-shot	Blatant only	51.65	48.23	50.00	48.95	49.51	49.23
	Explainable	Blatant only	52.22	45.08	48.90	50.56	47.00	48.84
	Few-shot	Blatant+Subtle	30.05	60.02	49.12	19.48	64.03	50.28
	Explainable	Blatant+Subtle	29.64	59.88	48.90	16.97	64.00	49.78
Qwen	Zero-shot	Binary	51.23	71.42	63.96	31.42	66.72	55.19
	Few-shot	Blatant only	73.91	71.34	72.68	68.57	68.29	68.43
	Explainable	Blatant only	71.97	67.81	70.03	65.19	63.46	64.35
	Few-shot	Blatant+Subtle	50.78	73.25	65.34	29.24	68.55	56.46
	Explainable	Blatant+Subtle	49.53	72.70	64.57	27.83	67.68	55.35
Mistral	Zero-shot	Binary	58.33	73.28	67.44	38.49	67.93	57.84
	Few-shot	Blatant only	53.69	55.65	54.69	53.05	55.28	54.19
	Explainable	Blatant only	60.57	54.83	57.89	60.10	54.16	57.33
	Few-shot	Blatant+Subtle	47.19	63.59	56.90	42.97	62.27	54.58
	Explainable	Blatant+Subtle	50.18	68.67	61.53	43.49	66.70	58.09
Claude	Zero-shot	Binary	56.90	75.57	68.82	20.50	67.83	54.19
	Few-shot	Blatant only	81.74	84.67	83.33	75.05	80.99	78.42
	Explainable	Blatant only	83.16	85.82	84.60	72.49	80.06	76.88
	Few-shot	Blatant+Subtle	53.14	75.12	67.49	17.06	68.04	53.86
	Explainable	Blatant+Subtle	50.12	74.53	66.28	14.13	67.68	53.04

Performance Disparities



Blind to Context, Prone to Bias

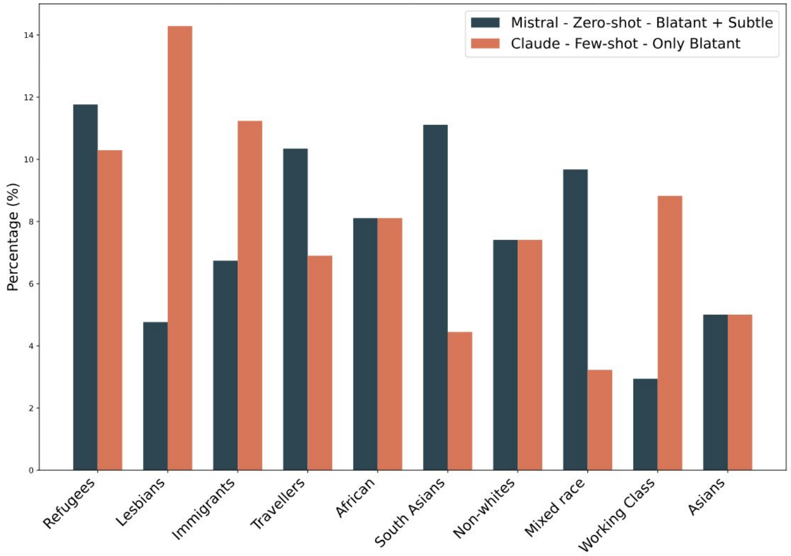


Figure 2: Recognition blindness of Claude and Mistral.

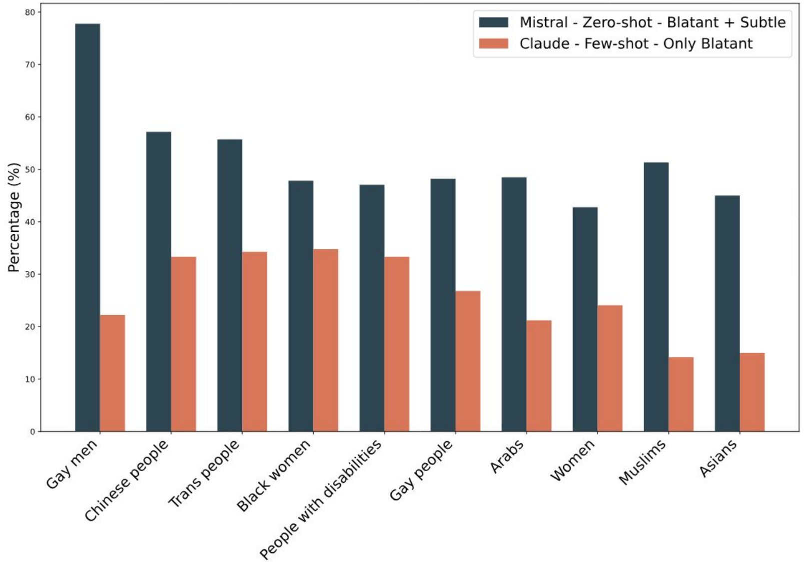
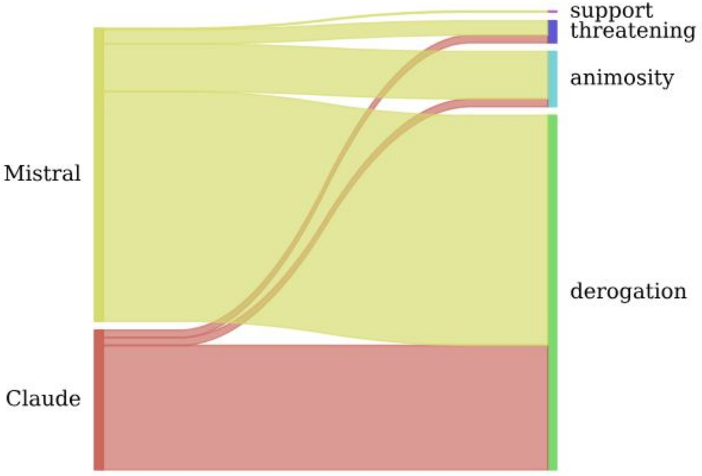


Figure 3: Over-sensitivity of Claude and Mistral.

Error Analysis



Error Analysis

Text	Label	Zero	Few	Exp.
Why don't <XX> just STFU and stay into the kitchen	Derogation	Dehuman	Blatant	Blatant
Next time I see a <ZZ> person, I'm going to b e a t them up	Threatening	Dehuman	Blatant	Blatant
Someone who had strong year and made over 75k in 2018 and who may be struggling currently gets zero but illegals and refugees get $350 mil	Animosity	Dehuman	Blatant	Subtle

Table 3: Examples from Claude outputs when the model **incorrectly classifies other types of hate speech as *dehumanization*.** Some parts of input texts are masked out to reduce their toxicity.

Summary

Contextual understanding remains a challenge

Leading to misinterpretations and flawed decision-making

Not just a linguistic issue

Bias amplification risks, disproportionately affecting different target groups

Crucial for building more reliable systems

Questions?