Contextual Language Models: Using Spatio-temporal and Social Context to improve Language Models

Shubhanshu Mishra

(work done during PhD at UIUC and with collaborators at Twitter)

All views are my own and do not represent the views of my past and current employers.

Key Papers

- Jinning Li, Shubhanshu Mishra, Ahmed El-Kishky, Sneha Mehta, and Vivek Kulkarni. 2022. NTULM: Enriching Social Media Text Representations with Non-Textual Units. In Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), pages 69–82, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An Approach for Socially Sensitive Pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mishra, Shubhanshu. 2020. “Information Extraction from Digital Social Trace Data with Applications to Social Media and Scholarly Communication Data.” University of Illinois at Urbana-Champaign. https://shubhanshu.com/phd_thesis/
- Mishra, S., & Diesner, J. (2018, July 3). Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. Proceedings of the 29th on Hypertext and Social Media. HT ’18: 29th ACM Conference on Hypertext and Social Media. https://doi.org/10.1145/3209542.3209562

Language Modeling and Representation Learning

Text = [The, cat, sat, on, the, mat]

Goal:

f(text) = P(text) ← ***Language Modeling Objective***

or

f(text) = score(text) s.t.

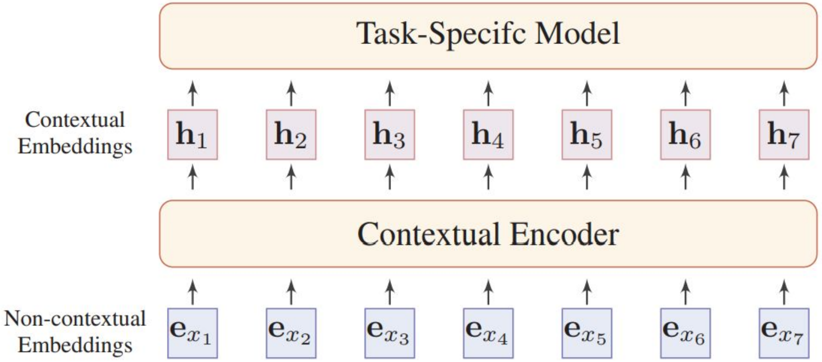
score(good text) > score(bad text) ← ***Contrastive Objective***

Hypothesis:

If we can learn a good ***f(text)***, we can use it to improve our performance on downstream tasks ***g(f(text)) = y*** or to generate or complete text.

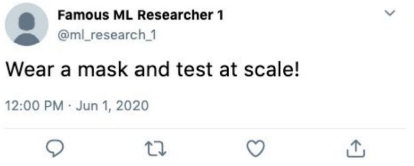
Application of language models

Content Representation


Task-Specifc Model
Contextual Embeddings
$\mathbf{h}_1$ $\mathbf{h}_2$ $\mathbf{h}_3$ $\mathbf{h}_4$ $\mathbf{h}_5$ $\mathbf{h}_6$ $\mathbf{h}_7$
Contextual Encoder
Non-contextual Embeddings
$\mathbf{e}_{x_1}$ $\mathbf{e}_{x_2}$ $\mathbf{e}_{x_3}$ $\mathbf{e}_{x_4}$ $\mathbf{e}_{x_5}$ $\mathbf{e}_{x_6}$ $\mathbf{e}_{x_7}$

Content Recommendation

Query Tweet

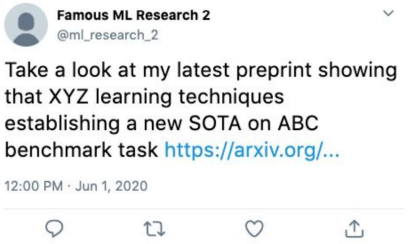


Famous ML Researcher 1
@ml_research_1

Wear a mask and test at scale!

12:00 PM · Jun 1, 2020

Collaborative Filtering Candidate

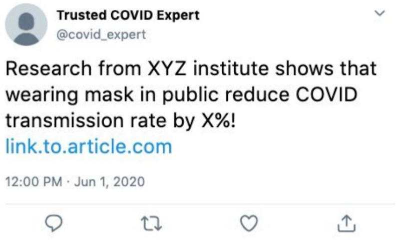


Famous ML Research 2
@ml_research_2

Take a look at my latest preprint showing that XYZ learning techniques establishing a new SOTA on ABC benchmark task https://arxiv.org/...

12:00 PM · Jun 1, 2020

Content-Based Candidate

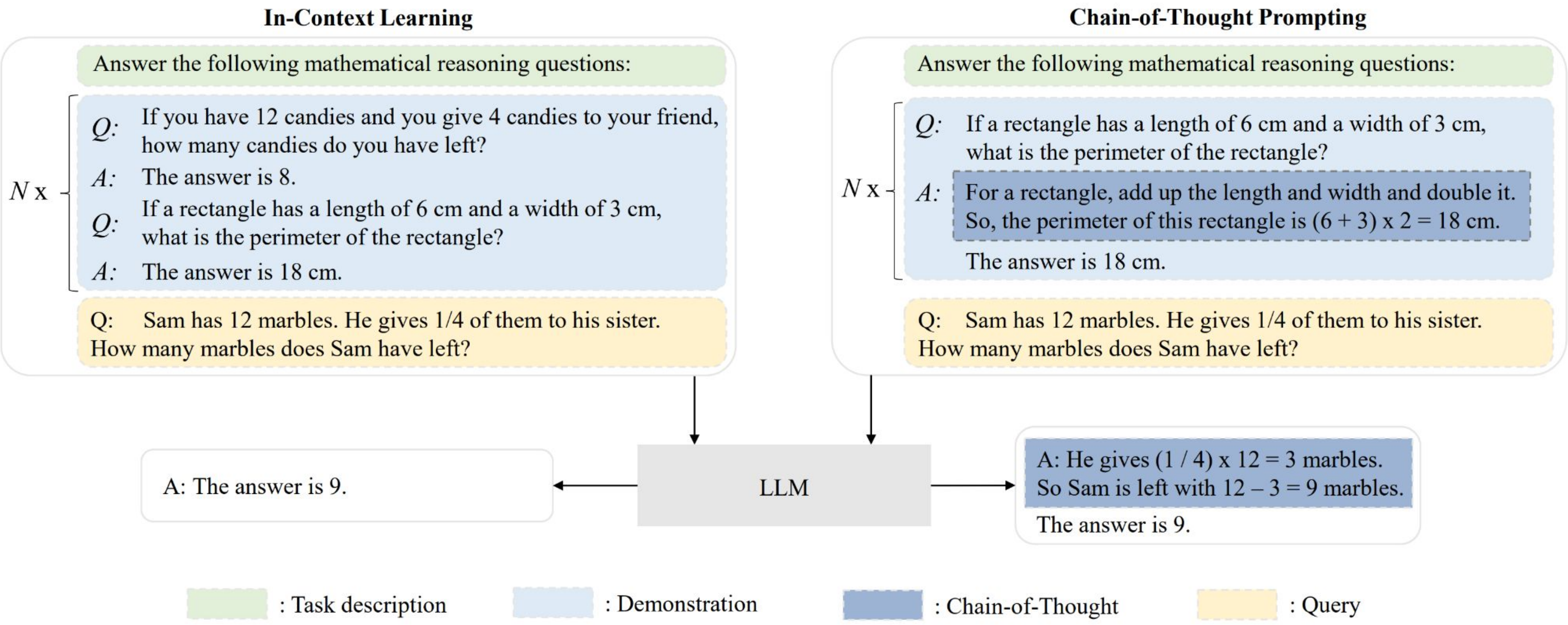


Trusted COVID Expert
@covid_expert

Research from XYZ institute shows that wearing mask in public reduce COVID transmission rate by X%! link.to.article.com

12:00 PM · Jun 1, 2020

Application of generative language models



In-Context Learning

Answer the following mathematical reasoning questions:

Q: If you have 12 candies and you give 4 candies to your friend, how many candies do you have left?

A: The answer is 8.

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: The answer is 18 cm.

Q: Sam has 12 marbles. He gives 1/4 of them to his sister. How many marbles does Sam have left?

A: The answer is 9.

Chain-of-Thought Prompting

Answer the following mathematical reasoning questions:

Q: If a rectangle has a length of 6 cm and a width of 3 cm, what is the perimeter of the rectangle?

A: For a rectangle, add up the length and width and double it. So, the perimeter of this rectangle is (6 + 3) x 2 = 18 cm.

The answer is 18 cm.

Q: Sam has 12 marbles. He gives 1/4 of them to his sister. How many marbles does Sam have left?

A: He gives (1 / 4) x 12 = 3 marbles. So Sam is left with 12 – 3 = 9 marbles.

The answer is 9.

Source: [2303.18223] A Survey of Large Language Models

Language Modeling - Old to New

n-gram

Word2Vec

RNN

LSTM

ELMo

BERT

GPT

T5

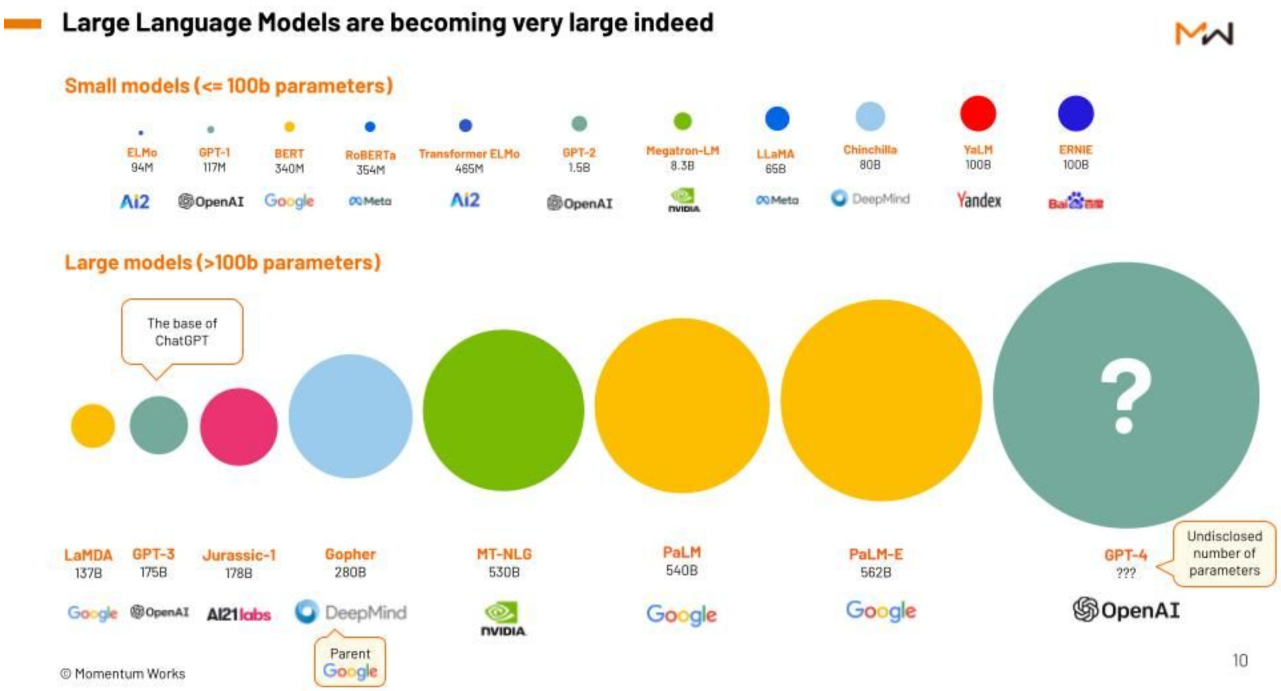
ChatGPT

LLaMA

RWKV-LM

Good survey on progress of LMs: https://lilianweng.github.io/posts/2019-01-31-lm/
Good link on details of language models: https://lena-voita.github.io/nlp_course/language_modeling.html

State of Language Modeling



Large Language Models are becoming very large indeed

Small models (<= 100b parameters)

ELMo 94M
GPT-1 117M
BERT 340M
RoBERTa 354M
Transformer ELMo 465M
GPT-2 1.5B
Megatron-LM 8.3B
LLaMA 65B
Chinchilla 80B
YaLM 100B
ERNIE 100B

Large models (>100b parameters)

The base of ChatGPT

LaMDA 137B
GPT-3 175B
Jurassic-1 178B
Gopher 280B
MT-NLG 530B
PaLM 540B
PaLM-E 562B
GPT-4 ???

Undisclosed number of parameters

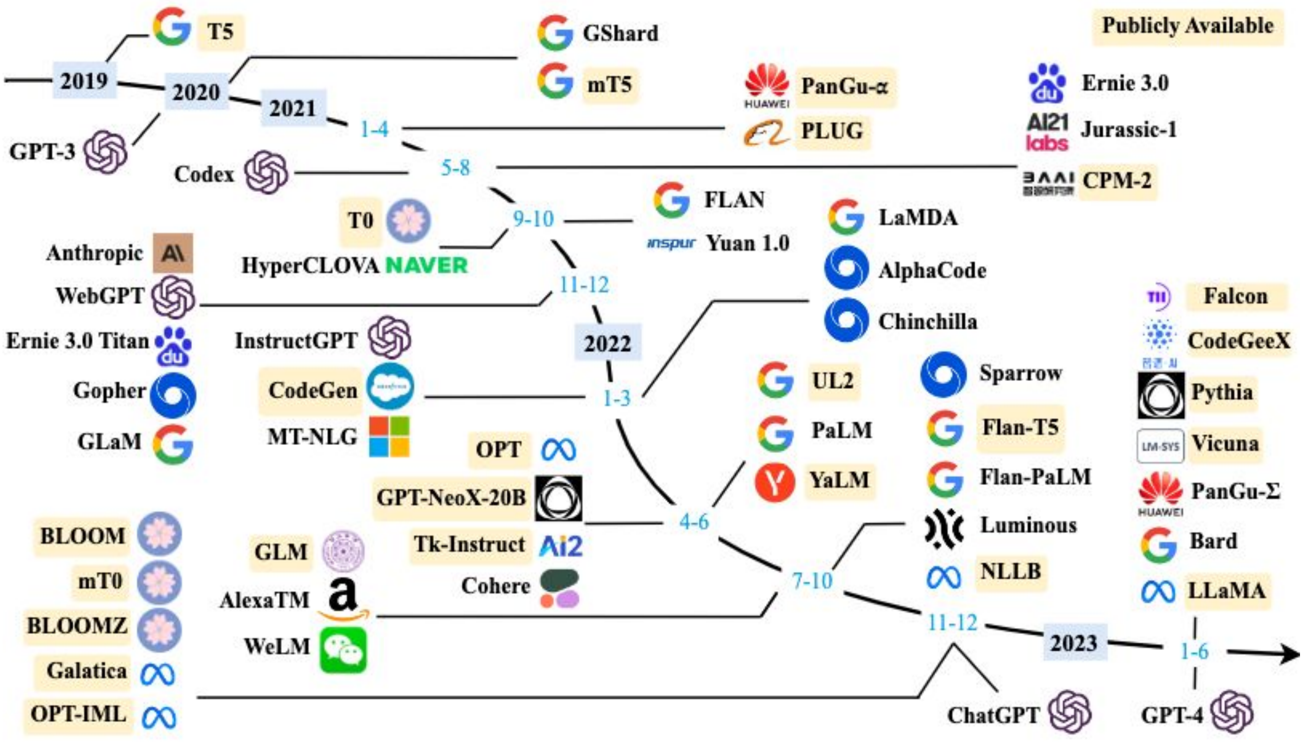
Parent Google

© Momentum Works

10

Source: https://momentum.asia/product/the-future-by-chatgpt/

Recent history of Language Models



Source: https://github.com/RUCAIBox/LLMSurvey

Scaling Laws

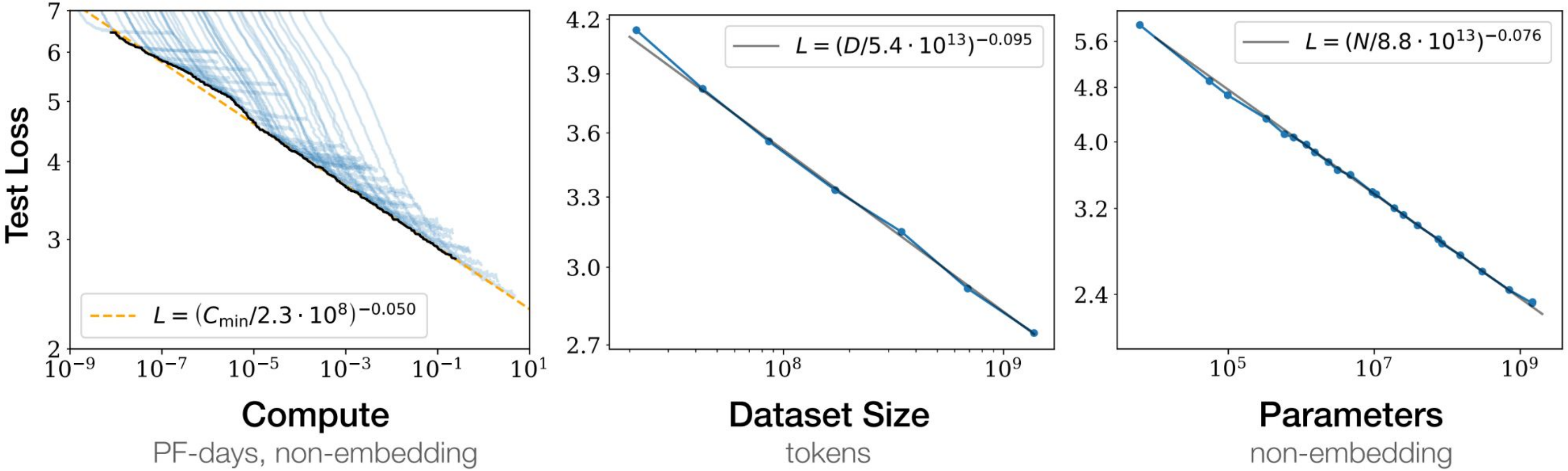
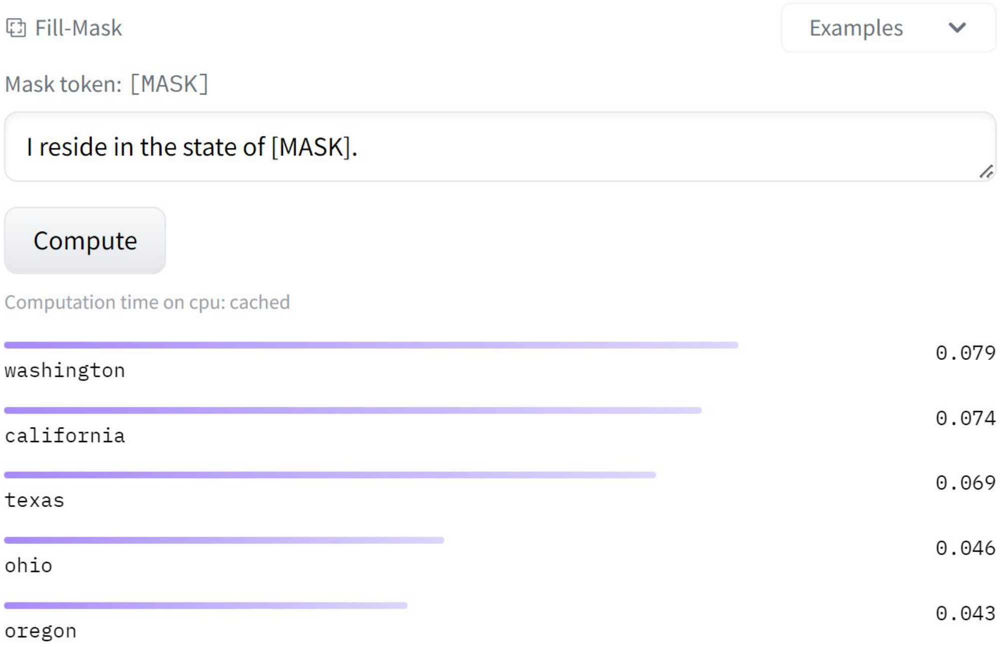


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Source: https://arxiv.org/abs/2001.08361

Need for Context

Real world data is contextual



Fill-Mask

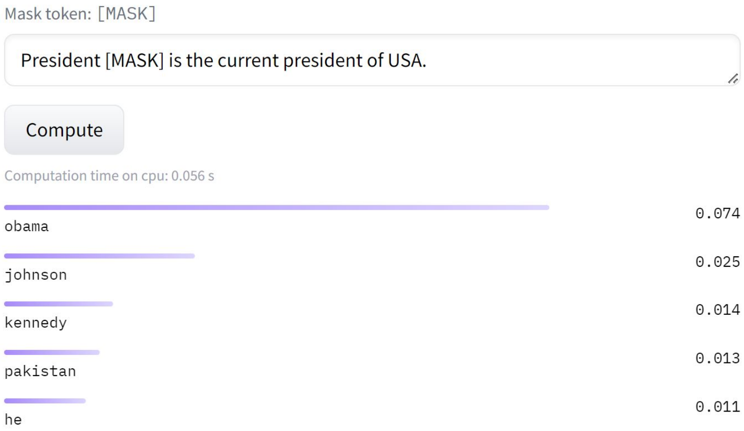
Mask token: [MASK]

I reside in the state of [MASK].

Computation time on cpu: cached

washington	0.079
california	0.074
texas	0.069
ohio	0.046
oregon	0.043

Requires location context



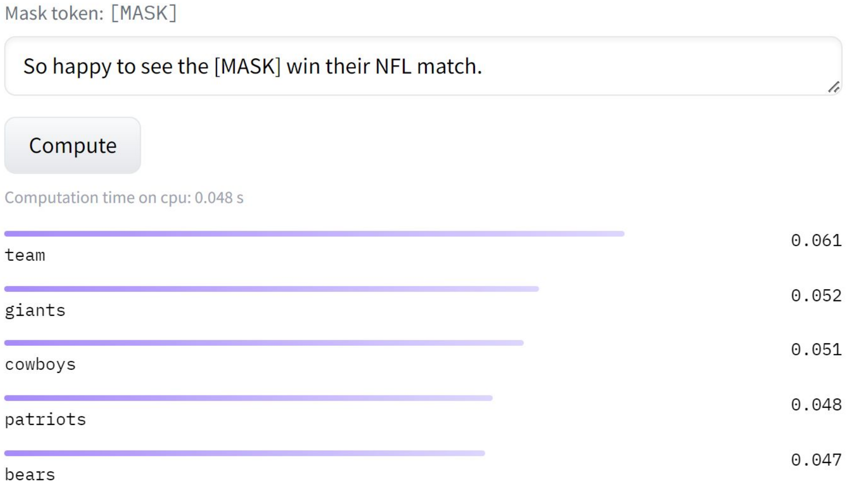
Mask token: [MASK]

President [MASK] is the current president of USA.

Computation time on cpu: 0.056 s

obama	0.074
johnson	0.025
kennedy	0.014
pakistan	0.013
he	0.011

Requires temporal context



Mask token: [MASK]

So happy to see the [MASK] win their NFL match.

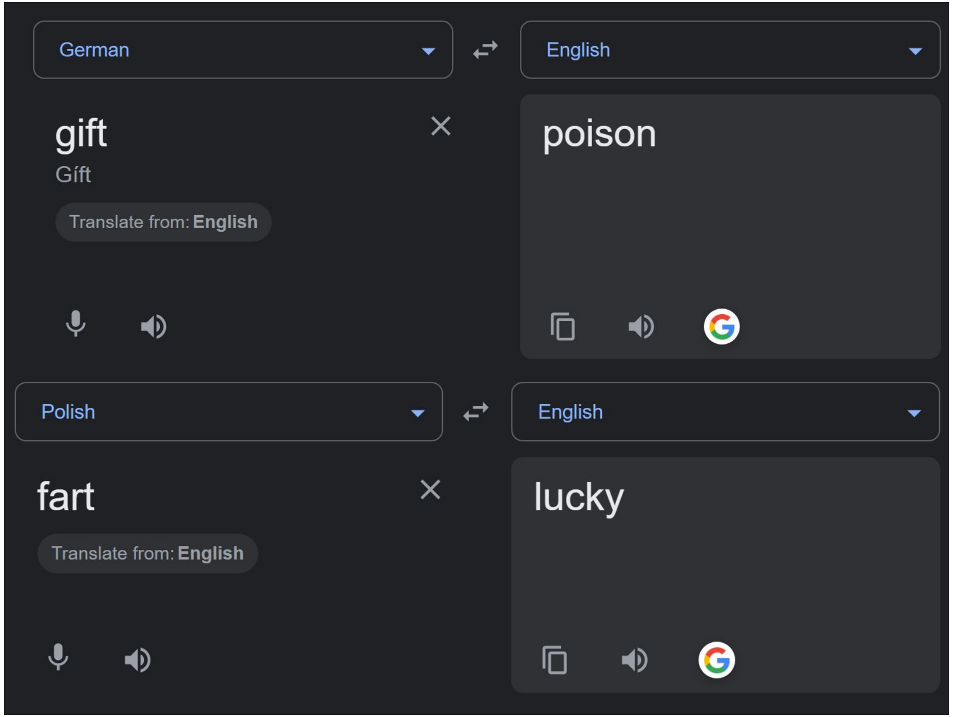
Computation time on cpu: 0.048 s

team	0.061
giants	0.052
cowboys	0.051
patriots	0.048
bears	0.047

Requires social context

Need for context in Translation

When translating a language it is also important to learn which words from one language map to which words in the other language to avoid embarrassing errors like the following:



German

English

gift

Gift

Translate from: English

poison

Polish

English

fart

Translate from: English

lucky

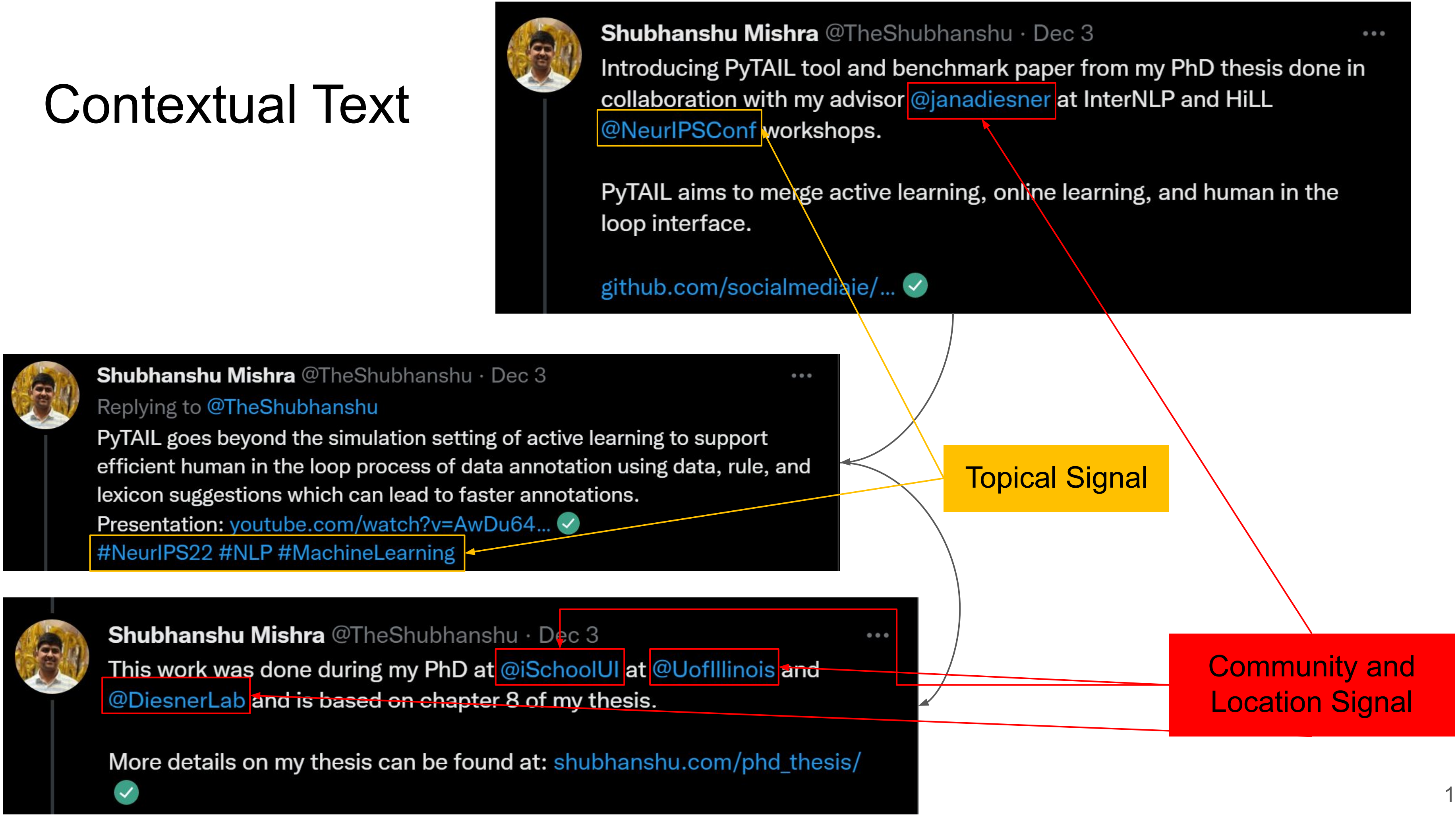
Me (visiting a german friend): I brought this drink as a gift for you

Guests who know only german look in awe.

A polish person (learning English) visits an English friend: I hope this day brings you lots of fart

English friend: Ewww

Contextual Text



Shubhanshu Mishra @TheShubhanshu · Dec 3

Introducing PyTAIL tool and benchmark paper from my PhD thesis done in collaboration with my advisor @janadiesner at InterNLP and HiLL @NeurIPSConf workshops.

PyTAIL aims to merge active learning, online learning, and human in the loop interface.

github.com/socialmediaie/... ✓

Shubhanshu Mishra @TheShubhanshu · Dec 3

Replying to @TheShubhanshu

PyTAIL goes beyond the simulation setting of active learning to support efficient human in the loop process of data annotation using data, rule, and lexicon suggestions which can lead to faster annotations.

Presentation: youtube.com/watch?v=AwDu64... ✓

#NeurIPS22 #NLP #MachineLearning

Shubhanshu Mishra @TheShubhanshu · Dec 3

This work was done during my PhD at @iSchoolUI at @UofIllinois and @DiesnerLab and is based on chapter 8 of my thesis.

More details on my thesis can be found at: shubhanshu.com/phd_thesis/ ✓

Topical Signal

Community and Location Signal

Challenges of encoding context via Text based prompts

1. You need a good way to represent context via text based prompt which is informative.
    a. Requires trial and error.
    b. Difficult for images, audio, video. You can use signal to text and then add it as prompt. [1]
2. Inappropriate handling of rare words because of tokenization issues
    a. URL, user handles, rare brand names will lead to token based splitting and the context is lost.
3. Text is often the wrong abstraction for context which is embedded in a graph.
    a. A user is known more by their engagement signal than by their name, description.
    b. Naive Incontext Learning based on text will not help here.
4. Transformers have O(N^2) training and inference cost so using long text based context is more costly than using a single embedding based context.
5. Using Context allows us to infuse domain knowledge.

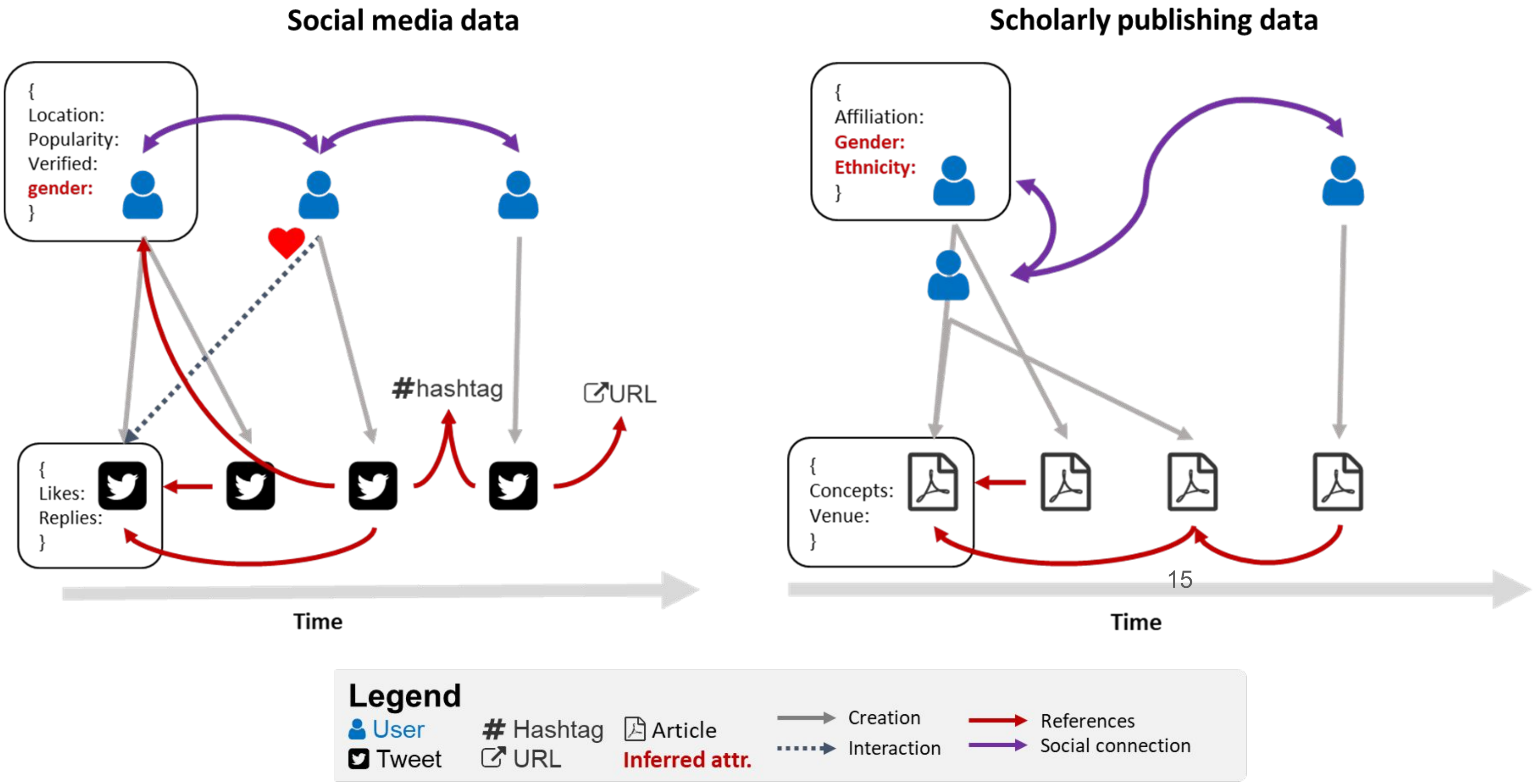
Digital Social Trace Data https://shubhanshu.com/phd_thesis/

A representation for real world contextual text data

Digital Social Trace Data (DSTD) are digital activity traces generated by individuals as part of a social interactions, such as interactions on social media websites like Twitter, Facebook; or in scientific publications.

Can help with building better Content Models for Web Scale Content

Digital Social Trace Data (DSTD)



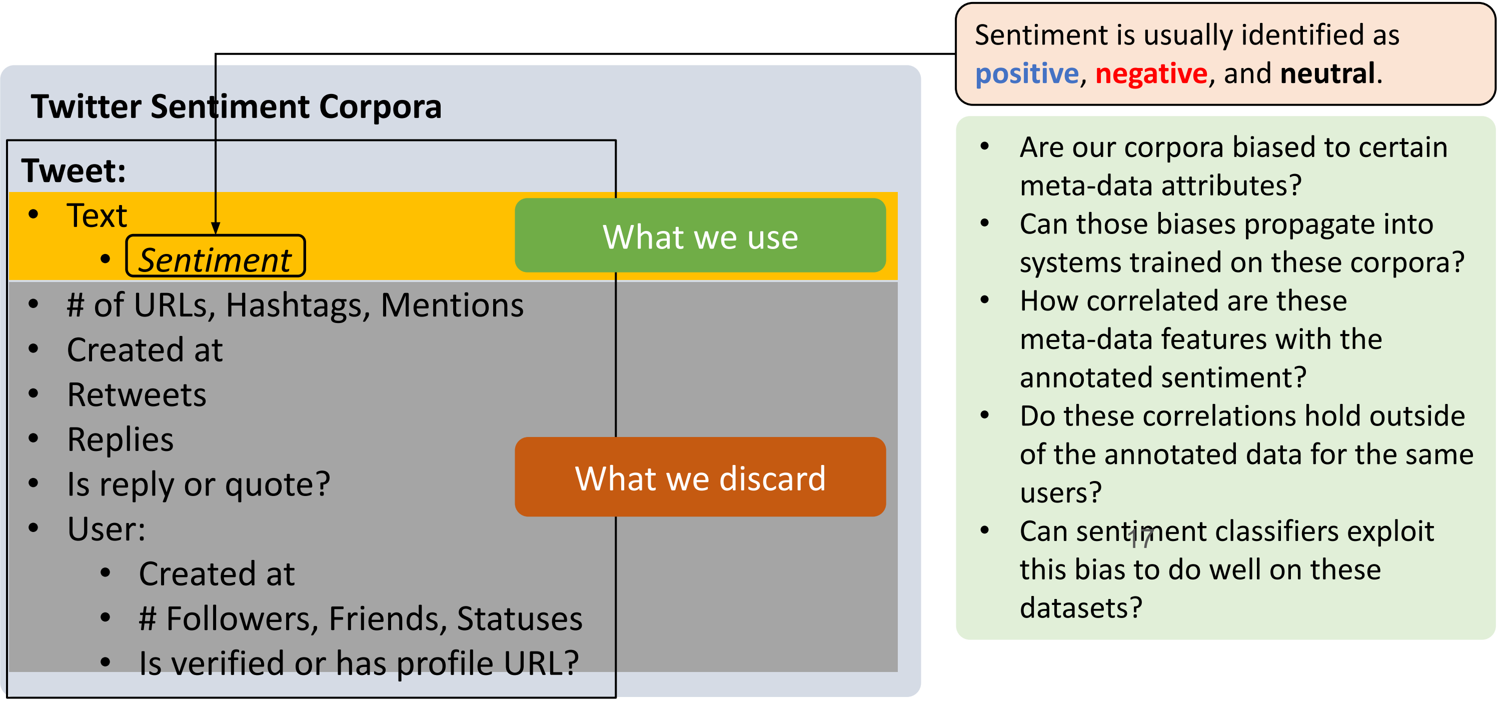
Social media data

Scholarly publishing data

https://shubhanshu.com/phd_thesis

Most Language Models are missing this holistic context around the text

Improving sentiment classification using user and tweet metadata



Sentiment is usually identified as **positive**, **negative**, and **neutral**.

Twitter Sentiment Corpora

Tweet:

- Text
  - *Sentiment*
- # of URLs, Hashtags, Mentions
- Created at
- Retweets
- Replies
- Is reply or quote?
- User:
  - Created at
  - # Followers, Friends, Statuses
  - Is verified or has profile URL?

What we use

What we discard

- Are our corpora biased to certain meta-data attributes?
- Can those biases propagate into systems trained on these corpora?
- How correlated are these meta-data features with the annotated sentiment?
- Do these correlations hold outside of the annotated data for the same users?
- Can sentiment classifiers exploit this bias to do well on these datasets?

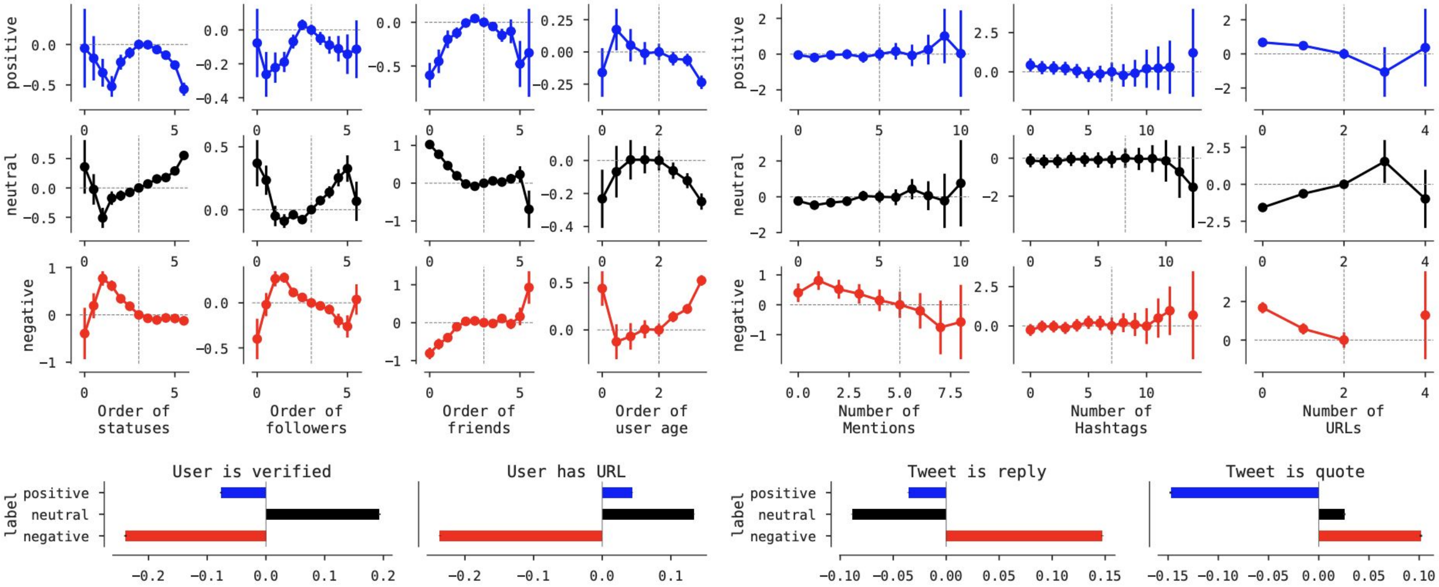
Mishra, S., & Diesner, J. (2018, July 3). Detecting the Correlation between Sentiment and User-level as well as Text-Level Meta-data from Benchmark Corpora. Proceedings of the 29th on Hypertext and Social Media. HT ’18: 29th ACM Conference on Hypertext and Social Media. https://doi.org/10.1145/3209542.3209562

Types of metadata and what they quantify

Quantification	**User metadata**
Activity level	# Statuses
Social Interest of the user	# Friends
Social status	# Followers
Account age	# days since account creation to posted tweet
Profile authenticity	Presence of URL on the profile or if the profile is verified

Quantification	**Tweet metadata**
Topical variety	# hashtags
Reference to sources	# URLs
Reference to network	# user mentions
Part of conversation	Is reply
Reference to conversation	Is quote

User metadata v/s Sentiment



(a) User-level meta-data **(b) Tweet-level meta-data**

Figure 3: Meta-data features vs. sentiment classes. Y-axis in top plots and X-axis in bottom plots, is log-odds ratio, with respect to point at dashed lines.

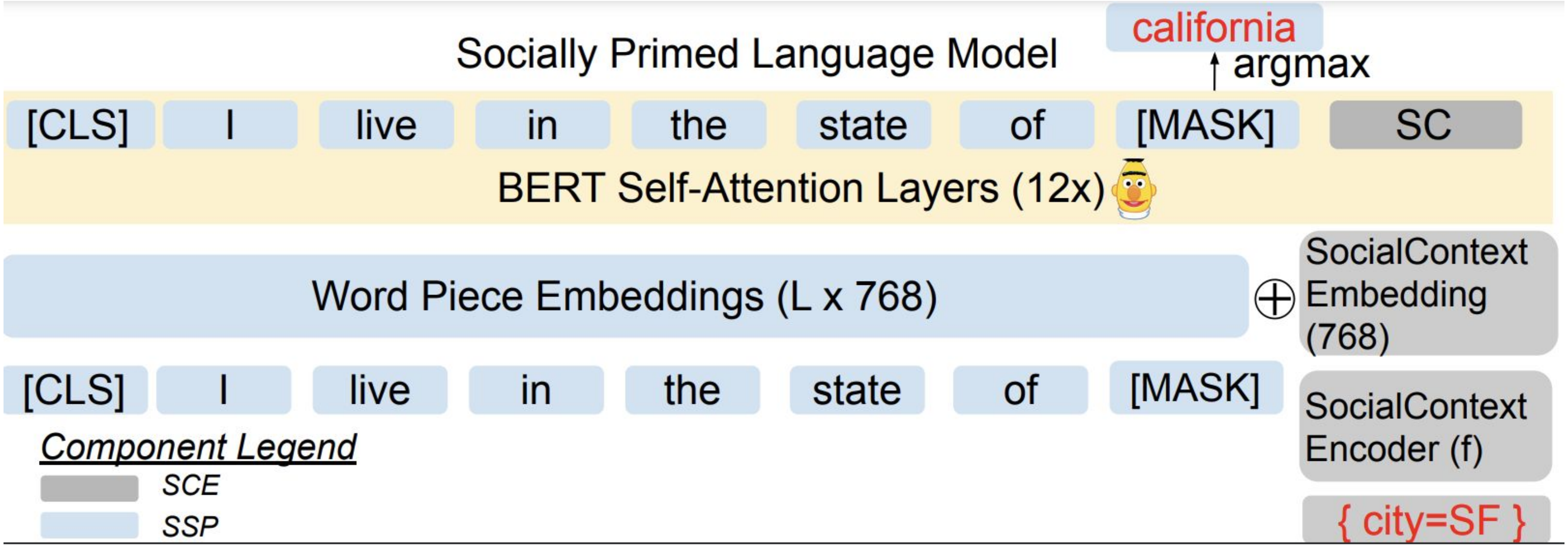
Using metadata features can improve sentiment classification

Dataset	Model	Acc.	P	R	F1	KLD
Airline	meta	63.9	61.1	36.8	32.8	0.663
	text	80.0	78.3	69.0	72.4	0.026
	joint	80.3	76.6	72.0	**74.0**	0.005
Clarin	meta	45.7	42.1	40.9	37.8	0.238
	text	64.1	64.5	62.2	62.9	0.012
	joint	64.1	64.0	63.0	**63.4**	0.000
GOP	meta	59.9	54.3	37.5	33.6	0.776
	text	66.4	63.7	51.4	53.6	0.111
	joint	65.6	59.9	56.5	**57.8**	0.006
Healthcare	meta	56.7	36.8	39.4	35.1	0.717
	text	64.2	71.3	49.5	51.0	0.233
	joint	65.6	61.6	58.3	**59.5**	0.007
Obama	meta	39.3	37.0	35.1	32.0	0.282
	text	61.5	64.8	59.7	60.9	0.030
	joint	62.3	63.2	61.6	**62.2**	0.002
SemEval	meta	47.0	31.0	36.2	33.0	0.845
	text	65.5	64.1	58.0	59.5	0.032
	joint	65.6	62.7	60.5	**61.4**	0.001

Boost in F1 is mostly due to better recall. Precision is lower.

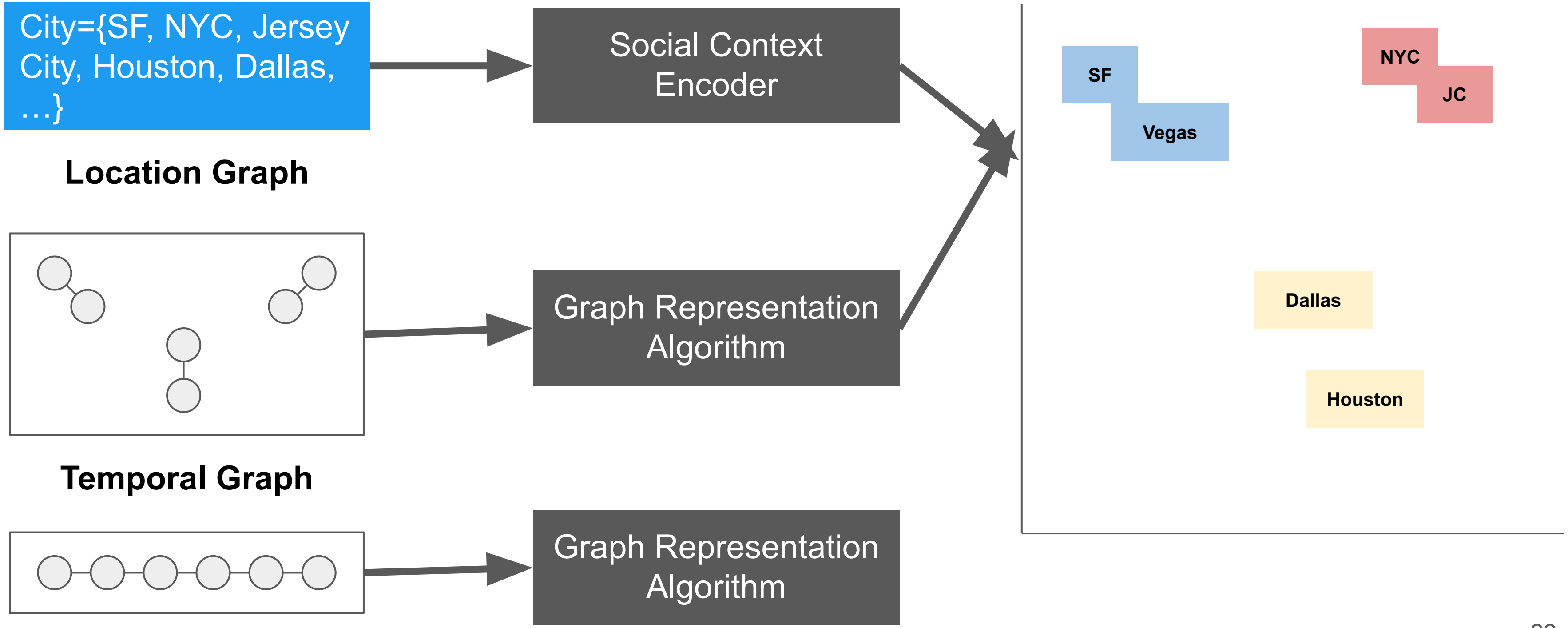
MESC might be helping with tweets with high OOV rates, where text classifiers don't do well.

LMSOC: An Approach for Socially Sensitive Pretraining


Socially Primed Language Model
california
argmax
[CLS]
I
live
in
the
state
of
[MASK]
SC
BERT Self-Attention Layers (12x)
Word Piece Embeddings (L x 768)
⊕
SocialContext Embedding (768)
[CLS]
I
live
in
the
state
of
[MASK]
SocialContext Encoder (f)
{ city=SF }
Component Legend
SCE
SSP

Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An Approach for Socially Sensitive Pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics

Social Context Encoder



City={SF, NYC, Jersey City, Houston, Dallas, …}

Social Context Encoder

Location Graph

Graph Representation Algorithm

Temporal Graph

Graph Representation Algorithm

Evaluating LMSOC

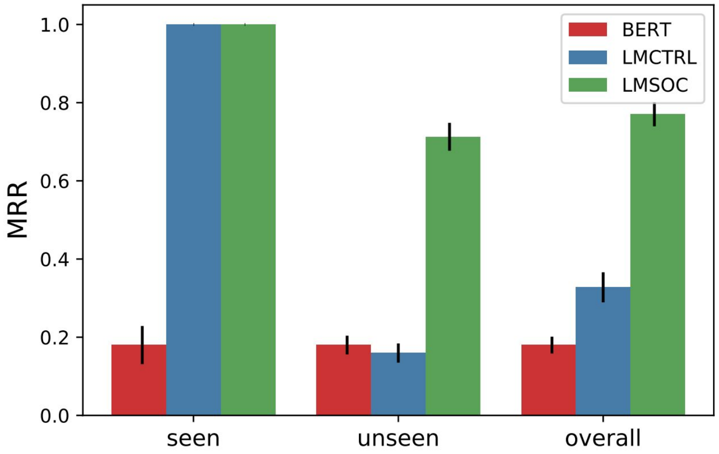


Figure 2: Performance of models on the synthetic data set as measured in terms of mean reciprocal rank (MRR, higher is better). See Section 3.1 for details.

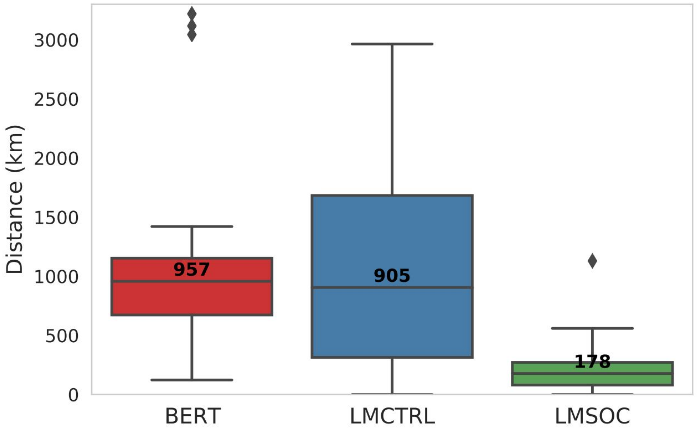
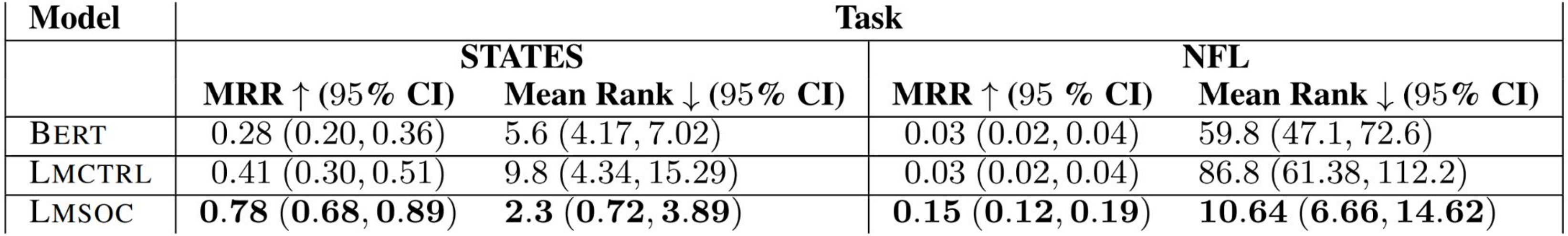


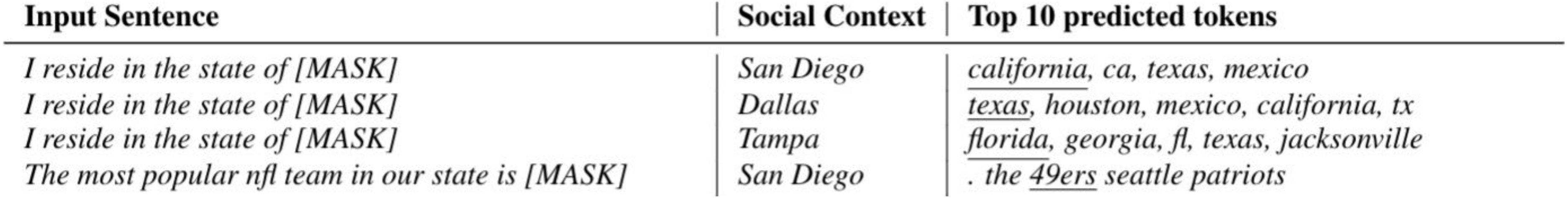
Figure 3: Descriptive statistics of the distances of the top cities from the input city predicted by various models on the **CLOSECITY** task (lower is better). See Section 3.2.2 for details.

Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An Approach for Socially Sensitive Pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics

Evaluating LMSOC



Model	**Task**			
	STATES		**NFL**	
	MRR ↑ (95% CI)	**Mean Rank ↓ (95% CI)**	**MRR ↑ (95 % CI)**	**Mean Rank ↓ (95% CI)**
BERT	0.28 (0.20, 0.36)	5.6 (4.17, 7.02)	0.03 (0.02, 0.04)	59.8 (47.1, 72.6)
LMCTRL	0.41 (0.30, 0.51)	9.8 (4.34, 15.29)	0.03 (0.02, 0.04)	86.8 (61.38, 112.2)
LMSOC	**0.78 (0.68, 0.89)**	**2.3 (0.72, 3.89)**	**0.15 (0.12, 0.19)**	**10.64 (6.66, 14.62)**



Input Sentence	**Social Context**	**Top 10 predicted tokens**
I reside in the state of [MASK]	*San Diego*	*california, ca, texas, mexico*
I reside in the state of [MASK]	*Dallas*	*texas, houston, mexico, california, tx*
I reside in the state of [MASK]	*Tampa*	*florida, georgia, fl, texas, jacksonville*
The most popular nfl team in our state is [MASK]	*San Diego*	*. the 49ers seattle patriots*

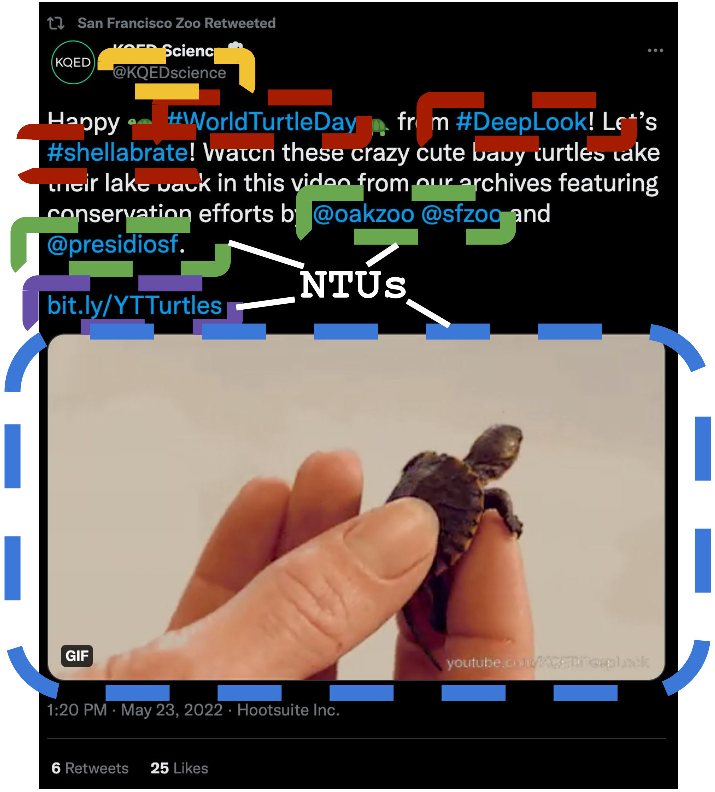
Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An Approach for Socially Sensitive Pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics

Bringing it all together

NTULM: Enriching Social Media Text Representations with Non-Textual Units

Motivation: Non-Textual Units

Non-Textual Units (NTUs) are the social contexts which appear alongside a social media post, e.g. *Hashtag*, *URL*, *author*, *user mentions* and *media*



San Francisco Zoo Retweeted

KQED Science

@KQEDscience

Happy 🐢 #WorldTurtleDay 🐢 from #DeepLook! Let's #shellabrate! Watch these crazy cute baby turtles take their lake back in this video from our archives featuring conservation efforts by @oakzoo @sfzoo and @presidiosf.

bit.ly/YTTurtles

GIF

youtube.com/KQEDDeepLook

1:20 PM · May 23, 2022 · Hootsuite Inc.

6 Retweets 25 Likes

https://twitter.com/KQEDscience/status/1528833154492026880

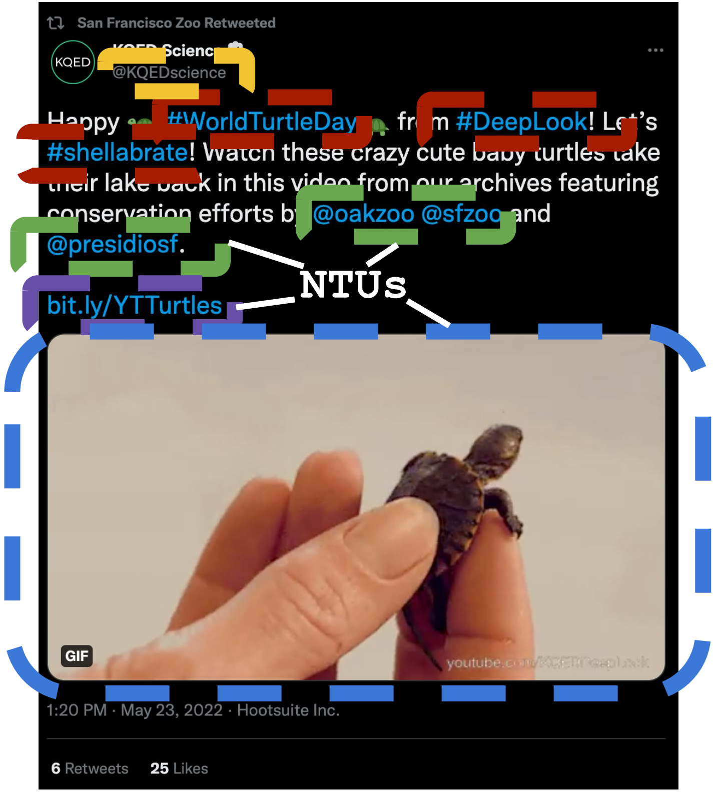
Challenge: Existing models and NTUs

NTUs embedded in the text are broken up by tokenizers diminishing their signal.

Non embedded NTUs are not included.

NTUs have a global context outside of the text.

```
[happy, [UNK], #, world, ##tur, ##tled, ##ay, [UNK],
from, #, deep, ##lo, ##ok, !, let, , s, #, shell,
##ab, ##rate, !, watch, these, crazy, cute, baby,
turtles, take, their, lake, back, in, this, video,
from, our, archives, featuring, conservation, efforts,
by, @, oak, ##zoo, @, sf, ##zoo, and, @, pre, ##si,
##dio, ##sf, ., http, :, /, /, bit, ., l, ##y, /, y,
##tt, ##urt, ##les]
(Result from tokenizer of bert-base-uncased)
```



San Francisco Zoo Retweeted

KQED Science
@KQEDscience

Happy 🐢 #WorldTurtleDay 🐢 from #DeepLook! Let's #shellabrate! Watch these crazy cute baby turtles take their lake back in this video from our archives featuring conservation efforts by @oakzoo @sfzoo and @presidiosf.

bit.ly/YTTurtles

1:20 PM · May 23, 2022 · Hootsuite Inc.

6 Retweets 25 Likes

https://twitter.com/KQEDscience/status/1528833154492026880

Intuition: Our approach for Non-Textual Units

Inject average NTU embeddings into the Transformer alongside token embeddings.

Pre-compute NTU embeddings using heterogeneous networks, e.g. social engagements for users and Hashtags

```
[happy, [UNK], #, world, ##tur, ##tled, ##ay, [UNK], from, #,
deep, ##lo, ##ok, !, let, , s, #, shell, ##ab, ##rate, !, watch,
these, crazy, cute, baby, turtles, take, their, lake, back, in,
this, video, from, our, archives, featuring, conservation,
efforts, by, @, oak, ##zoo, @, sf, ##zoo, and, @, pre, ##si,
##dio, ##sf, ., http, :, /, /, bit, ., l, ##y, /, y, ##tt, ##urt,
##les] + [@KQEDscience, #WorldTurtleDay, #DeepLook, #shellabrate,
@oakzoo, @sfzoo, @presidiosf, bit.ly/YTTurtles, Media 1]
```

NTULM Framework

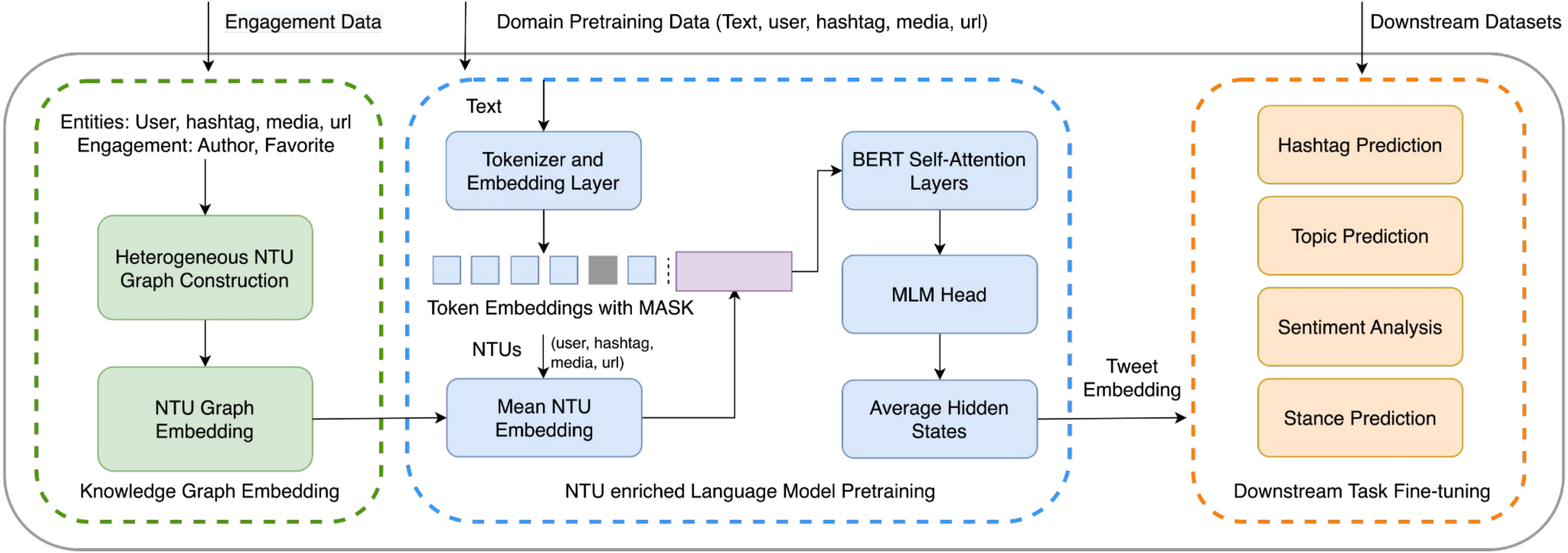
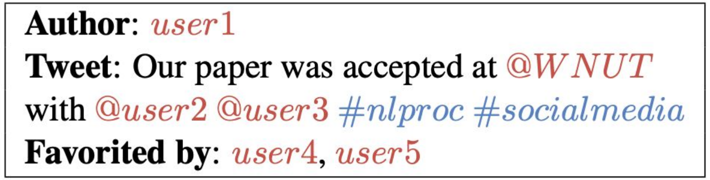

Engagement Data
Domain Pretraining Data (Text, user, hashtag, media, url)
Downstream Datasets
Entities: User, hashtag, media, url
Engagement: Author, Favorite
Heterogeneous NTU Graph Construction
NTU Graph Embedding
Knowledge Graph Embedding
Text
Tokenizer and Embedding Layer
Token Embeddings with MASK
NTUs
(user, hashtag, media, url)
Mean NTU Embedding
BERT Self-Attention Layers
MLM Head
Average Hidden States
NTU enriched Language Model Pretraining
Tweet Embedding
Hashtag Prediction
Topic Prediction
Sentiment Analysis
Stance Prediction
Downstream Task Fine-tuning

Fig 1: Framework of NTULM

Knowledge Graph Embedding

- **Graph nodes**: author, Hashtag
- **Graph edges**: connect user-Hashtag if user authors, favorites, or is co-mentioned with a Hashtag
- **Training**: TwHIN framework (El-Kishky et al)



Author: *user1*
Tweet: Our paper was accepted at *@WNUT* with *@user2* *@user3* *#nlproc* *#socialmedia*
Favorited by: *user4*, *user5*

Table 1: Example tweet with engagement data of author, mentions, Hashtags, and favorites

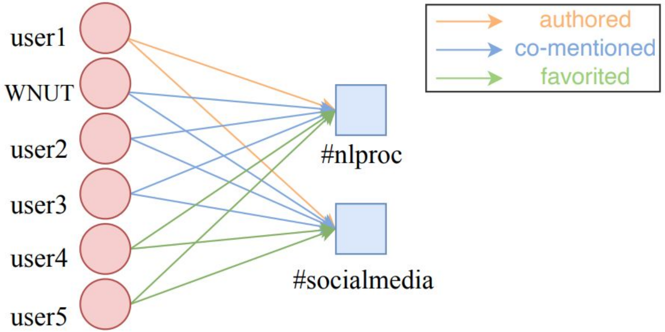
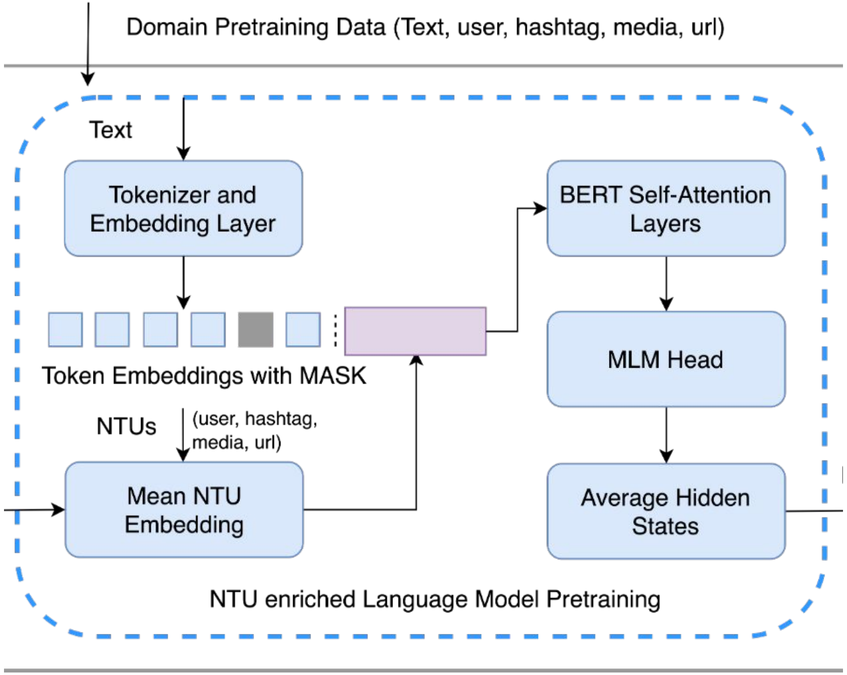


Figure 2: Graph construction with the example data in Table 1 for training NTULM user-Hashtag embeddings.

Ahmed El-Kishky, Thomas Markovich, Serim Park, Chetan Verma, Baekjin Kim, Ramy Eskander, Yury Malkov, Frank Portman, Sofía Samaniego, Ying Xiao, and Aria Haghighi. 2022. TwHIN: Embedding the Twitter Heterogeneous Information Network for Personalized Recommendation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22). Association for Computing Machinery, New York, NY, USA, 2842–2850. https://doi.org/10.1145/3534678.3539080

NTULM: Masked Language Modeling

- Tweet with NTUs, use average NTU embeddings
- Linear projection to map the average NTU embedding from graph space to LM space
- Concatenate NTU embedding to token embeddings
- Average embedding of NTU type for OOV NTUs
- Fine-tune NTULM via MLM


Domain Pretraining Data (Text, user, hashtag, media, url)
Text
Tokenizer and Embedding Layer
BERT Self-Attention Layers
Token Embeddings with MASK
MLM Head
NTUs
(user, hashtag, media, url)
Mean NTU Embedding
Average Hidden States
NTU enriched Language Model Pretraining

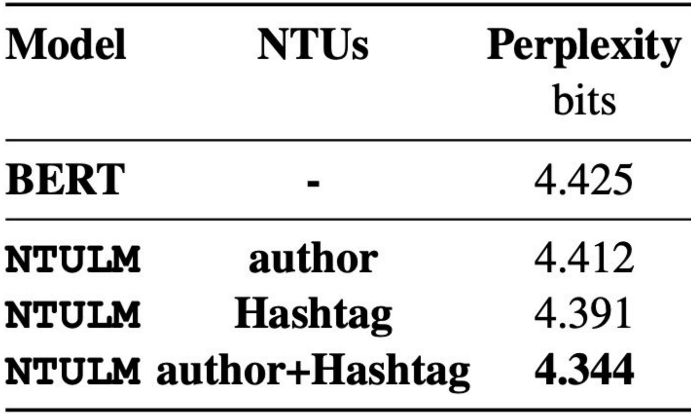
Experiments - Dataset

NTU heterogeneous network: Tweets (2018-01-01~2022-07-01) with Hashtags and their engagements with users, consisting of 60M Hashtags, 255M users, 5B authorship edges, 3B favorite edges, and 0.9B co-mention edges. We only considered users with 10 - 100 unique Hashtags interactions

MLM fine tuning: 1M Tweets sampled from (2022-06-01~2022-06-15).
We also fine-tune BERT without NTUs on these Tweets.

Downstream Tasks: TweetEval, SemEval, SocialMediaIE, Hashtag Pred, Topic

Results: Masked Language Modeling

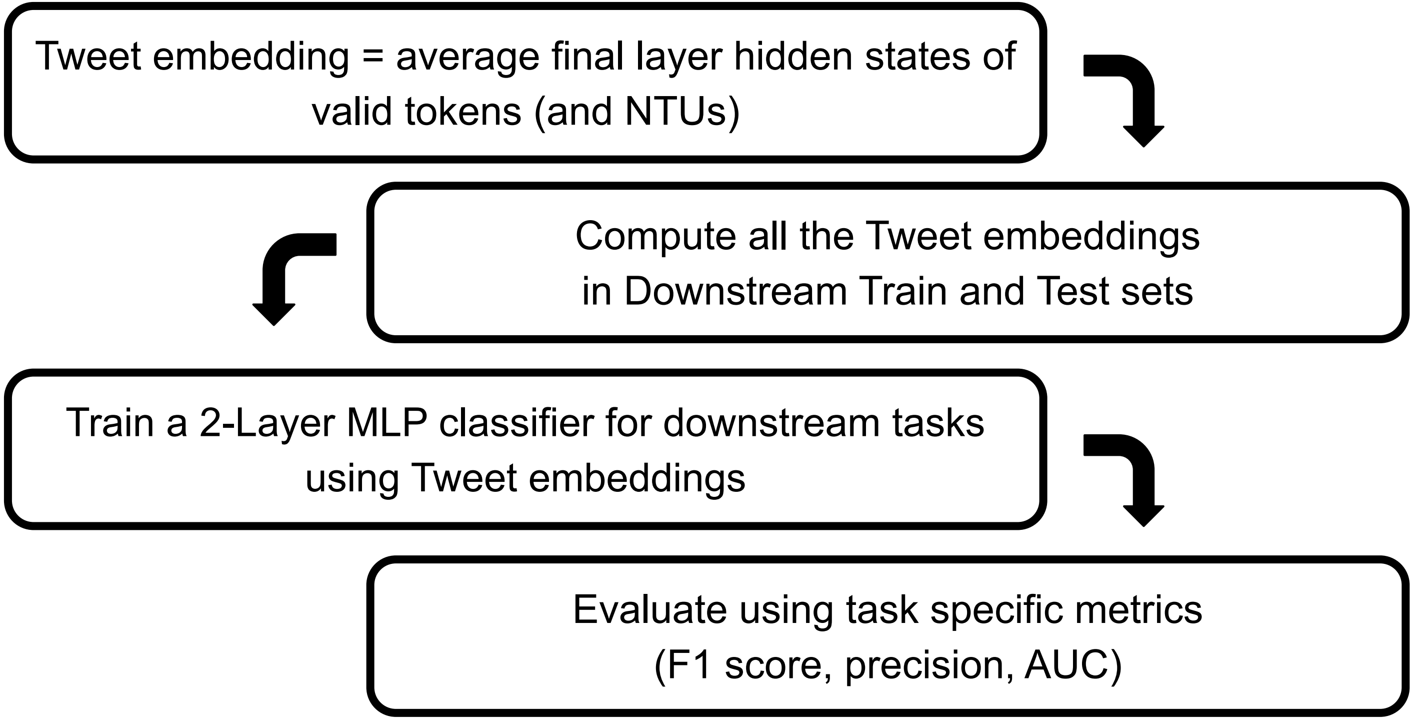


Model	NTUs	Perplexity bits
BERT	-	4.425
NTULM	author	4.412
NTULM	Hashtag	4.391
NTULM	author+Hashtag	**4.344**

Incorporating NTU embedding improves perplexity

Hashtag embedding is more effective than user embedding, combination is best

Evaluation on Downstream Tasks



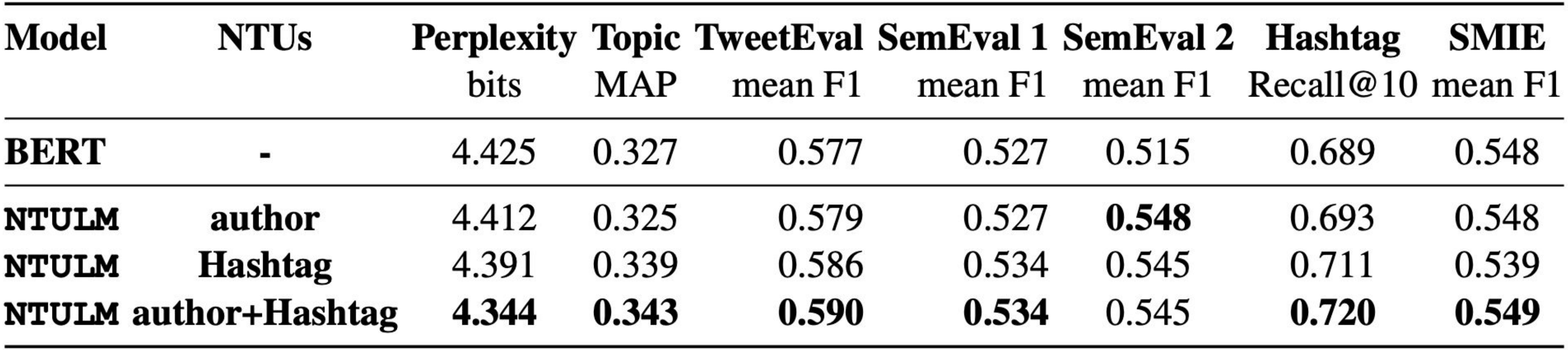
Tweet embedding = average final layer hidden states of valid tokens (and NTUs)

Compute all the Tweet embeddings in Downstream Train and Test sets

Train a 2-Layer MLP classifier for downstream tasks using Tweet embeddings

Evaluate using task specific metrics (F1 score, precision, AUC)

Results: All tasks



Model	**NTUs**	**Perplexity**	**Topic**	**TweetEval**	**SemEval 1**	**SemEval 2**	**Hashtag**	**SMIE**
		bits	MAP	mean F1	mean F1	mean F1	Recall@10	mean F1
BERT	-	4.425	0.327	0.577	0.527	0.515	0.689	0.548
NTULM	**author**	4.412	0.325	0.579	0.527	**0.548**	0.693	0.548
NTULM	**Hashtag**	4.391	0.339	0.586	0.534	0.545	0.711	0.539
NTULM	**author+Hashtag**	**4.344**	**0.343**	**0.590**	**0.534**	0.545	**0.720**	**0.549**

Incorporating NTU embedding improves downstream task performance

Hashtag embedding is more effective than user embedding, combination is best

NTU Overlap in downstream datasets

Dataset	Hashtag overlap	User overlap
Hashtag	99%	10%
SemEval	92%	21%
Social Media IE	95%	22%
Topic	99%	14%
TweetEval	98%	0%
Grand Total	95%	14%

Downstream Hashtags more likely to overlap with NTU embeddings than users.

Why is NTULM effective?

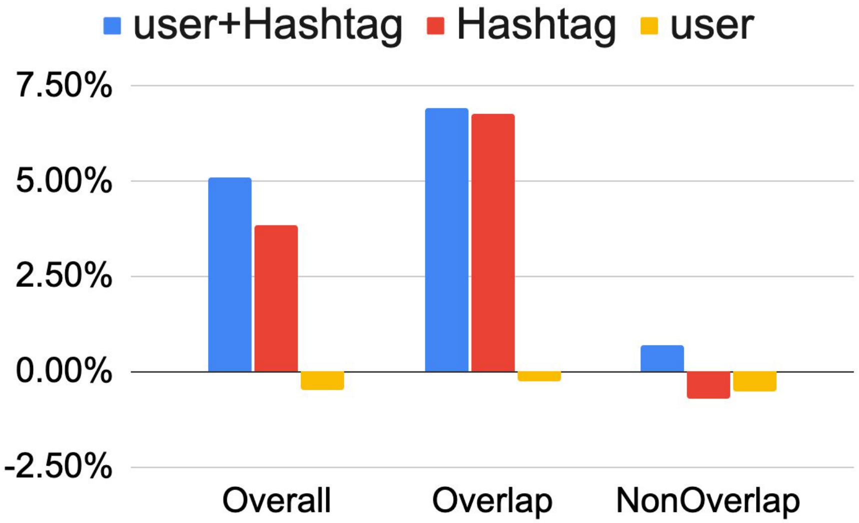
Hypothesis:

- If NTU is available, NTULM should help.
- If NTU is absent, NTULM should be similar to BERT.

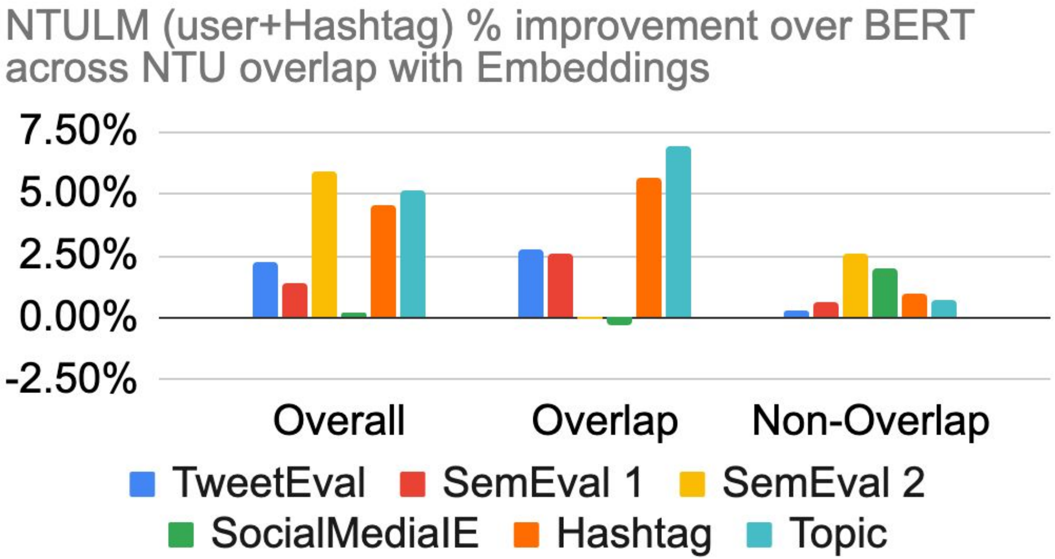
Observation:

- Hypothesis holds
- Gains with Hashtag NTU are much better than user.

Topic Task % improvement over baseline BERT model



Results: Overlap performance

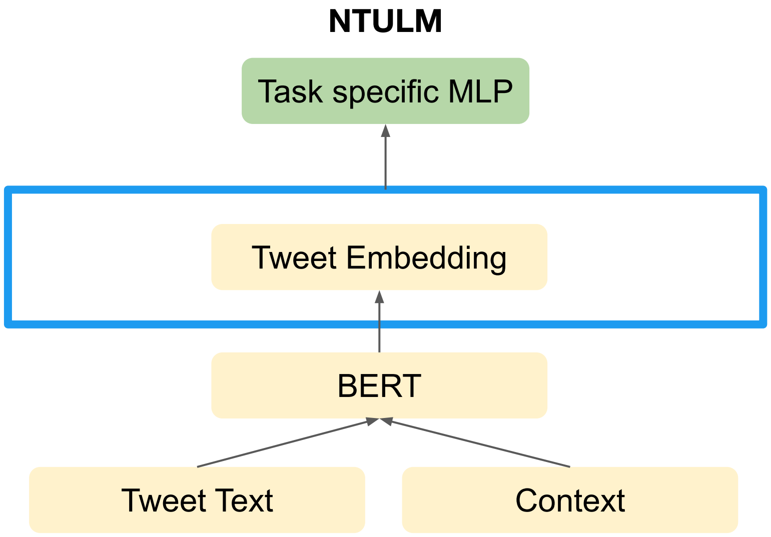


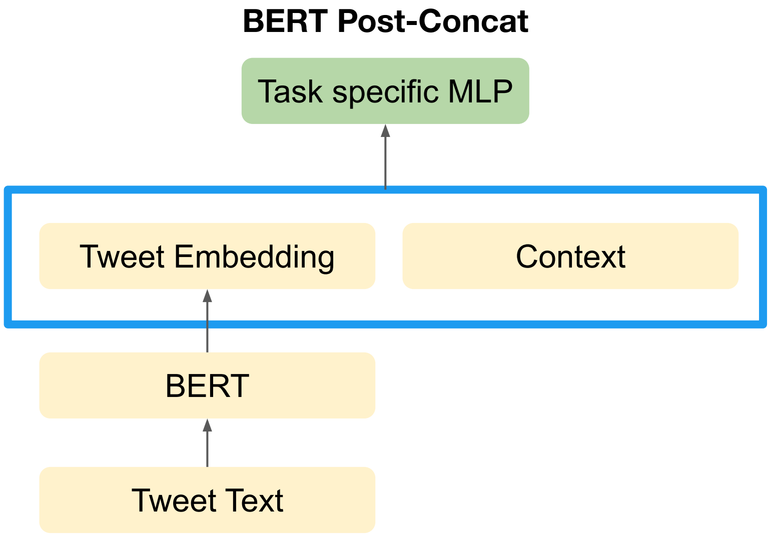
NTULM improved over BERT more when we have no OOV NTUs

Even for no NTUs, NTULM learns good text based embeddings which show small improvements.

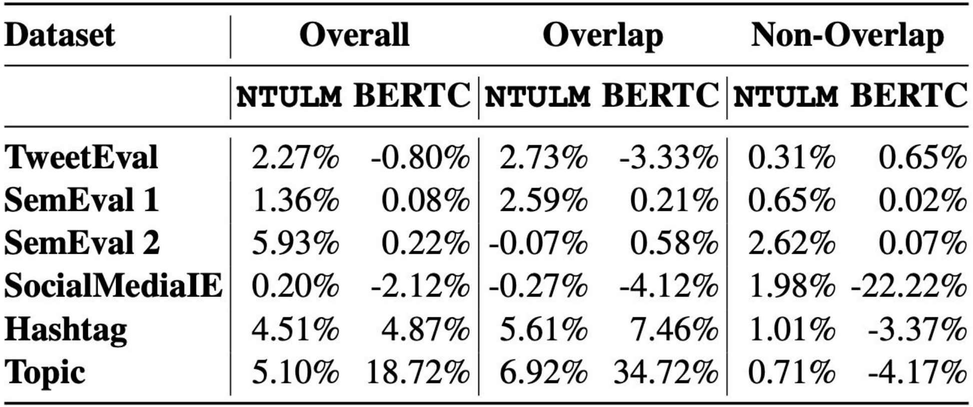
NTULM v/s BERT and Context separate

Alternative way to add context embedding: concatenate the context embedding after the BERT encoder? (named BERT Post-Concat or BERTC)

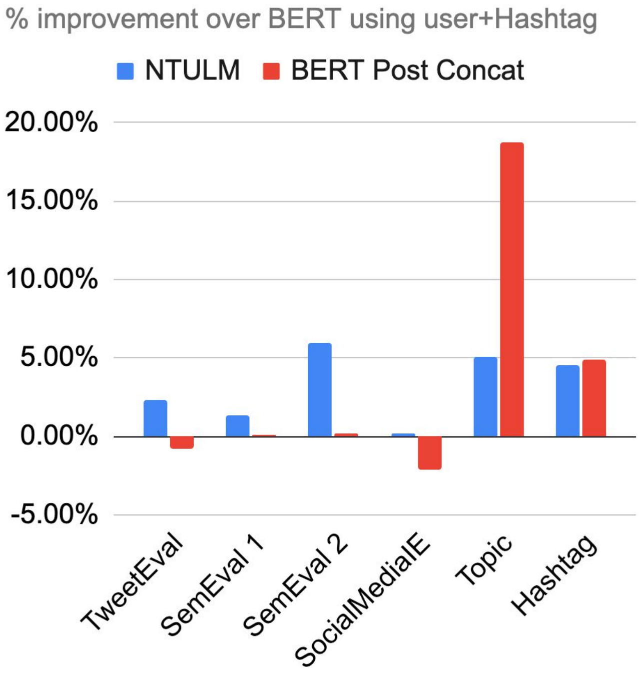




NTULM v/s BERT and Context separate



Dataset	Overall		Overlap		Non-Overlap	
	NTULM	BERTC	NTULM	BERTC	NTULM	BERTC
TweetEval	2.27%	-0.80%	2.73%	-3.33%	0.31%	0.65%
SemEval 1	1.36%	0.08%	2.59%	0.21%	0.65%	0.02%
SemEval 2	5.93%	0.22%	-0.07%	0.58%	2.62%	0.07%
SocialMediaIE	0.20%	-2.12%	-0.27%	-4.12%	1.98%	-22.22%
Hashtag	4.51%	4.87%	5.61%	7.46%	1.01%	-3.37%
Topic	5.10%	18.72%	6.92%	34.72%	0.71%	-4.17%



- **NTULM** integrates contexts embedding before attention layer, enabling the BERT encoder to automatically learn the attention of context embeddings.
- **BERTC** directly attach the context embedding after encoder, making it over-dependent on context embedding (affects the language model itself)

Vision for high business impact

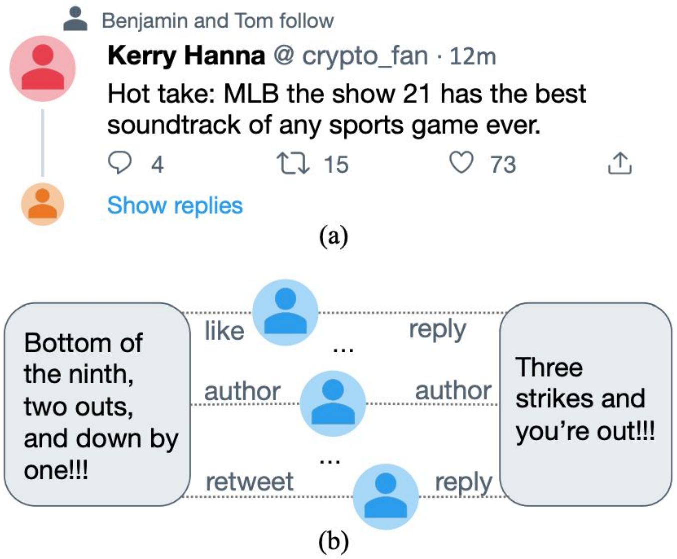
- Many text based user inputs are contextual.
- Using these inputs with their context can lead to better representation.
- This can help address:
  - The cold-start problem (as embeddings based on text), and
  - The popularity bias (as text embeddings are contextual)
- Social Graph can be integrated directly as part of language models to have better item and session representations which can power search and recommendation systems.

Recap

- Context is important when modeling language
- Spatio temporal and social context when utilized can lead to improved performance of language models on downstream tasks
- NTULM shows how to integrate social context of Non Textual Units into language models
- NTULM led to significant improvements on a variety of tasks over other baselines
- Improving coverage of NTUs may further improve NTULM.

Some newer works

TwHIN BERT: Tweets which are co-engaged are similar


Benjamin and Tom follow

Kerry Hanna @ crypto_fan · 12m

Hot take: MLB the show 21 has the best soundtrack of any sports game ever.

Bottom of the ninth, two outs, and down by one!!!

Three strikes and you're out!!!

Figure 1: (a) This mock-up shows a short-text Tweet and social engagements such as Faves, Retweets, Replies, Follows that create a social context to Tweets and signify Tweet appeal to engaging users. (b) Co-engagement is a strong indicator of Tweet similarity.

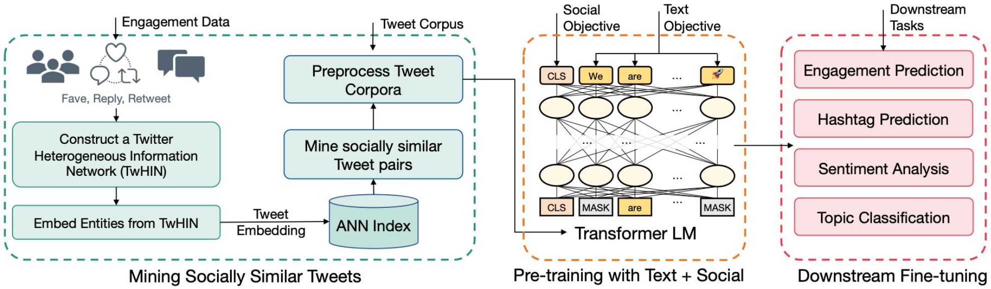
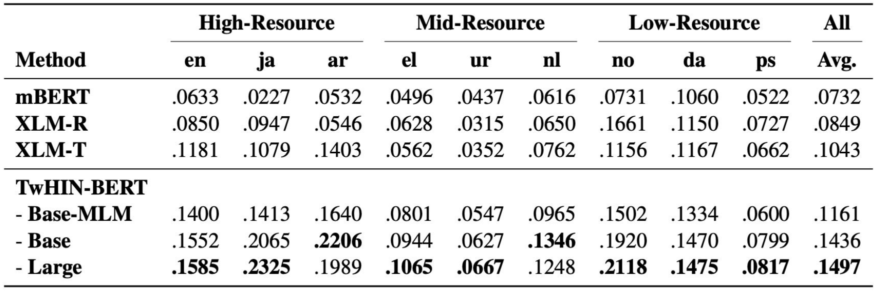


Figure 2: We outline the end-to-end TwHIN-BERT process. This three-step process involves (1) mining socially similar Tweet pairs by embedding a Twitter Heterogeneous Information Network (2) training TwHIN-BERT using a joint social and MLM objective and finally (3) fine-tuning TwHIN-BERT on downstream tasks.

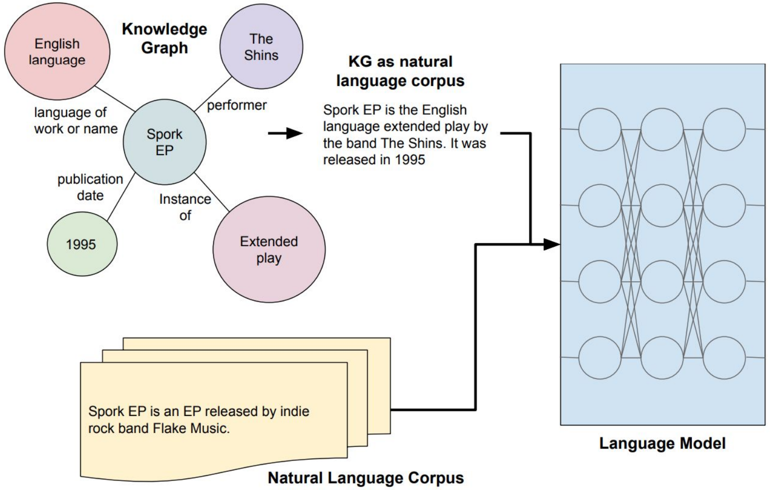
Table 1: Engagement prediction HITS@10 on high, mid, low-resource, and average of all languages.

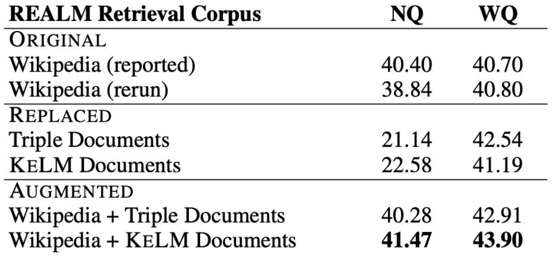


	High-Resource			Mid-Resource			Low-Resource			All
Method	**en**	**ja**	**ar**	**el**	**ur**	**nl**	**no**	**da**	**ps**	**Avg.**
mBERT	.0633	.0227	.0532	.0496	.0437	.0616	.0731	.1060	.0522	.0732
XLM-R	.0850	.0947	.0546	.0628	.0315	.0650	.1661	.1150	.0727	.0849
XLM-T	.1181	.1079	.1403	.0562	.0352	.0762	.1156	.1167	.0662	.1043
TwHIN-BERT										
- Base-MLM	.1400	.1413	.1640	.0801	.0547	.0965	.1502	.1334	.0600	.1161
- Base	.1552	.2065	**.2206**	.0944	.0627	**.1346**	.1920	.1470	.0799	.1436
- Large	**.1585**	**.2325**	.1989	**.1065**	**.0667**	.1248	**.2118**	**.1475**	**.0817**	**.1497**

[2209.07562] TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations

KELM: Integrating Knowledge Graphs with Language Model Pre-training Corpora


Knowledge Graph
English language
The Shins
language of work or name
performer
Spork EP
publication date
Instance of
1995
Extended play
KG as natural language corpus
Spork EP is the English language extended play by the band The Shins. It was released in 1995
Spork EP is an EP released by indie rock band Flake Music.
Natural Language Corpus
Language Model



REALM Retrieval Corpus	**NQ**	**WQ**
ORIGINAL		
Wikipedia (reported)	40.40	40.70
Wikipedia (rerun)	38.84	40.80
REPLACED		
Triple Documents	21.14	42.54
KELM Documents	22.58	41.19
AUGMENTED		
Wikipedia + Triple Documents	40.28	42.91
Wikipedia + KELM Documents	**41.47**	**43.90**

Table 7: Exact Match (EM) accuracy of REALM on NQ and WQ. Pretraining corpus used is CC-News.

Performance on Natural Questions and Web Questions benchmark

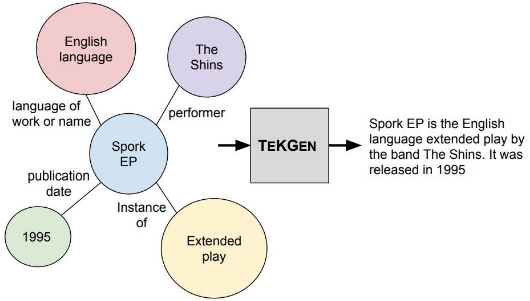
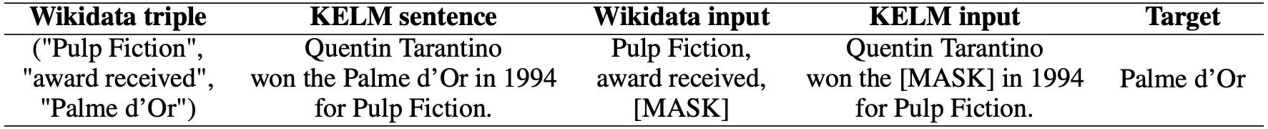


Figure 1: An example of generating text from KG. First, the entity subgraphs on the left are created and then converted to the sentence on the right.

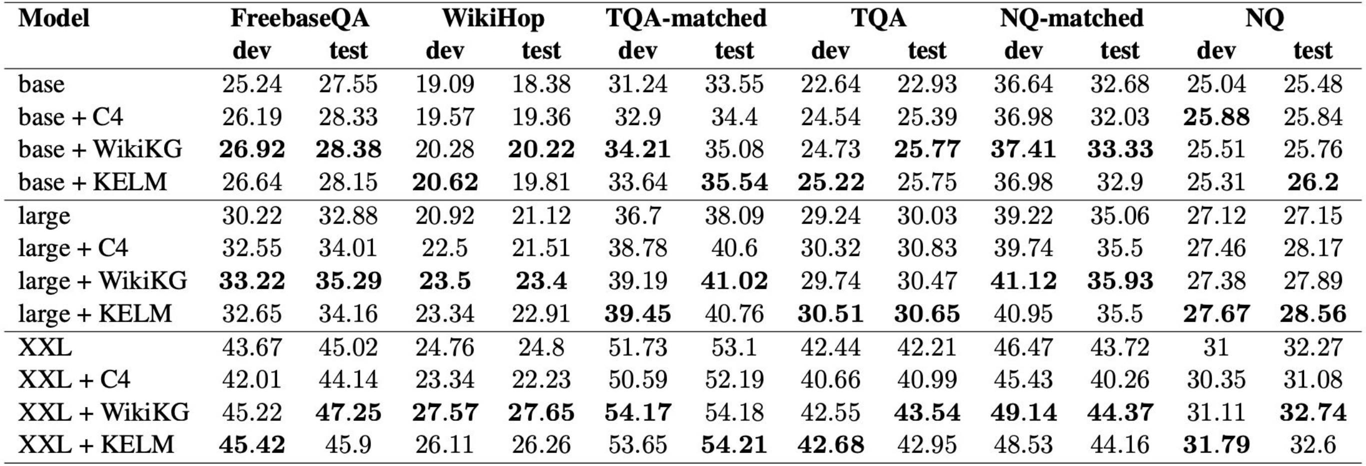
[2010.12688] Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training

SKILL: Structured Knowledge Infusion for Large Language Models

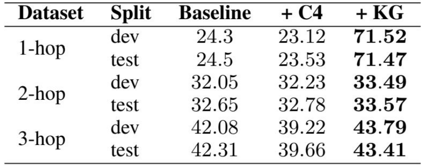


Wikidata triple	KELM sentence	Wikidata input	KELM input	Target
("Pulp Fiction", "award received", "Palme d'Or")	Quentin Tarantino won the Palme d'Or in 1994 for Pulp Fiction.	Pulp Fiction, award received, [MASK]	Quentin Tarantino won the [MASK] in 1994 for Pulp Fiction.	Palme d'Or

Table 1: Example inputs for SKILL pre-training with Wikidata and KELM corpora.



Model	FreebaseQA		WikiHop		TQA-matched		TQA		NQ-matched		NQ	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
base	25.24	27.55	19.09	18.38	31.24	33.55	22.64	22.93	36.64	32.68	25.04	25.48
base + C4	26.19	28.33	19.57	19.36	32.9	34.4	24.54	25.39	36.98	32.03	**25.88**	25.84
base + WikiKG	**26.92**	**28.38**	20.28	**20.22**	**34.21**	35.08	24.73	**25.77**	**37.41**	**33.33**	25.51	25.76
base + KELM	26.64	28.15	**20.62**	19.81	33.64	**35.54**	**25.22**	25.75	36.98	32.9	25.31	**26.2**
large	30.22	32.88	20.92	21.12	36.7	38.09	29.24	30.03	39.22	35.06	27.12	27.15
large + C4	32.55	34.01	22.5	21.51	38.78	40.6	30.32	30.83	39.74	35.5	27.46	28.17
large + WikiKG	**33.22**	**35.29**	**23.5**	**23.4**	39.19	**41.02**	29.74	30.47	**41.12**	**35.93**	27.38	27.89
large + KELM	32.65	34.16	23.34	22.91	**39.45**	40.76	**30.51**	**30.65**	40.95	35.5	**27.67**	**28.56**
XXL	43.67	45.02	24.76	24.8	51.73	53.1	42.44	42.21	46.47	43.72	31	32.27
XXL + C4	42.01	44.14	23.34	22.23	50.59	52.19	40.66	40.99	45.43	40.26	30.35	31.08
XXL + WikiKG	45.22	**47.25**	**27.57**	**27.65**	**54.17**	54.18	42.55	**43.54**	**49.14**	**44.37**	31.11	**32.74**
XXL + KELM	**45.42**	45.9	26.11	26.26	53.65	**54.21**	**42.68**	42.95	48.53	44.16	**31.79**	32.6



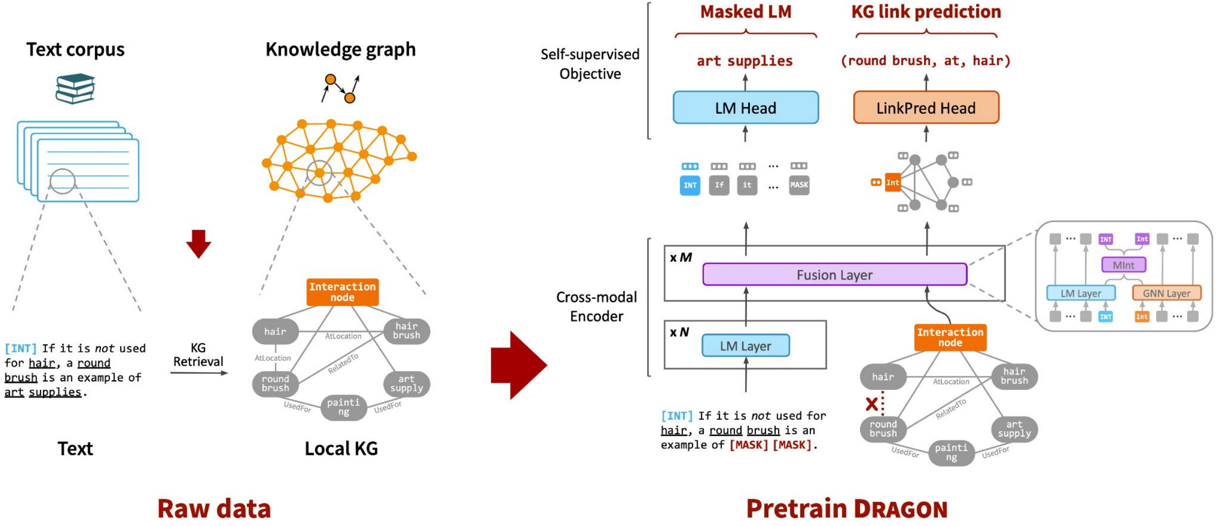
Dataset	Split	Baseline	+ C4	+ KG
1-hop	dev	24.3	23.12	**71.52**
	test	24.5	23.53	**71.47**
2-hop	dev	32.05	32.23	**33.49**
	test	32.65	32.78	**33.57**
3-hop	dev	42.08	39.22	**43.79**
	test	42.31	39.66	**43.41**

Table 3: Exact match scores achieved by fine-tuning different T5.1.1-large checkpoints on MetaQA task.

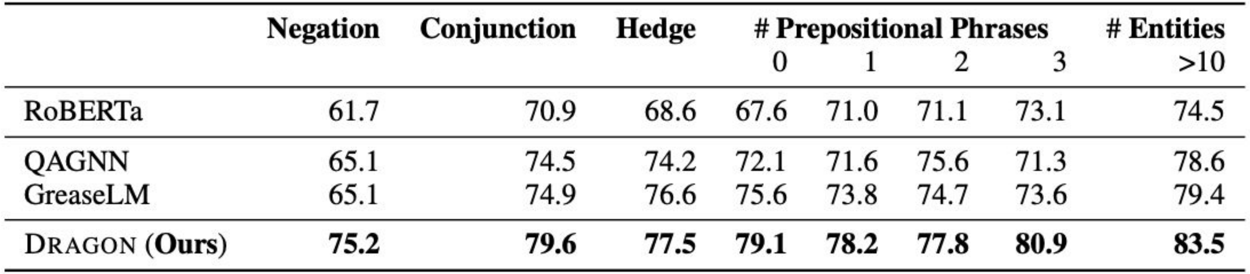
Good performance on WikiMovies KG (MetaQA) using multi-hop reasoning.

1-hop: "What films does Paresh Rawal appear in?"
2-hop: "Who are the directors of the films written by Laura Kerr?"
3-hop: "Who directed the movies written by the writer of Millennium Actress?"

DRAGON 🐲: Deep Bidirectional Language-Knowledge Graph Pretraining

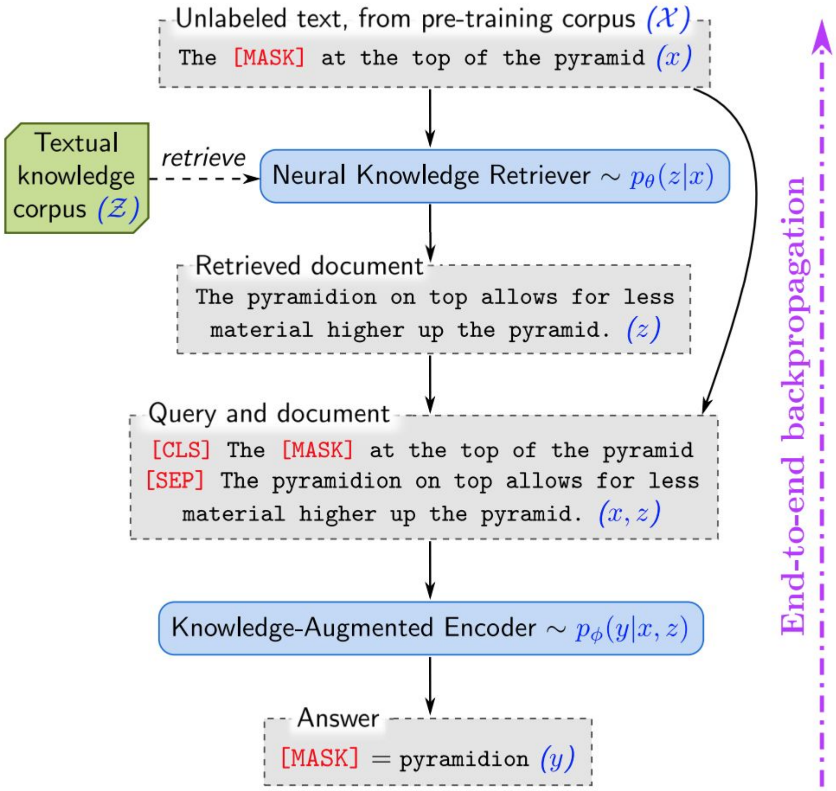


- LLM + GNN
- Joint optimization
- Very good on Entity based question answering



	Negation	Conjunction	Hedge	# Prepositional Phrases				# Entities
				0	1	2	3	>10
RoBERTa	61.7	70.9	68.6	67.6	71.0	71.1	73.1	74.5
QAGNN	65.1	74.5	74.2	72.1	71.6	75.6	71.3	78.6
GreaseLM	65.1	74.9	76.6	75.6	73.8	74.7	73.6	79.4
DRAGON (Ours)	**75.2**	**79.6**	**77.5**	**79.1**	**78.2**	**77.8**	**80.9**	**83.5**

REALM: Retrieval-Augmented Language Model Pre-Training



Unlabeled text, from pre-training corpus $(\mathcal{X})$

The [MASK] at the top of the pyramid (x)

Textual knowledge corpus $(\mathcal{Z})$

retrieve

Neural Knowledge Retriever $\sim p_\theta(z|x)$

Retrieved document

The pyramidion on top allows for less material higher up the pyramid. (z)

Query and document

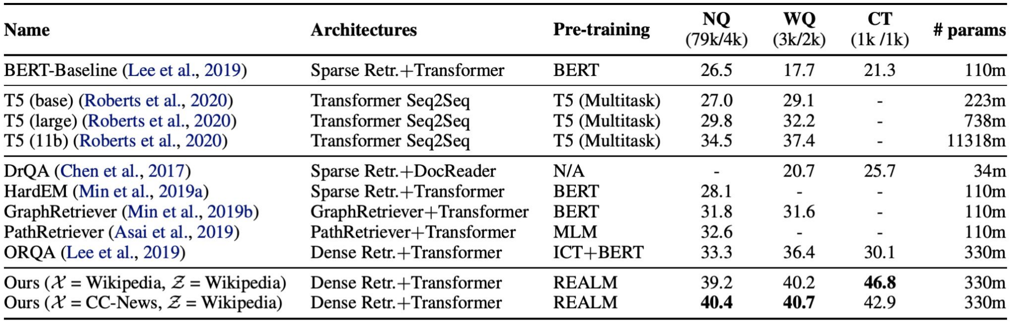
[CLS] The [MASK] at the top of the pyramid [SEP] The pyramidion on top allows for less material higher up the pyramid. (x, z)

Knowledge-Augmented Encoder $\sim p_\phi(y|x, z)$

Answer

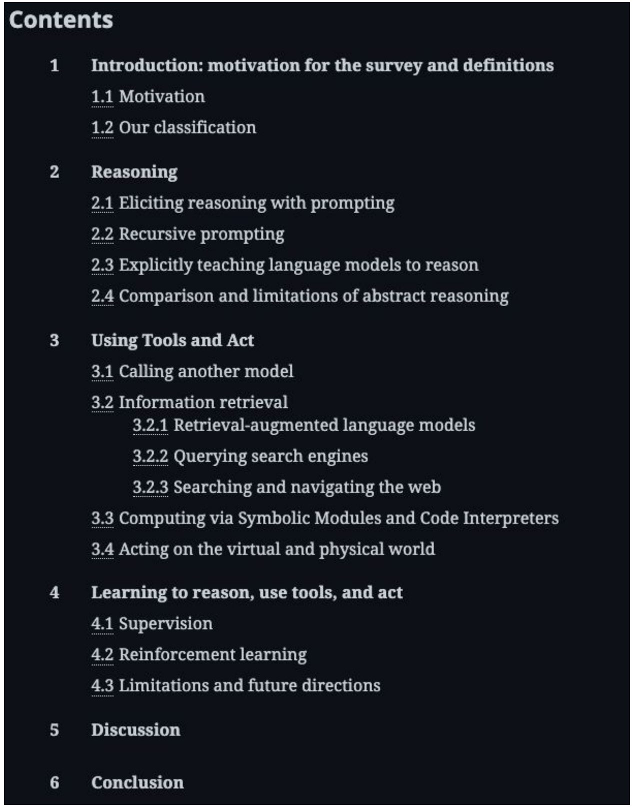
[MASK] = pyramidion (y)

End-to-end backpropagation



Name	**Architectures**	**Pre-training**	**NQ**	**WQ**	**CT**	**# params**
			(79k/4k)	(3k/2k)	(1k /1k)	
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
ORQA (Lee et al., 2019)	Dense Retr.+Transformer	ICT+BERT	33.3	36.4	30.1	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

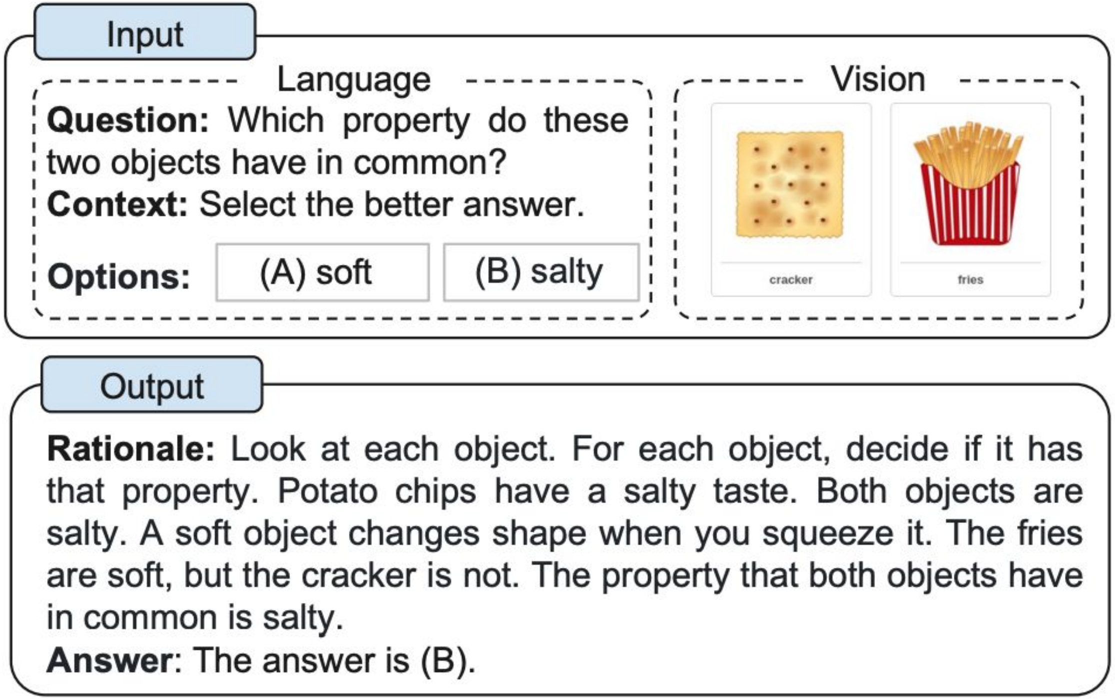
Augmented Language Models: a Survey



Contents

Really great survey on different ways of augmenting LLMs to support reasoning, usage of external tools, and retrieval models using external KG.

Multimodal Chain-of-Thought Reasoning in Language Models



Question: Which property do these two objects have in common?
Context: Select the better answer.

Options: (A) soft (B) salty

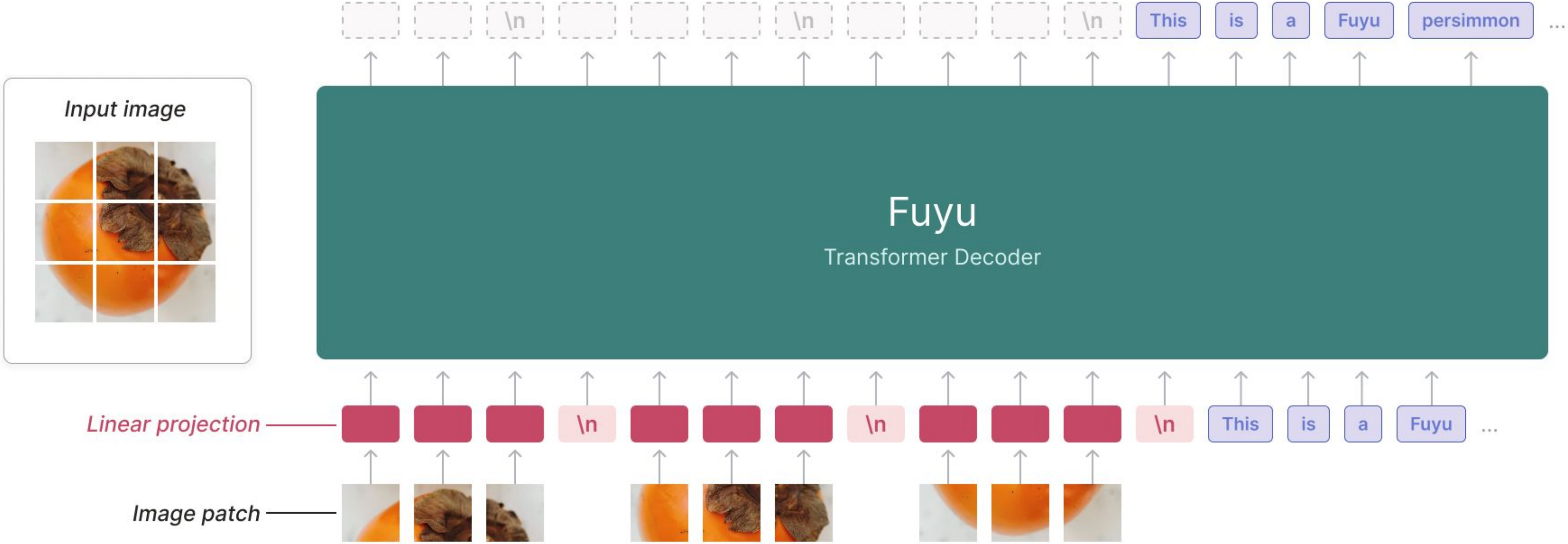
Rationale: Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.
Answer: The answer is (B).

Figure 1. Example of the multimodal CoT task.

[2302.00923] Multimodal Chain-of-Thought Reasoning in Language Models

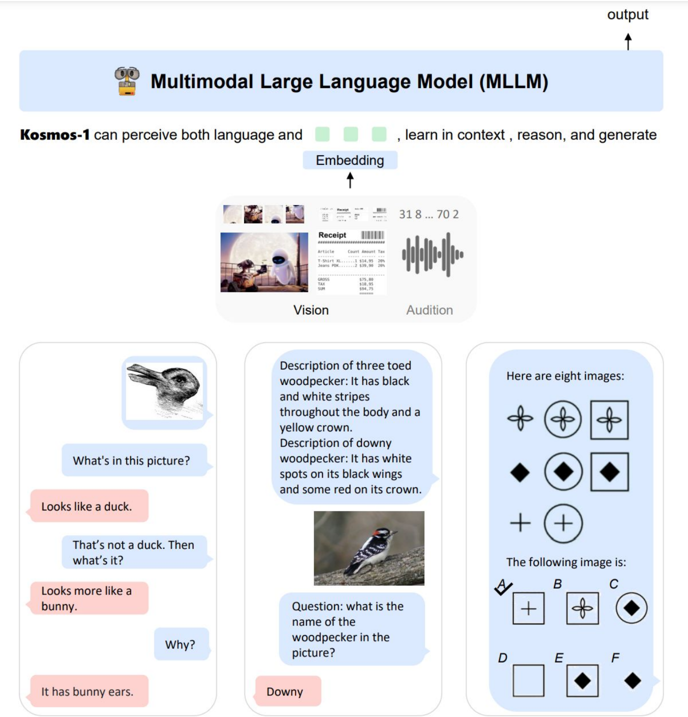
1B parameter model

Fuyu MLLM



Fuyu-8B: A Multimodal Architecture for AI Agents

KOSMOS-1



Kosmos-1 can perceive both language and 🟩 🟩 🟩 , learn in context , reason, and generate

What's in this picture?

Looks like a duck.

That's not a duck. Then what's it?

Looks more like a bunny.

Why?

It has bunny ears.

Description of three toed woodpecker: It has black and white stripes throughout the body and a yellow crown.
Description of downy woodpecker: It has white spots on its black wings and some red on its crown.

Question: what is the name of the woodpecker in the picture?

Downy

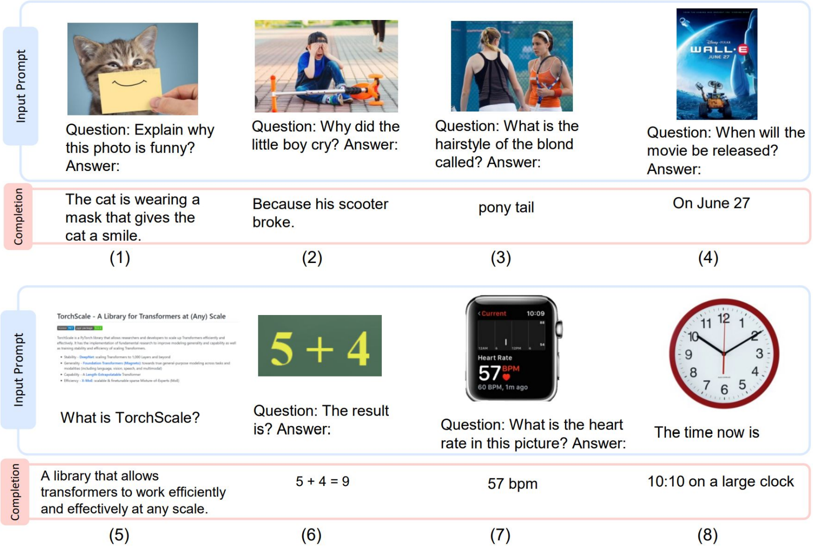
Here are eight images:

The following image is:

A ✓ B C D E F

Figure 1: Kosmos-1 is a multimodal large language model (MLLM) that is capable of perceiving multimodal input, following instructions, and performing in-context learning for not only language tasks but also multimodal tasks. In this work, we align vision with large language models (LLMs), advancing the trend of going from LLMs to MLLMs.

<s> paragraph **<image> Image Embedding </image>** paragraph </s>



(1) Question: Explain why this photo is funny? Answer: — The cat is wearing a mask that gives the cat a smile.

(2) Question: Why did the little boy cry? Answer: — Because his scooter broke.

(3) Question: What is the hairstyle of the blond called? Answer: — pony tail

(4) Question: When will the movie be released? Answer: — On June 27

(5) What is TorchScale? — A library that allows transformers to work efficiently and effectively at any scale.

(6) Question: The result is? Answer: — 5 + 4 = 9

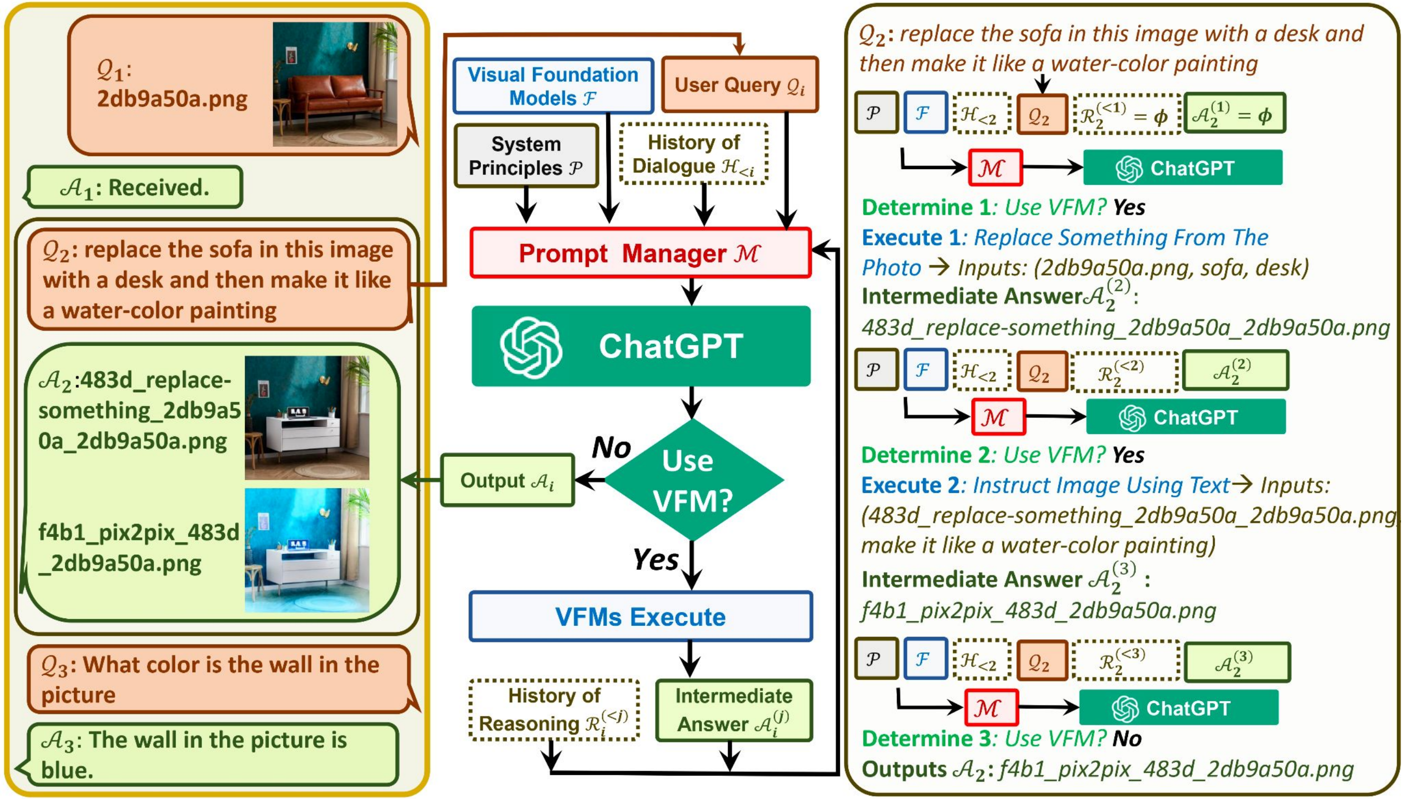
(7) Question: What is the heart rate in this picture? Answer: — 57 bpm

(8) The time now is — 10:10 on a large clock

Figure 2: Selected examples generated from Kosmos-1. Blue boxes are input prompt and pink boxes are Kosmos-1 output. The examples include (1)-(2) visual explanation, (3)-(4) visual question answering, (5) web page question answering, (6) simple math equation, and (7)-(8) number recognition.

[Language Is Not All You Need: Aligning Perception with Language Models](#)

Visual ChatGPT

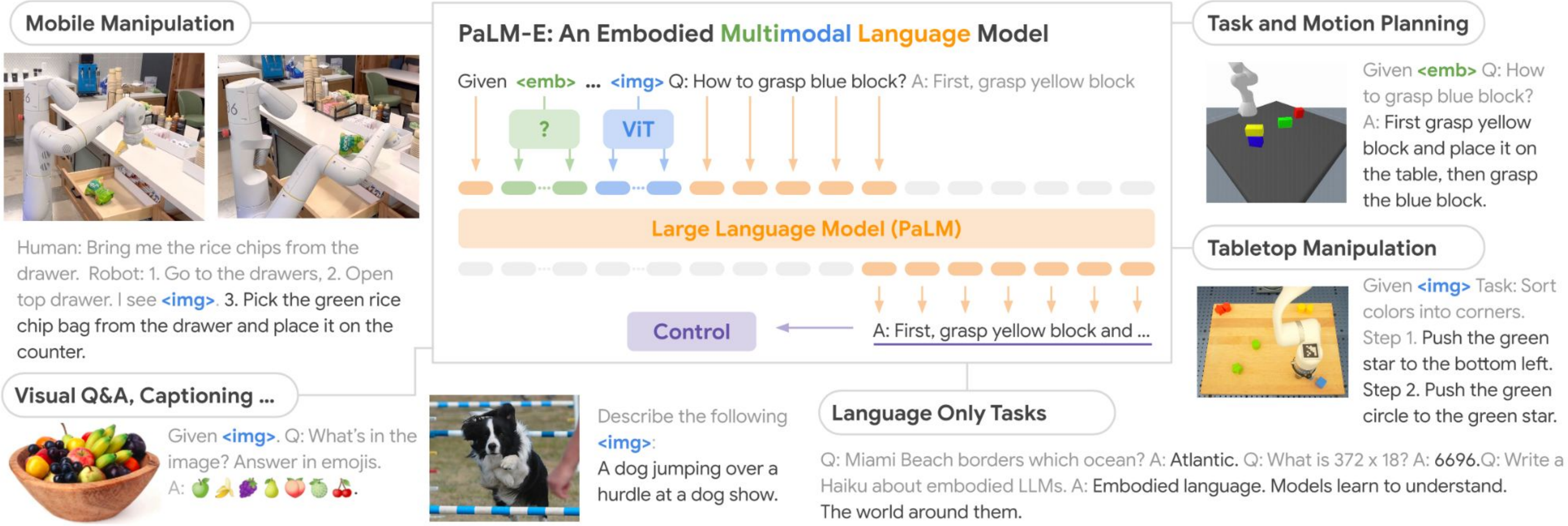


Q_1: 2db9a50a.png
A_1: Received.
Q_2: replace the sofa in this image with a desk and then make it like a water-color painting
A_2:483d_replace-something_2db9a50a_2db9a50a.png
f4b1_pix2pix_483d_2db9a50a.png
Q_3: What color is the wall in the picture
A_3: The wall in the picture is blue.
Visual Foundation Models $\mathcal{F}$
User Query Q_i
System Principles $\mathcal{P}$
History of Dialogue $\mathcal{H}_{<i}$
Prompt Manager $\mathcal{M}$
ChatGPT
Use VFM?
No
Yes
Output A_i
VFMs Execute
History of Reasoning $\mathcal{R}_i^{(<j)}$
Intermediate Answer $A_i^{(j)}$
Q_2: replace the sofa in this image with a desk and then make it like a water-color painting
$\mathcal{P}$ $\mathcal{F}$ $\mathcal{H}_{<2}$ Q_2 $\mathcal{R}_2^{(<1)} = \phi$ $A_2^{(1)} = \phi$
$\mathcal{M}$ ChatGPT
Determine 1: Use VFM? Yes
Execute 1: Replace Something From The Photo → Inputs: (2db9a50a.png, sofa, desk)
Intermediate Answer $A_2^{(2)}$: 483d_replace-something_2db9a50a_2db9a50a.png
$\mathcal{P}$ $\mathcal{F}$ $\mathcal{H}_{<2}$ Q_2 $\mathcal{R}_2^{(<2)}$ $A_2^{(2)}$
$\mathcal{M}$ ChatGPT
Determine 2: Use VFM? Yes
Execute 2: Instruct Image Using Text→ Inputs: (483d_replace-something_2db9a50a_2db9a50a.png, make it like a water-color painting)
Intermediate Answer $A_2^{(3)}$: f4b1_pix2pix_483d_2db9a50a.png
$\mathcal{P}$ $\mathcal{F}$ $\mathcal{H}_{<2}$ Q_2 $\mathcal{R}_2^{(<3)}$ $A_2^{(3)}$
$\mathcal{M}$ ChatGPT
Determine 3: Use VFM? No
Outputs A_2: f4b1_pix2pix_483d_2db9a50a.png

[2303.04671] Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models

GitHub - microsoft/visual-chatgpt: VisualChatGPT

PaLM-E: An Embodied Multimodal Language Model



PaLM-E: An Embodied Multimodal Language Model
?
ViT
Large Language Model (PaLM)
Control

Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block

A: First, grasp yellow block and ...

Mobile Manipulation

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Task and Motion Planning

Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation

Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

Visual Q&A, Captioning ...

Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

Describe the following <img>: A dog jumping over a hurdle at a dog show.

Language Only Tasks

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696.Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

Questions

Reach out for questions:

- Twitter: @TheShubhanshu
- LinkedIn: https://www.linkedin.com/in/shubhanshumishra
- Webpage: https://shubhanshu.com/

Links

- [2208.10174] KEEP: An Industrial Pre-Training Framework for Online Recommendation via Knowledge Extraction and Plugging
- [PDF] Training Large-Scale News Recommenders with Pretrained Language Models in the Loop | Semantic Scholar
- [2101.12294] Combining pre-trained language models and structured knowledge
- Prismer: A Vision-Language Model with Multi-Modal Experts - https://twitter.com/DrJimFan/status/1633868130932891648
- MultiDiffusion
- [2101.12294] Combining pre-trained language models and structured knowledge
- KELM: Integrating Knowledge Graphs with Language Model Pre-training Corpora - [2010.12688] Knowledge Graph Based
- [2210.03629] ReAct: Synergizing Reasoning and Acting in Language Models
  - https://github.com/hwchase17/langchain/blob/master/langchain/agents/conversational/prompt.py
- [2302.00923] Multimodal Chain-of-Thought Reasoning in Language Models
- Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems