Large Content And Behavior Models To Understand, Simulate, And Optimize Content And Behavior

ICLR-2024 Spotlight

Ashmit Khandelwal*[1] Aditya Agrawal*[1] Aanisha Bhattacharyya* Yaman K Singla*
Somesh Singh Uttaran Bhattacharya Ishita Dasgupta Stefano Petrangeli
Rajiv Ratn Shah Changyou Chen Balaji Krishnamurthy
Adobe, BITS, Pilani, IIIT-Delhi, State University of New York at Buffalo

behavior-in-the-wild.github.io/LCBM

Shannon's Theory Of Communication: Highlights

LEVEL A. How accurately can the symbols of communication be transmitted? (The technical problem.)

↓

LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)

↓

LEVEL C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

Principle scaled solution characteristic of each level

The Internet

Data generated by internet used to build LLM

LLMs

??

The problems we need to solve at the third level

- Simulate Human Behavior For A Certain Content
- Generate Content To Elicit A Given Behavior
- Use Behavior To Extract Signals About Content
- Explain Human Behavior

Idea: Why don't we use LLMs transfer learning power to solve the third level?

Large Language Models As Foundation To Enable Transfer Learning

[...] treat every text processing problem as a “text-to-text” problem, i.e. taking text as input and producing new text as output. The main utility of transfer learning is the possibility of leveraging pre-trained models in data-scarce settings

- T5, Raffel et al, 2020

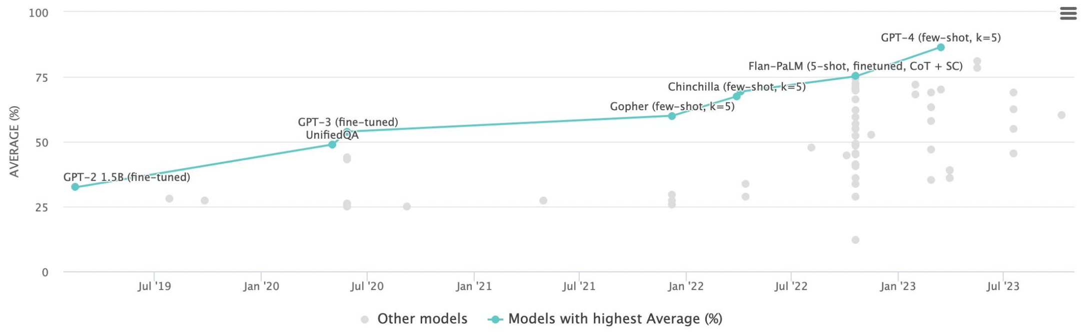
We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task.

- Improving Language Understanding By Generative Pre-training, Radford et al, 2018

Multi-task Language Understanding on MMLU

Leaderboard Dataset

View Average (%) by Date



Filter: few-shot fine-tuned zero-shot untagged

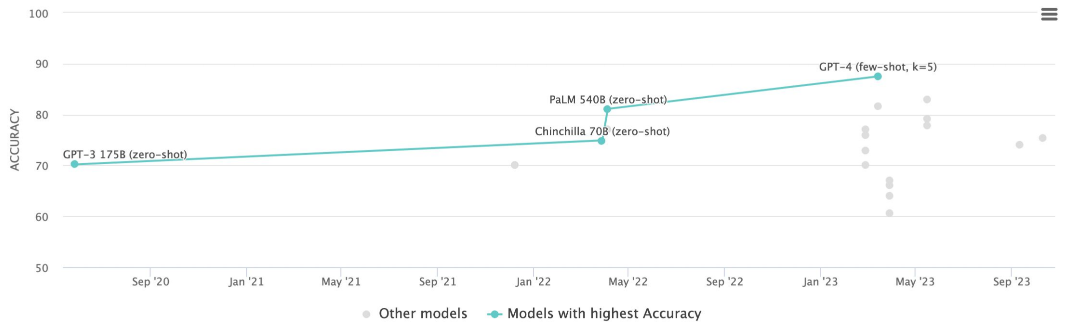
Edit Leaderboard

Rank	Model	Average (%)	Humanities	STEM	Social Sciences	Other	Parameters (Billions)	Tokens (Billions)	Paper	Code	Result	Year	Tags
1	GPT-4 (few-shot, k=5)	86.5							GPT-4 Technical Report			2023	few-shot
2	Flan-PaLM 2 (L)	81.2							PaLM 2 Technical Report			2023	
3	PaLM 2 (large)	78.3							PaLM 2 Technical Report			2023	
4	Flan-PaLM (5-shot, finetuned,	75.2					540		Scaling Instruction-Finetuned Language			2022	fine-tuned

Common Sense Reasoning on WinoGrande

Leaderboard Dataset

View Accuracy by Date



Filter: untagged

Edit Leaderboard

Rank	Model	Accuracy	Paper	Code	Result	Year	Tags
1	GPT-4 (few-shot, k=5)	87.5	GPT-4 Technical Report			2023	
2	PaLM 2-L (one-shot)	83.0	PaLM 2 Technical Report			2023	
3	GPT-3.5 (few-shot, k=5)	81.6	GPT-4 Technical Report			2023	
4	PaLM 540B (zero-shot)	81.1	PaLM: Scaling Language Modeling with Pathways			2022	
5	PaLM 2-M	79.2	PaLM 2 Technical Report			2023	

Transfer Learning Power of LLMs in Action
Vision Modality Fit Into Text-to-text Paradigm

🌋 LLaVA: Large Language and Vision Assistant

[Project Page] [Code] [Model] | 📚 [LLaVA] [LLaVA-v1.5]

LLaVA Chatbot

Which event is being talked about in this image?

The event being talked about in this image is a seminar on optimizing communication to achieve third level of Shannon's theory of communication.

Where is the event happening?

The event is happening at the NUS Computing building.

Enter text and press ENTER

Send

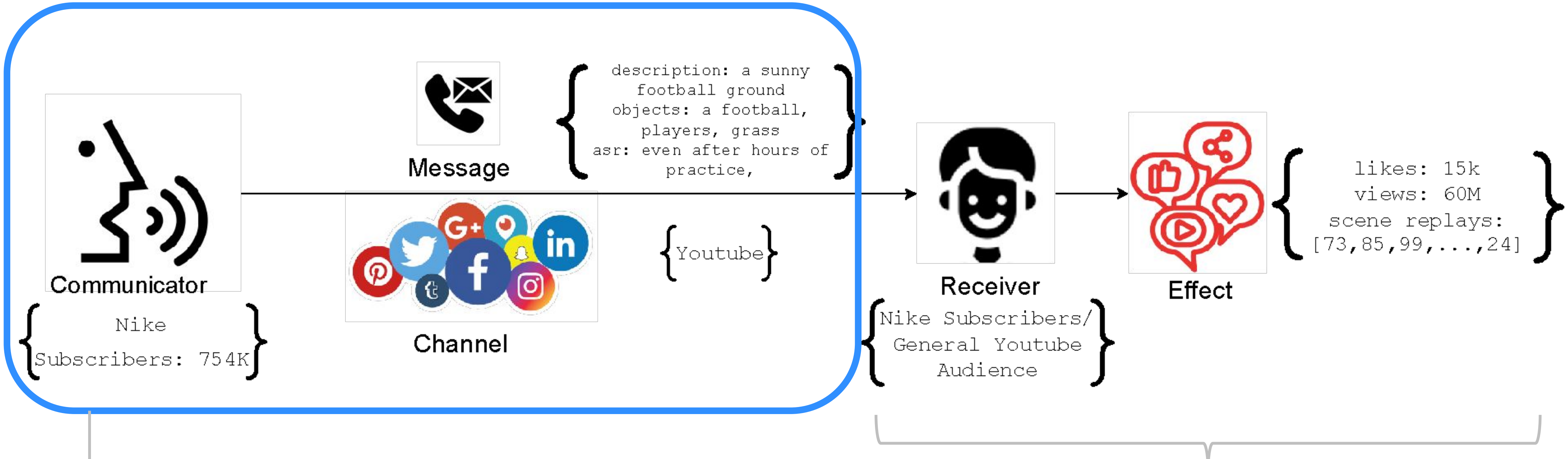
👍 Upvote | 👎 Downvote | ⚠ Flag | 🔄 Regenerate | 🗑 Clear

Liu, Haotian et al., Visual Instruction Tuning, *NeurIPS*, 2023

So,

1. Have LLMs transfer learnt Behavior already?
2. If not, why? And how do we make them transfer learn Behavior using the famous text-to-text paradigm?

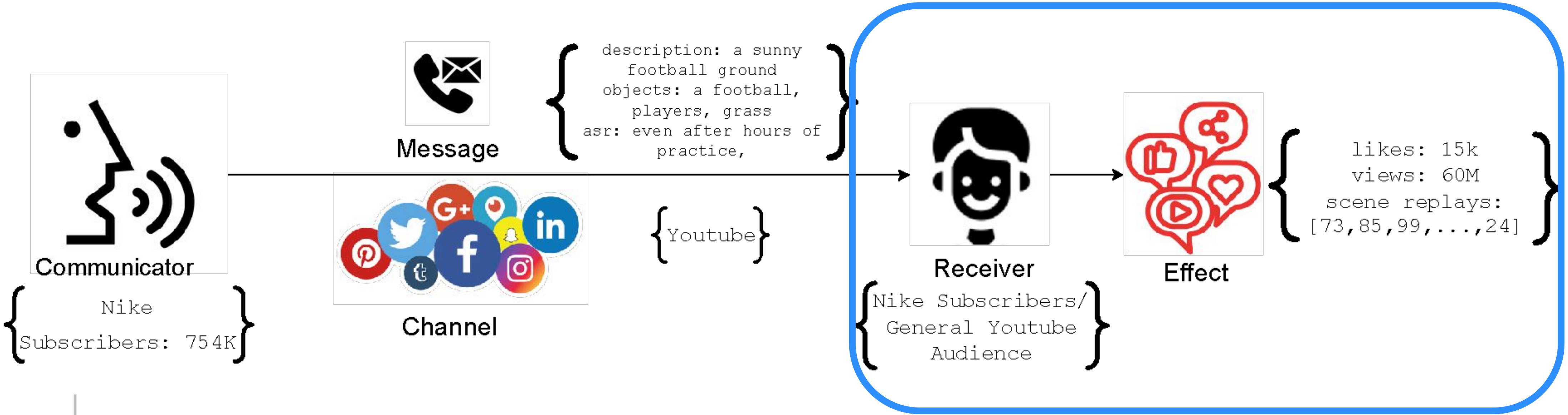
Human Communication Process And How LLMs Only See Half The Picture?



Removed as noise while preparing data for LLMs

- ~200 TBs — Common Crawl (Web-scale dump of webpages)
- ~200 TBs — Books
- ~50 TBs — Scientific Papers
- ~1 TBs — Reddit
- ~1 TBs — Discussion Forums such as StackExchange, hackernews, etc
- ~3 TBs — Github (Web-scale dump of code)

Human Communication Process And How We Are Seeing Only Half The Picture?



~200 TBs	Common Crawl (Web-scale dump of webpages)	Timestamp, Person, Effect
~200 TBs	Books	Timestamp, Person, Effect
~50 TBs	Scientific Papers	Timestamp, Person, Effect
~1 TBs	Reddit	Timestamp, Community, Upvotes, Comments
~1 TBs	Discussion Forums such as StackExchange, hackernews, etc	Timestamp, Community, Upvotes, Comments
~3 TBs	Github (Web-scale dump of code)	

How Can We See The Full Picture?

Large Content And Behavior Models (LCBMs)

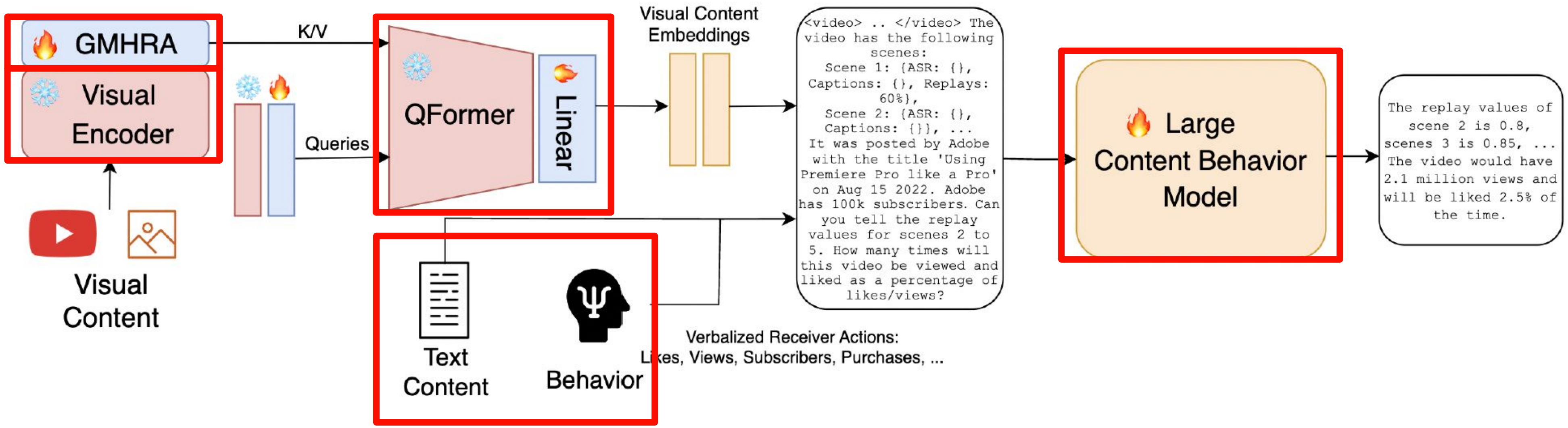


Figure 4: Encoding and predicting content (images, videos, and text) and behavior in the language space. Strategy to behavior instruction fine-tune (BFT) LLMs to create LCBMs. We capture visual concepts through the visual encoder (EVA-CLIP), and world knowledge is through an LLM (Llama). To leverage the rich knowledge of LLMs, we use GMHRA and QFormer to convert visual tokens of ViT to language tokens that Llama can understand. Further, we find that verbalizing the visual stimulus helps Llama to gather information more explicitly than what is provided by ViT+QFormer. We fine-tune the combined model end-to-end to predict 1) behavior given content and 2) content given behavior. Snowflake and fire symbols denote the frozen and unfrozen parts of the architecture.

What Happens If We See The Full Picture?

LCBM Results vs 15x Larger SOTA LLMs – Behavior and Content Simulation

Model	#Params	Training type	Training	RMSE	R^2	Accuracy
LCBM	13B	BFT	Replay values 3-masked	1.31	0.87	15.89
LCBM		BFT	Replay values 5-masked	1.48	0.82	19.93
LCBM		BFT	Replay values 7-masked	1.71	0.78	15.20
LCBM		BFT	Replay values 11-masked	1.55	0.82	13.94
GPT-4	$>$100B†	ICL	10-shot	3.50	-0.01	7.84
GPT-4		ICL	2-shot	3.58	-0.03	5.39
GPT-3.5	175B	ICL	3-shot	64.40	-256.96	2.48
GPT-3.5		ICL	2-shot	64.88	-375.83	1.27
Random	-	-	-	4.67	0	3.94

Model	#Params	Accuracy
Vicuna	13B	19.30%
LCBM	13B	48.68%
GPT-3.5	175B	34.98%
Random	-	4%

Youtube Content Simulation

Model	#Params	Training	Past		Future		Random Window Size 5		Random Window Size 7		All Masked	
			RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy	RMSE	Accuracy
LCBM	13B	3-BFT	8.12	55.10	15.05	42.42	8.55	61.41	9.91	55.10	-	-
LCBM		5-BFT	11.53	52.06	12.02	53.06	8.13	64.83	9.22	60.26	31.34	17.16
LCBM		7-BFT	16.17	35.61	15.14	44.11	9.02	59.22	10.47	53.84	-	-
LCBM		11-BFT	18.25	30.95	15.05	41.44	10.01	55.15	10.49	52.61	-	-
GPT-4	$>$100B†	10-shot-ICL	34.45	20.55	19.51	36.08	22.99	26.99	27.25	17.27	38.52	14.26
GPT-4		2-shot-ICL	35.05	19.34	18.07	39.33	17.42	38.10	21.26	28.05	37.60	13.73
GPT-3.5	175B	3-shot-ICL	34.10	19.06	24.71	27.14	24.52	24.81	26.30	18.74	38.77	13.47
GPT-3.5		2-shot-ICL	33.36	18.02	26.44	25.42	23.35	25.35	24.68	21.24	37.16	13.39
Random	-	-	34.10	10.00	34.10	10.00	34.10	10.00	34.10	10.00	34.10	10.00

Youtube Behavior Simulation

LCBM Shows Signs of Behavior Domain Adaptation

Model	#Params	Training type	Training	Time Separated	Brand Separated
GPT-3.5	175B	ICL	Few-shot	58.84	64.19
LCBM	13B	BFT	Twitter	74.3	97.69
LCBM	13B	BFT	Twitter and YouTube data	76.87	92.19

Table 9: **Behavior Simulation and Behavior Domain Adaptation**‡. Two-way classification accuracies for like prediction on Twitter. Given content, channel, and time, predict behavior (High, Low). We note that LCBM trained on Twitter and YouTube performs better than the one trained only on Twitter, showing signs of performance improvement by domain adaptation. BFT denotes behavior fine-tuning, and ICL stands for in-context learning. The best results over four runs are reported for all models. Best models are denoted in green and runner-ups in blue.

Model	Training	Test	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-l
GPT-3.5	ICL	Brand Separated	53.95	42.36	31.84	24.28	15.24
		Time Separated	57.69	45.11	33.67	25.52	15.27
LCBM	BFT on Twitter	Brand Separated	62.29	46.59	33.98	25.64	14.44
		Time Separated	70	54.4	41.43	32.48	17.38
LCBM	BFT on Twitter + Youtube	Brand Separated	64.28	48.1	35.17	26.63	14.83
		Time Separated	70.23	54.54	41.52	32.54	17.45

Table 10: **Content Simulation and Behavior Domain Adaptation**‡. Given behavior, channel, time, tweet media caption as prompt, predict content (tweet text). We note that LCBM trained on Twitter and YouTube performs better than the one trained only on Twitter, showing signs of performance improvement by domain adaptation. BFT denotes behavior fine-tuning, and ICL stands for in-context learning. The best results over four runs are reported for all models. Best models are denoted in green and runner-ups in blue.

LCBM Shows Signs of Behavior Domain Adaptation

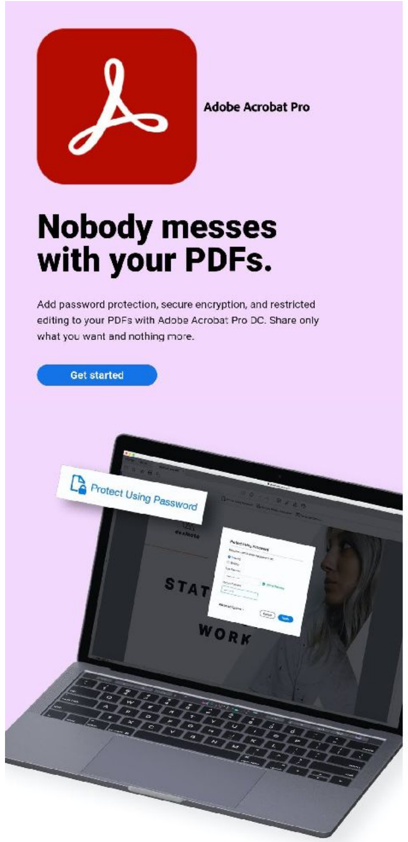
LCBM Results vs 15x Larger SOTA LLMs

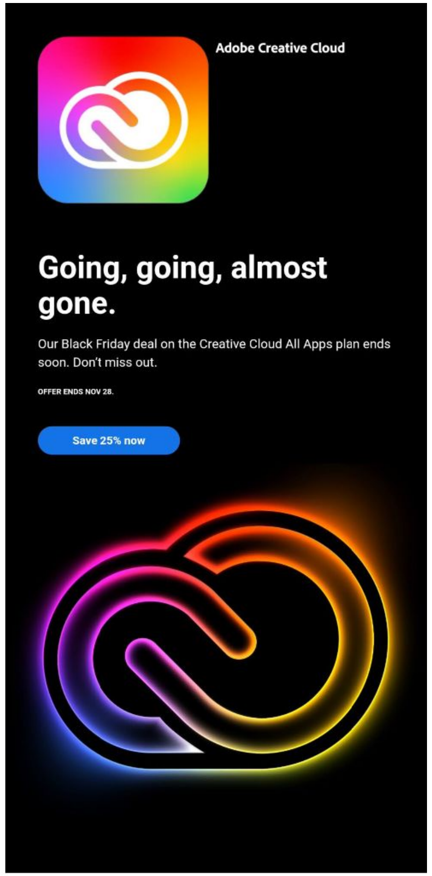
in-house Email Marketing

LCBM Type	Fine-tuned on YouTube?	Trained On: Unique Emails	Trained On: Unique Segments	Trained On: Email-Segment Pairs	Tested On	RMSE	R^2
Domain-Adapted	Yes	100	10	1k	Different Segment (emails could be same)	14.47	0.64
In-Domain	No	600	560k	350k		25.28	0.55
Domain-Adapted	Yes	100	10	1k	Different Segments & Different Emails	27.28	0.54
In-Domain	No	600	560k	350k		29.28	0.5

LVU Benchmark

Training	Model	Testing	MSE
Trained	R101-slowfast+NL (Wu & Krahenbuhl, 2021)	Test set	0.386
Trained	VideoBERT (Sun et al., 2019)	Test set	0.32
Trained	Qian et al. (2021)	Test set	0.353
Trained	Xiao et al. (2022)	Test set	0.444
Trained	Object Transformers (Wu & Krahenbuhl, 2021)	Test set	0.23
Zero-shot	LCBM (Ours)	Test set	0.14
Zero-shot	GPT-3.5	Test set	0.03
Zero-shot	Vicuna	Complete dataset	0.44
Zero-shot	LCBM (Ours)	Complete dataset	0.30
Zero-shot	GPT-3.5	Complete dataset	0.02





LCBM Shows Signs of Behavior Domain Adaptation

LCBM Results vs 15x Larger SOTA LLMs

Video Title: Robber hangs from the 23rd-floor window to escape the police in China
Posted by: South China Morning Post
Link: youtu.be/P7iDWZiT3Kg

Comments from Video:
(not shown to LLMs/LCBM)

Commenter: I feel bad for the dude. Nobody in the right mind robs a local shop.. To then hang out a window. I wonder what he is going through.

Commenter: Kinda feel sorry for him when he cried, he must have been in a bad situation to do what he did but it does not justify his actions.

Commenter: Aw, i felt his desperation. sorry young man but you did wrong. :(

Question: Would the average sentiment of the comments of the audience be positive, neutral, or negative? Explain why.

LCBM: The average sentiment of the comments of the audience watching the video would be negative. This is because the video shows a man jumping out of a high-rise building to escape the police. The audience is likely to feel sympathy and sadness towards the man's situation.

Vicuna: The audience would likely have a mix of positive and negative sentiment towards the video. Viewers may feel sympathetic towards the person trying to escape and praise their bravery.

GPT-3.5: The video depicts a dangerous and potentially life-threatening situation, which may cause some viewers to feel uneasy or uncomfortable. However, others may find the events exciting or thrilling to watch.

Video Title: The new Porsche 718 Spyder. Perfectly irrational.
Posted by: Porsche
Link: youtu.be/573YaSAP5H0

Comments from Video:
(not shown to LLMs/LCBM)

Commenter: Dream car. Mid engine, natural aspiration, b6, rwd, manual, lightweight, open top

Commenter: All the makings of a epic sports car 👌

Commenter: This car is perfect. I'm in love. New dream car, no questions asked!

Commenter: I add my vote and wishes to yours ... a great car I'd love to own and pilot

Question: Given the video, would the average audience sentiment be positive, neutral, or negative? Explain why.

LCBM: Positive average sentiment of the audience. The video features a luxury car driving along a scenic coastal road, which is would evoke a sense of awe and admiration in the viewers. Additionally, the Porsche 718 in the video is an iconic sports car that is widely recognized and respected in the automotive community, contributing to the positive sentiment.

Vicuna: The average sentiment of the comments would be positive. The video showcases a beautiful scenic drive through the mountains, which is likely to evoke a sense of awe and enjoyment in the viewers. Additionally, the presence of a luxury car in the video may also contribute to a positive sentiment, as it may be seen as a symbol of success and wealth.

GPT-3.5: The sentiment of the audience watching the video is positive because the video showcases a new sports car, which is exciting and desirable to car enthusiasts.

Figure 5: A few examples showing LCBM's ability to understand and explain human behavior of audience sentiment. We also compare it against other models like Vicuna and GPT-3.5.

Thank You

Paper Page

behavior-in-the-wild.github.io/LCBM