University of Sheffield | Healthy Lifespan Institute

Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection

Dylan Phelps, Thomas Pickard, Maggie Mi,
Edward Gow-Smith, Aline Villavicencio
University of Sheffield

University of Sheffield | Healthy Lifespan Institute

Introduction

University of Sheffield | Healthy Lifespan Institute

Motivation

Motivation

Motivation

Motivation

Contributions

In this paper we will:

- Evaluate the performance of multiple LLMs across multiple idiomaticity detection datasets

Contributions

In this paper we will:

- Evaluate the performance of multiple LLMs across multiple idiomaticity detection datasets
- Evaluate both closed source and open source models

Contributions

In this paper we will:

- Evaluate the performance of multiple LLMs across multiple idiomaticity detection datasets
- Evaluate both closed source and open source models
- Experiment with different prompts and prompting styles

Contributions

In this paper we will:

- Evaluate the performance of multiple LLMs across multiple idiomaticity detection datasets
- Evaluate both closed source and open source models
- Experiment with different prompts and prompting styles
- Discuss some of the practical considerations encountered whilst running our experiments

University of Sheffield | Healthy Lifespan Institute

Methodology

Tasks and Datasets

FLUTE:

- NLI dataset, covers a large range of figurative language
- 250 instances of 69 MWEs

Semeval 2022 Task 2a:

- Binary Classification task
- 2342 instances of 150 MWEs over 3 languages

MAGPIE:

- Classification Task
- 4,840 instances of 1134 MWEs

Construction Artifacts and Contamination

Recent paper on Construction Artifacts has shown that datasets can still perform well when the expression is hidden

- May affect our comparisons to fine-tuned models

Construction Artifacts and Contamination

Recent paper on Construction Artifacts has shown that datasets can still perform well when the expression is hidden

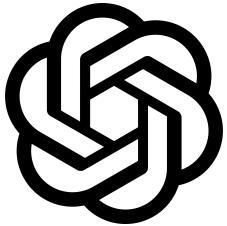
- May affect our comparisons to fine-tuned models

LLMs suffer from data contamination, so may have seen data before

- Seen data ‘in the wild’ and without idiomaticity labels

Models

SAAS Models



GPT 3.5 Turbo
GPT 4
GPT 4 Turbo

Gemini 1.5 Pro

Models

SAAS Models

GPT 3.5 Turbo
GPT 4
GPT 4 Turbo

Gemini

Gemini 1.5 Pro

Open Access Models

Phi-2
LLaMa2 (7B & 13B)
Mistral-7B (CH)

Models

SAAS Models

GPT 3.5 Turbo
GPT 4
GPT 4 Turbo

Gemini 1.5 Pro

Open Access Models

Phi-2
LLaMa2 (7B & 13B)
Mistral-7B (CH)

FLAN-T5 (Small, Base, Large, XL, XXL)

University of Sheffield | Healthy Lifespan Institute

Results

Overview of Results

Gemini, GPT-4, and GPT-4-turbo show similar results on SemEval and FLUTE.

Scaling we expect between GPT-3.5 and the other models.

	SemEval	FLUTE	MAGPIE
GPT-3.5-Turbo	0.645	0.820	0.559
GPT-4-turbo	0.668	0.936	0.860
GPT-4	0.636	0.936	0.896
Gemini 1.0 Pro	0.672	0.924	0.721

Overview of Results

Gemini, GPT-4, and GPT-4-turbo show similar results on SemEval and FLUTE.

Scaling we expect between GPT-3.5 and the other models.

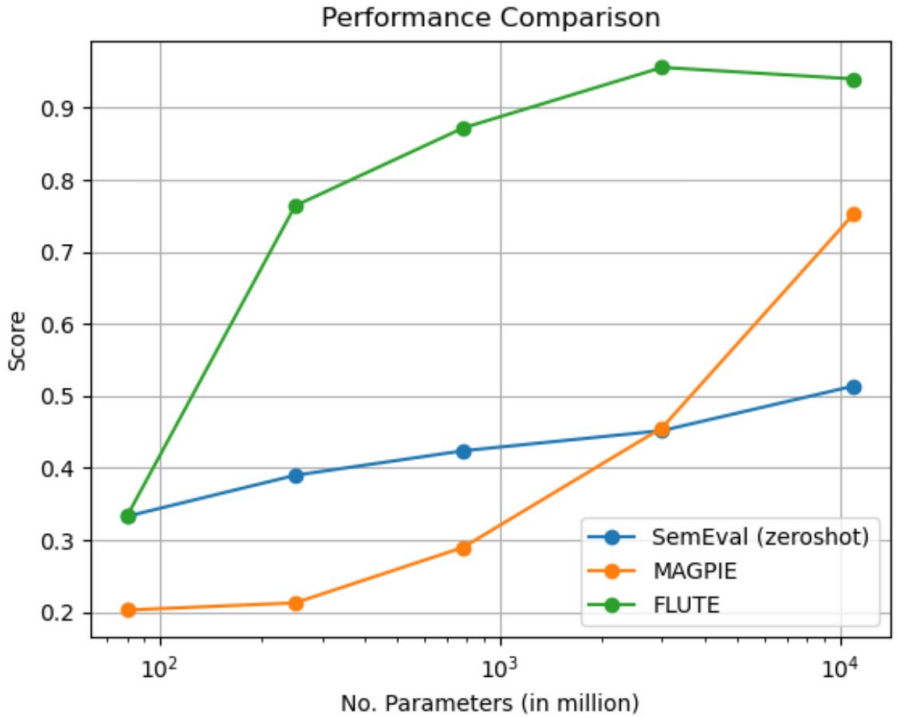
	SemEval	FLUTE	MAGPIE
GPT-4	0.636	0.936	0.896
Phi-2	0.447	0.458	0.531
Llama2 (7B-chat)	0.479	0.373	0.314
Llama2 (13B-chat)	0.505	0.602	0.483
CapybaraHermes-2.5-Mistral-7B	0.539	0.812	0.587

Scaling

We also see the scaling we expect with the FLAN-T5 models

	SemEval	FLUTE	MAGPIE
GPT-4	0.636	0.936	0.896
Flan-T5-Small	0.333	0.333	0.203
Flan-T5-Base	0.390	0.764	0.213
Flan-T5-Large	0.424	0.872	0.290
Flan-T5-XL	0.452	0.956	0.456
Flan-T5-XXL (11.3B)	0.514	0.940	0.753

Scaling


Performance Comparison
Score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
10^2
10^3
10^4
No. Parameters (in million)
SemEval (zeroshot)
MAGPIE
FLUTE

Prompting

Experiment in different prompting styles using GPT-3.5 on the English split of SemEval 2022 Task 2a.

[Base] Only return one letter (i or l).

Prompting

Experiment in different prompting styles using GPT-3.5 on the English split of SemEval 2022 Task 2a.

[Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or l).

Prompting

Experiment in different prompting styles using GPT-3.5 on the English split of SemEval 2022 Task 2a.

[Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or c).

Prompting

Experiment in different prompting styles using GPT-3.5 on the English split of SemEval 2022 Task 2a.

[Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or c).

You are an expert in idiomatic language. [Base] Only return one letter (i or l).

Prompting

Experiment in different prompting styles using GPT-3.5 on the English split of SemEval 2022 Task 2a.

[Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or l).

You are an expert in language use. [Base] Only return one letter (i or c).

You are an expert in idiomatic language. [Base] Only return one letter (i or l).

You are an expert in idiomatic language. [Base]

Prompting

Experiment in different prompting styles using GPT-3.5 on the English split of SemEval 2022 Task 2a.

	EN
Default	0.739
“Expert in language use”	0.635
“Expert in language use” + Idiomatic vs. Compositional	0.717
“Expert in Idiomatic Language”	0.538
No “Only return one letter (i or l).”	0.633

Language Prompting

Experimented with different ways to prompt the models of the language used in the example.

> Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally.

Language Prompting

Experimented with different ways to prompt the models of the language used in the example.

> Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally.

You will be given a sentence in Portuguese. Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally.

Language Prompting

Experimented with different ways to prompt the models of the language used in the example.

> Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally.
>
> You will be given a sentence in Portuguese. Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally.
>
> Desambigua se a expressão dada é usada idiomaticamente ou literalmente no contexto determinado, retornando 'i' se a expressão estiver sendo usada idiomaticamente ou 'l' se literalmente.

Language Prompting

Experimented with different ways to prompt the models of the language used in the example.

	GPT-3.5-turbo		Gemini 1.0		Flan-T5-XXL	
	PT	GL	PT	GL	PT	GL
Default	0.553	0.587	0.582	0.604	0.464	0.411
Language Prompt	0.554	0.604	0.561	0.640	0.479	0.457
Translated	0.541	0.512	0.549	0.665	0.573	0.477

Few Shot Prompting

We experiment also with adding one and few shot prompts to the models.

Disambiguate whether the given expression is used idiomatically or literally in the given context, returning 'i' if the expression is being used idiomatically or 'l' if literally.

For example, the expression {MWE} is used {label} in the sentence "{target}", so you would return '{label}'.

Expression: {}. Context: {}. Only return one letter (i or l).Return i if the expression is used idiomatically or l if it is literal.

Few Shot Prompting

We experiment also with adding one and few shot prompts to the models.

Model	Setting	EN	PT	GL	All
Gemini Pro 1.0	Zero-shot	**0.766**	0.590	0.600	0.672
	One-shot	0.706	0.625	0.711	0.688
	Few-shot	0.685	**0.642**	**0.745**	**0.693**
GPT-3.5-turbo	Zero-shot	**0.739**	**0.563**	**0.579**	**0.645**
	One-shot	0.645	0.542	0.553	0.594
	Few-shot	0.686	0.545	0.566	0.614
Flan-T5-XXL	Zeroshot	0.629	0.464	0.411	0.514
	Oneshot	0.810	0.665	0.732	0.749
	Fewshot	**0.845**	**0.713**	**0.828**	**0.805**
Best	Zero-shot	0.964	0.894	0.937	0.939

University of Sheffield | Healthy Lifespan Institute

Discussion

Practicalities

A number of practicalities of using SAAS LLMs were encountered whilst running these experiments:

- Evaluation cost - almost $20 to run GPT-4 on MAGPIE

Practicalities

A number of practicalities of using SAAS LLMs were encountered whilst running these experiments:

- Evaluation cost - almost $20 to run GPT-4 on MAGPIE
- Rate limits - not possible to run GPT-4 on MAGPIE in one go at the time of the experiments

Practicalities

A number of practicalities of using SAAS LLMs were encountered whilst running these experiments:

- Evaluation cost - almost $20 to run GPT-4 on MAGPIE
- Rate limits - not possible to run GPT-4 on MAGPIE in one go at the time of the experiments
- Safety features - can be sensitive to some topics and reject examples

Practicalities

A number of practicalities of using SAAS LLMs were encountered whilst running these experiments:

- Evaluation cost - almost $20 to run GPT-4 on MAGPIE
- Rate limits - not possible to run GPT-4 on MAGPIE in one go at the time of the experiments
- Safety features - can be sensitive to some topics and reject examples
- Service changes - regular updates to the models change results (for better and worse)

Conclusion

Overall, we've shown that:

- Closed LLMs perform well on idiom classification datasets
- Smaller local models have promising performance, but not on par with larger ones
- Fine-tuned encoder-only models are still SOTA
- Choice of prompts can highly impact results

Future Work

Our results are not comprehensive:

- Many more models and prompting styles to investigate
- Fine-tuning of the models may lead to better results and close the gaps to fine-tuned encoder models
- Other tasks that can be used to evaluate idiomatic language understanding

University of Sheffield | Healthy Lifespan Institute

Thank you!

Any questions:
drsphelps1@sheffield.ac.uk