

Choose your *prompt* “well-aligned”

Collaborators: Subhabrata, Eshaan, Manish, Jeevesh

Based on our work published in ACL'22, ACL'23

Tanmoy Chakraborty
IIT Delhi, India
tanchak@iitd.ac.in
Web: tanmoychak.com
Lab: lcs2.in

The uprising of Large Language Models

Transformers: the backbone of large language models (LLM)

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez*[†]
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

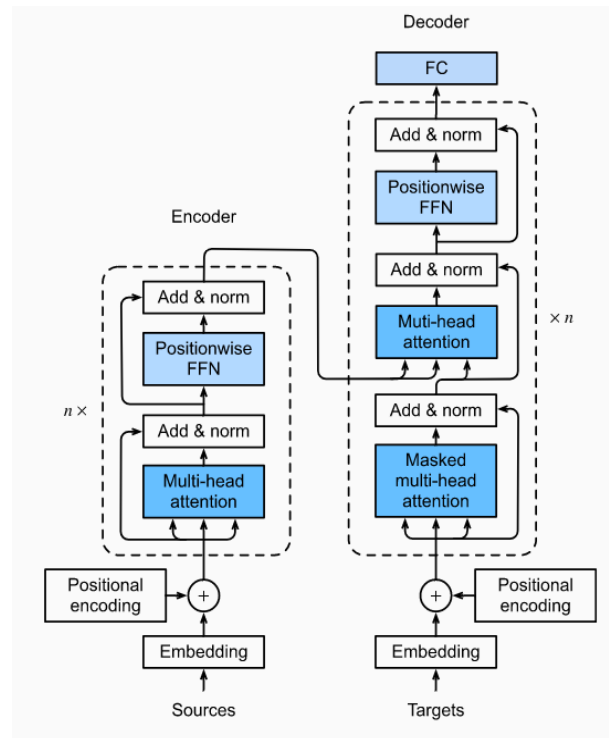
Illia Polosukhin*[‡]
illia.polosukhin@gmail.com

Attention is all you need

[A Vaswani, N Shazeer, N Parmar...](#) - Advances in neural ..., 2017 - proceedings.neurips.cc

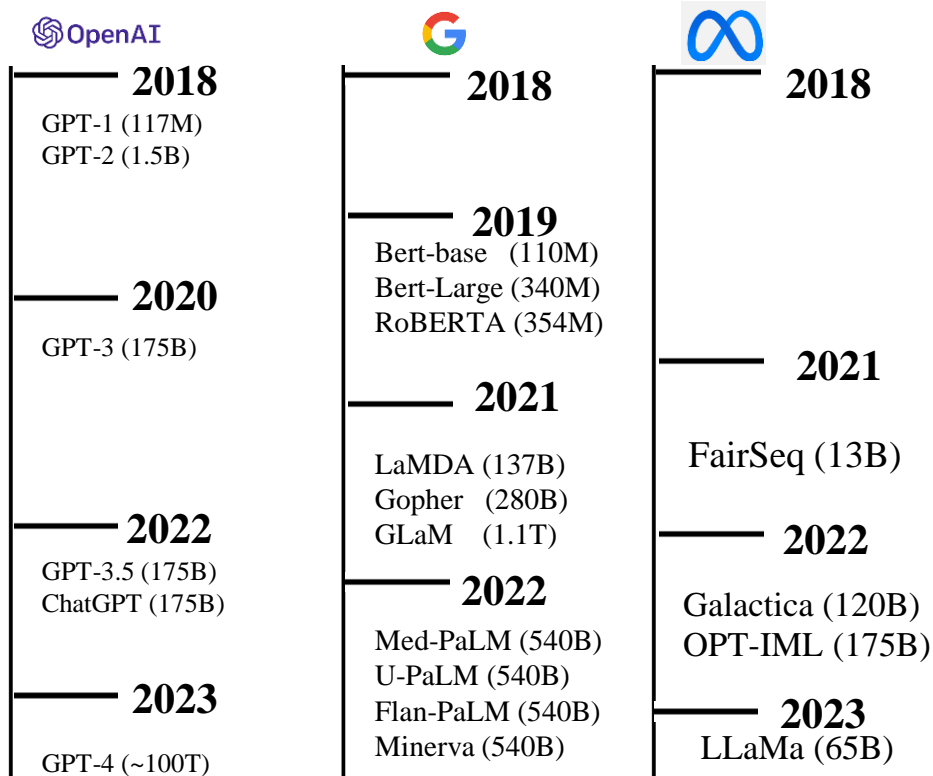
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attention mechanism. We propose a novel, simple network architecture based solely on an attention mechanism, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more ...

☆ Save 📄 Cite Cited by 69087 Related articles All 46 versions 🔗



The uprising of Large Language Models

Early pretrained LMs (BERT, RoBERTa, GPT) were mostly fine-tuned for downstream task

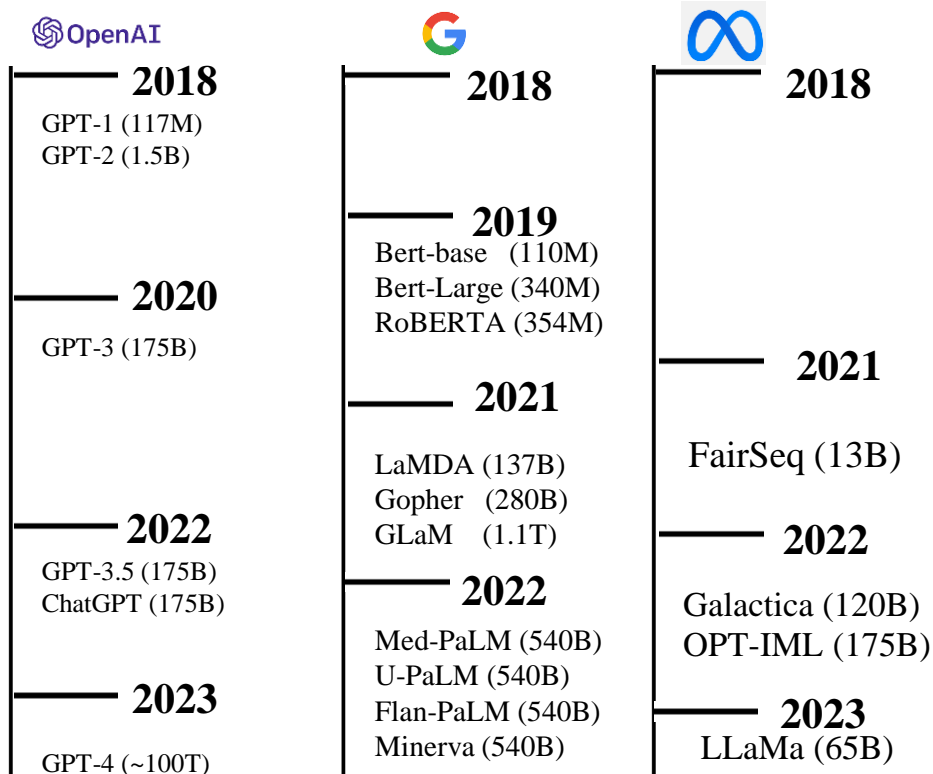


The uprising of Large Language Models

Early pretrained LMs (BERT, RoBERTa, GPT) were mostly fine-tuned for downstream task

- Add a classification head on top of LM and train it on labeled task data

Pretrain -> Finetune -> Predict



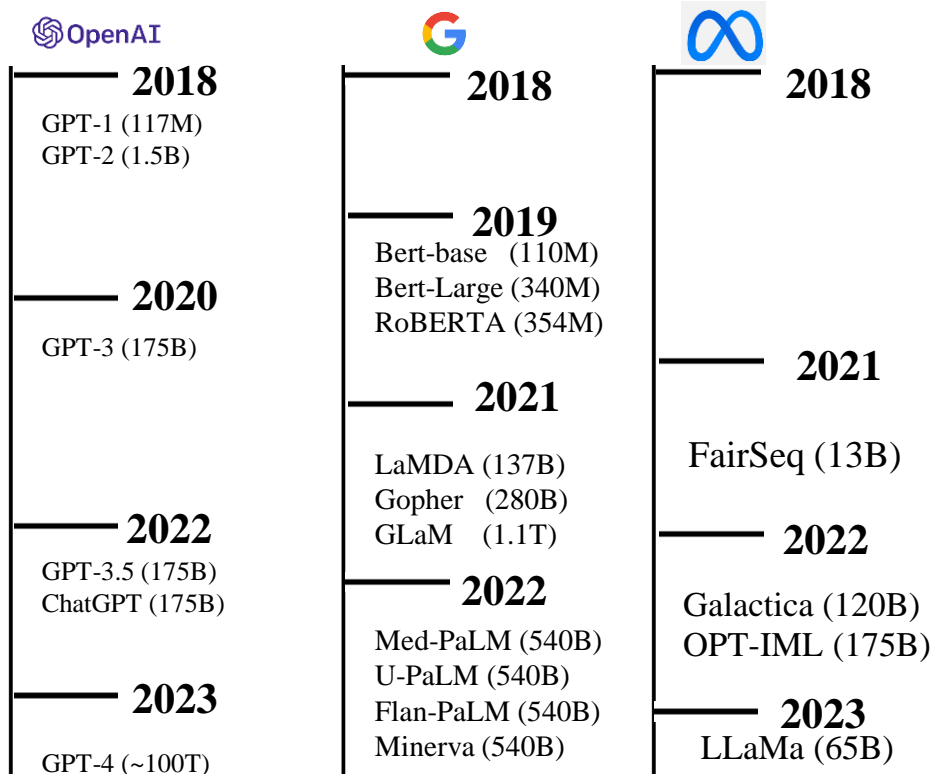
The uprising of Large Language Models

Early pretrained LMs (BERT, RoBERTa, GPT) were mostly fine-tuned for downstream task

- Add a classification head on top of LM and train it on labeled task data

Pretrain -> Finetune -> Predict

But GPT-3 showed a unique ability!



The uprising of Large Language Models

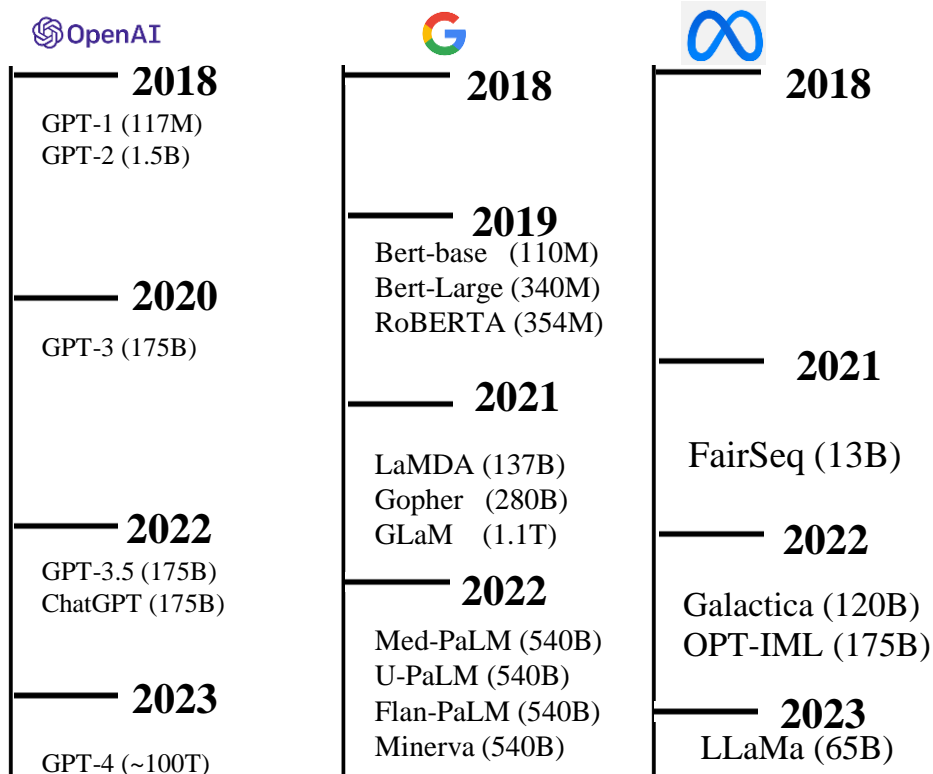
Early pretrained LMs (BERT, RoBERTa, GPT) were mostly fine-tuned for downstream task

- Add a classification head on top of LM and train it on labeled task data

Pretrain -> Finetune -> Predict

But GPT-3 showed a unique ability!

- Perform the task by generating natural language tokens, *aka*, *prompting*



The uprising of Large Language Models

Early pretrained LMs (BERT, RoBERTa, GPT) were mostly fine-tuned for downstream task

- Add a classification head on top of LM and train it on labeled task data

Pretrain -> Finetune -> Predict

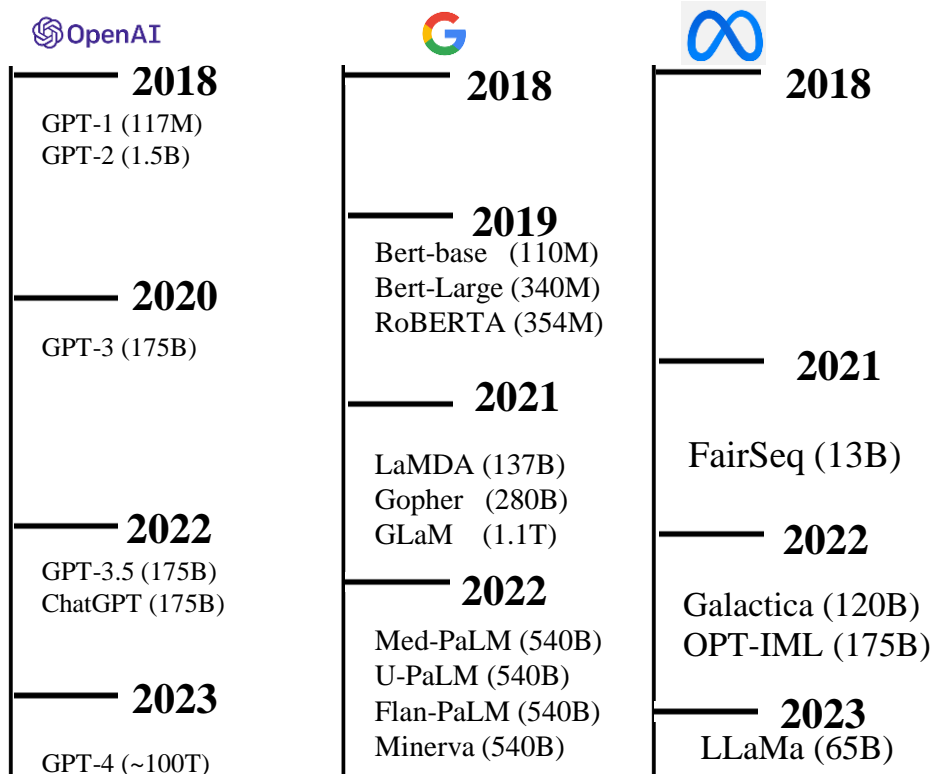
But GPT-3 showed a unique ability!

- Perform the task by generating natural language tokens, *aka*, **prompting**

That movie was great. Sentiment: Positive

It was a horrible day! Sentiment: Negative

This is an absolute mess. Sentiment:



The uprising of Large Language Models

Early pretrained LMs (BERT, RoBERTa, GPT) were mostly fine-tuned for downstream task

- Add a classification head on top of LM and train it on labeled task data

Pretrain -> Finetune -> Predict

But GPT-3 showed a unique ability!

- Perform the task by generating natural language tokens, *aka*, **prompting**

That movie was great. Sentiment: Positive

It

This is an absolute mess. Sentiment:

Pretrain -> Prompt -> Predict



2018

GPT-1 (117M)
GPT-2 (1.5B)

2020

GPT-3 (175B)

2022

GPT-3.5 (175B)
ChatGPT (175B)

2023

GPT-4 (~100T)



2018

2019

Bert-base (110M)
Bert-Large (340M)
RoBERTA (354M)

2021

LaMDA (137B)
Gopher (280B)
GLaM (1.1T)

2022

Med-PaLM (540B)
U-PaLM (540B)
Flan-PaLM (540B)
Minerva (540B)



2018

2021

FairSeq (13B)

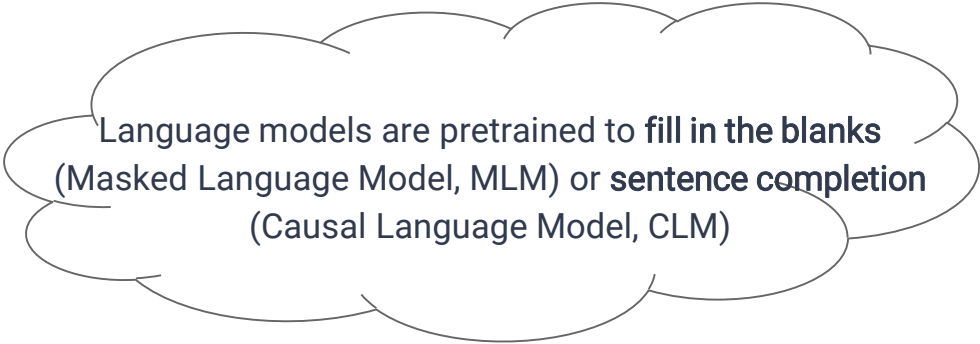
2022

Galactica (120B)
OPT-IML (175B)

2023

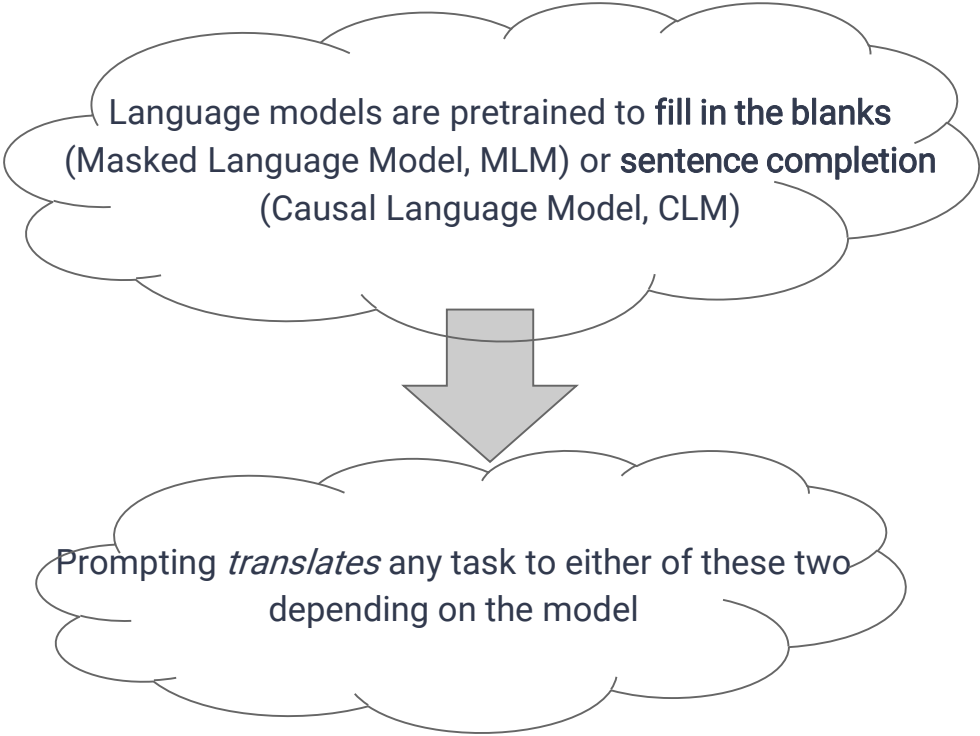
LLaMa (65B)

Prompting: A quick recap



Language models are pretrained to **fill in the blanks**
(Masked Language Model, MLM) or **sentence completion**
(Causal Language Model, CLM)

Prompting: A quick recap

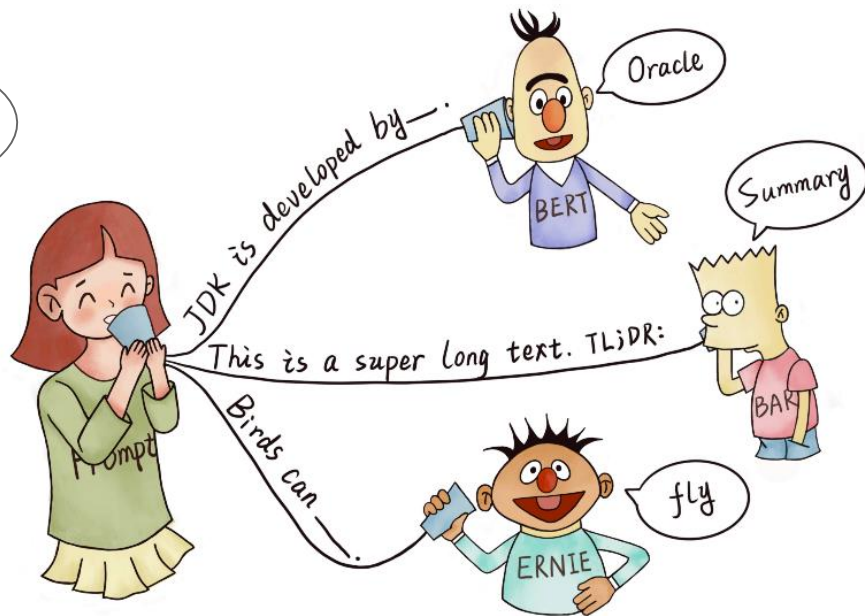
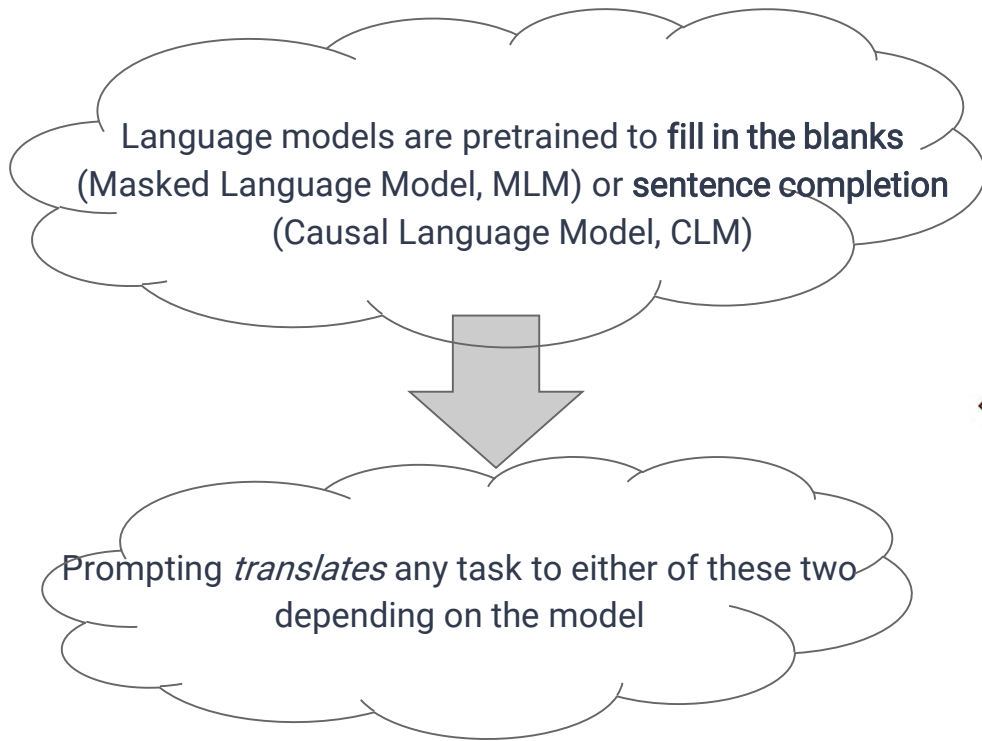


Language models are pretrained to **fill in the blanks** (Masked Language Model, MLM) or **sentence completion** (Causal Language Model, CLM)

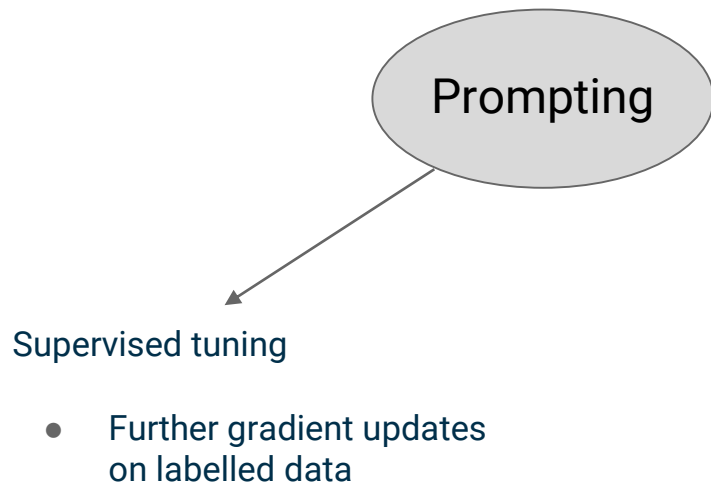
A large grey arrow points from the top cloud to the bottom cloud.

Prompting *translates* any task to either of these two depending on the model

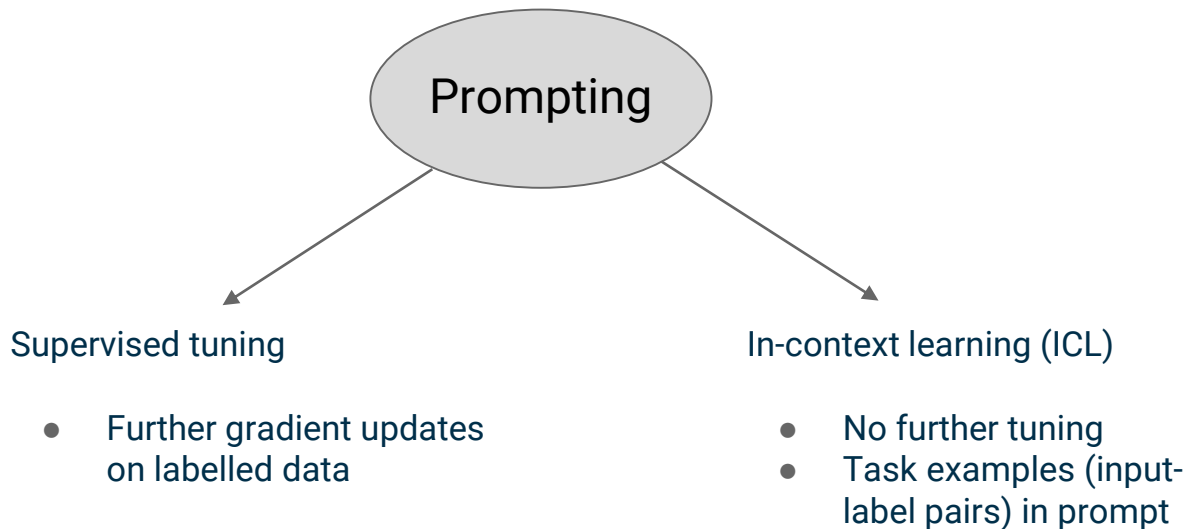
Prompting: A quick recap



Prompting: A quick recap



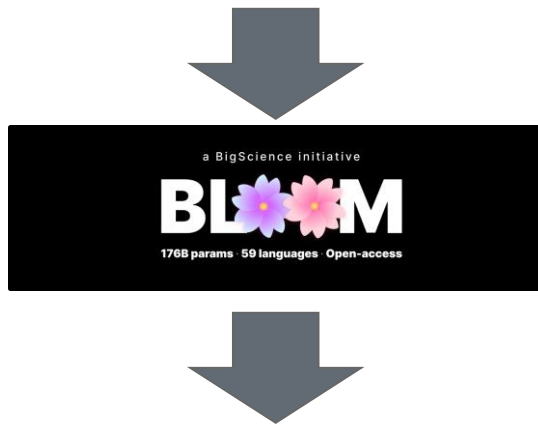
Prompting: A quick recap



Optimal prompts: An open problem

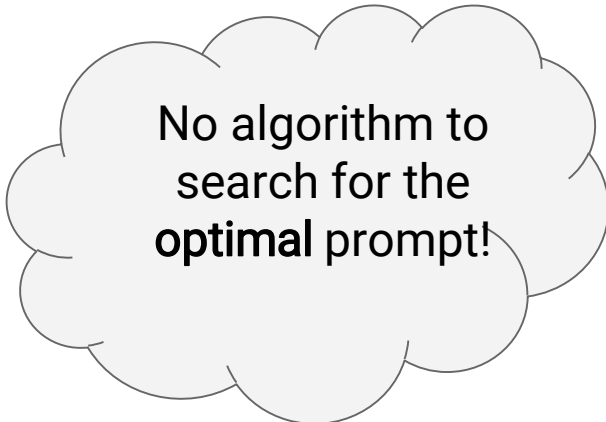
No algorithm to
search for the
optimal prompt!

Input: That movie was really great if you are dumb beyond imagination.
Sentiment expressed in the above sentence is



Input: That movie was really great if you are dumb beyond imagination.
Sentiment expressed in the above sentence is on the positive side of the range.

Optimal prompts: An open problem



No algorithm to
search for the
optimal prompt!

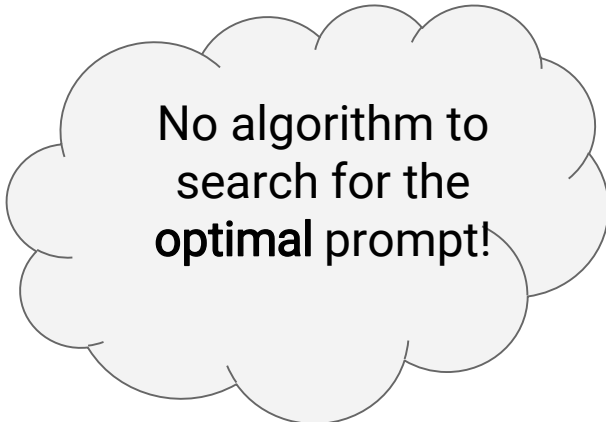
Consider the following sentence: “That movie was really great if you are dumb beyond imagination.” The author’s opinion expressed towards the movie in this sentence is



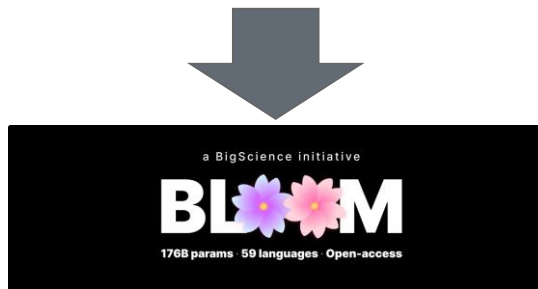
Consider the following sentence: “That movie was really great if you are dumb beyond imagination.” The author’s opinion expressed towards the movie in this sentence is **neutral**.

Optimal prompts: An open problem

Consider the following sentence: "That movie was really great if you are dumb beyond imagination." In the author's opinion, the movie is



No algorithm to
search for the
optimal prompt!



Consider the following sentence: "That movie was really great if you are dumb beyond imagination." In the author's opinion, the movie is **not good**.

Prompting as an alignment problem

Our hypothesis

Prompting works better when the task is similar to the pretraining objective/data

- MLM-based models are better suited to predict relation between pair of sentences

Prompting as an alignment problem

Our hypothesis

Prompting works better when the task is similar to the pretraining objective/data

- MLM-based models are better suited to predict relation between pair of sentences
- CLM-based models trained on codes (e.g., Codex) are superior to solve reasoning problems by generating codes

Prompting as an alignment problem

Our hypothesis

Prompting works better when the task is similar to the pretraining objective/data

- MLM-based models are better suited to predict relation between pair of sentences
- CLM-based models trained on codes (e.g., Codex) are superior to solve reasoning problems by generating codes

Our findings

Special unsupervised finetuning followed by prompt tuning aligned to the task

Dutta et al., ACL 2022

Prompting as an alignment problem

Our hypothesis

Prompting works better when the task is similar to the pretraining objective/data

- MLM-based models are better suited to predict relation between pair of sentences
- CLM-based models trained on codes (e.g., Codex) are superior to solve reasoning problems by generating codes

Our findings

Special unsupervised finetuning followed by prompt tuning aligned to the task

Dutta et al., ACL 2022

Aligning cross-lingual prompt-labels both in input and label space for in-context learning

Ongoing

Prompting as an alignment problem

Our hypothesis

Prompting works better when the task is similar to the pretraining objective/data

- MLM-based models are better suited to predict relation between pair of sentences
- CLM-based models trained on codes (e.g., Codex) are superior to solve reasoning problems by generating codes

Our findings

Special unsupervised finetuning followed by prompt tuning aligned to the task

Dutta et al., ACL 2022

Aligning cross-lingual prompt-labels both in input and label space for in-context learning

Ongoing

Argument Mining in a nutshell

Broadly, 2 tasks:

1. Identify the argument components (e.g, claims and premises)

EMTs, SAR, firefighters, police, etc. should receive “military discounts”. For those of you who don’t know, it’s common (at least in the US) for businesses, transit agencies, etc. to give small discounts to military veterans to thank them for their service. It seems that medical responders (even hospital staff, actually) and other emergency services do more good for society than soldiers and that such discounts should be given to them.

| other emergency services do more good for society than soldiers and that such discounts should be given to them.

What about teachers? Scientists? Doctors? Farmers? There are plenty of professions that do great things for society but people aren’t going around thanking them with discounts.

Argument Mining in a nutshell

Broadly, 2 tasks:

1. Identify the argument components (claims and premises)
2. Identify the relation between those

EMTs, SAR, firefighters, police, etc. should receive “military discounts”. For those of you who don’t know, it’s common (at least in the US) for businesses, transit agencies, etc. to give small discounts to military veterans to thank them for their service. It seems that medical responders (even hospital staff, actually) and other emergency services do more good for society than soldiers and that such discounts should be given to them.

Undercutter

Can we leverage large unlabeled data to train the model specific to the task?

Issues

- Very limited labeled data for finetuning
- No unified dataset for both the tasks

Relation Types

support
agreement
direct attack
undercutter attack
partial

Transfer Learning via Selective MLM (sMLM)

Discussion threads in *r/ChangeMyView*
subreddit

- Large, open source of public
argumentation

Transfer Learning via Selective MLM (sMLM)

Discussion threads in *r/ChangeMyView*
subreddit

- Large, open source of public argumentation
- Markers like *IMHO/IMO* helps identify argumentation signals

Transfer Learning via Selective MLM (sMLM)

Discussion threads in *r/ChangeMyView* subreddit

- Large, open source of public argumentation
- Markers like *IMHO/IMO* helps identify argumentation signals

Hypothesis: Can we make a pretrained language model aware of argumentative discourse by making it predict such markers?

Transfer Learning via Selective MLM (sMLM)

u/DurianMD:

CMV: Religion is not violent or not violent, its followers are.

So, my belief is that while religion can inform the views of people, it is far more likely that religion will be used to justify actions that would have been executed any way. I think that most Jewish people don't want to stone adulterers and most Muslims don't want to stone non believers.

u/recycled_kevlar:

Your stance relies on the assumption that religion has no influence on the actions of its followers beyond the superficial. Yet something must exist that allows this pattern to occur. Ill narrow it down to religion or culture. So, you are correct if you assume the culture dominates the religion, and you are incorrect if the reverse is true. With this in mind, I think its safe to assume the truth is somewhere in between, with both the religion and the culture somehow influencing the unrest we see.

u/DurianMD:

I suppose I was taking a harsh stance when I assumed that religion had no effect on behavior, when it obviously does. I still think the culture dominates religion to a great extent, however I cannot ignore that religion does have an effect on culture to some extent.

Hypothesis: Can we make a pretrained language model aware of argumentative discourse by making it predict such markers?

Transfer Learning via Selective MLM (sMLM)

u/DurianMD:

CMV: Religion is not violent or not violent, its followers are.

So, my belief is that while religion can inform the views of people, it is far more likely that religion will be used to justify actions that would have been executed any way. I think that most Jewish people don't want to stone adulterers and most Muslims don't want to stone non believers.

u/recycled_kevlar:

Your stance relies on the assumption that religion has no influence on the actions of its followers beyond the superficial. Yet something must exist that allows this pattern to occur. Ill narrow it down to religion or culture. So, you are correct if you assume the culture dominates the religion, and you are incorrect if the reverse is true. With this in mind, I think its safe to assume the truth is somewhere in between, with both the religion and the culture somehow influencing the unrest we see.

u/DurianMD:

I suppose I was taking a harsh stance when I assumed that religion had no effect on behavior, when it obviously does.

I still think the culture dominates religion to a great extent, however I cannot ignore that religion does have an effect on culture to some extent.

Hypothesis: Can we make a pretrained language model aware of argumentative discourse by making it predict such markers?

Transfer Learning via Selective MLM (sMLM)

u/DurianMD:

CMV: Religion is not violent or not violent, its followers are.

So, my belief is that while religion can inform the views of people, it is far more likely that religion will be used to justify actions that would have been executed any way. I think that most Jewish people don't want to stone adulterers and most Muslims don't want to stone non believers.

u/recycled_kevlar:

Your stance relies on the assumption that religion has no influence on the actions of its followers beyond the superficial. Yet something must exist that allows this pattern to occur. Ill

narrow it down to religion or culture. So, you are correct if you assume the culture dominates the religion, and you are incorrect if the reverse is true. With this in mind, I think its safe to assume the truth is somewhere in between, with both the religion and the culture somehow influencing the unrest we see.

u/DurianMD:

I suppose I was taking a harsh stance when I assumed that religion had no effect on behavior, when it obviously does.

I still think the culture dominates religion to a great extent, however I cannot ignore that religion does have an effect on culture to some extent.

Hypothesis: Can we make a pretrained language model aware of argumentative discourse by making it predict such markers?

Transfer Learning via Selective MLM (sMLM)

u/DurianMD:

CMV: Religion is not violent or not violent, its followers are.

So, my belief is that while religion can inform the views of people, it is far more likely that religion will be used to justify actions that would have been executed any way. I think that most Jewish people don't want to stone adulterers and most Muslims don't want to stone non believers.

u/recycled_kevlar:

Your stance relies on the assumption that religion has no influence on the actions of its followers beyond the superficial. Yet something must exist that allows this pattern to occur. Ill

narrow it down to religion or culture. So, you are correct if you assume the culture dominates the religion, and you are incorrect if the reverse is true. With this in mind, I think its safe to assume the truth is somewhere in between, with both the religion and the culture somehow influencing the unrest we see.

u/DurianMD:

I suppose I was taking a harsh stance when I assumed that religion had no effect on behavior, when it obviously does.

I still think the culture dominates religion to a great extent, however I cannot ignore that religion does have an effect on culture to some extent.

- Finetune a pretrained Transformer-based LM to predict these markers given the context

Transfer Learning via Selective MLM (sMLM)

u/DurianMD:

CMV: Religion is not violent or not violent, its followers are.

So, my belief is that while religion can inform the views of people, it is far more likely that religion will be used to justify actions that would have been executed any way. I think that most Jewish people don't want to stone adulterers and most Muslims don't want to stone non believers.

u/recycled_kevlar:

Your stance relies on the assumption that religion has no influence on the actions of its followers beyond the superficial. Yet something must exist that allows this pattern to occur. Ill

narrow it down to religion or culture. So, you are correct if you assume the culture dominates the religion, and you are incorrect if the reverse is true. With this in mind, I think its safe to assume the truth is somewhere in between, with both the religion and the culture somehow influencing the unrest we see.

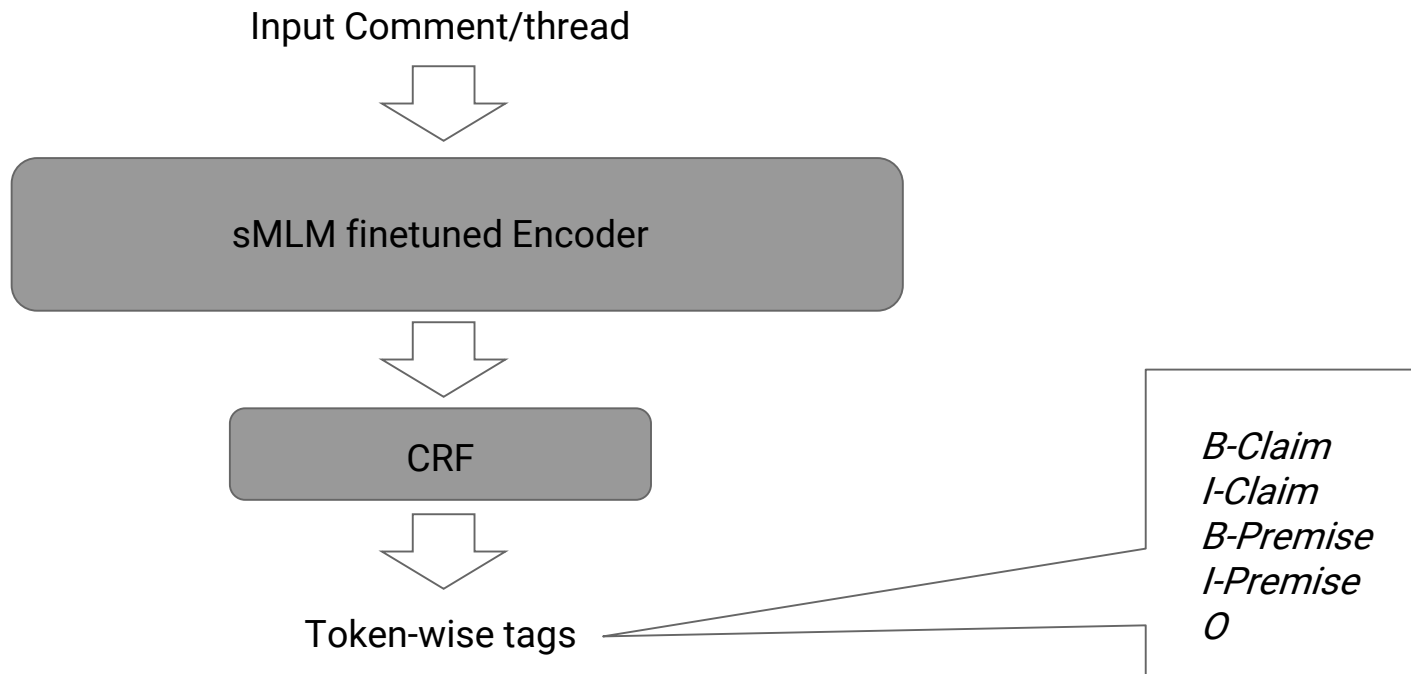
u/DurianMD:

I suppose I was taking a harsh stance when I assumed that religion had no effect on behavior, when it obviously does.

I still think the culture dominates religion to a great extent, however I cannot ignore that religion does have an effect on culture to some extent.

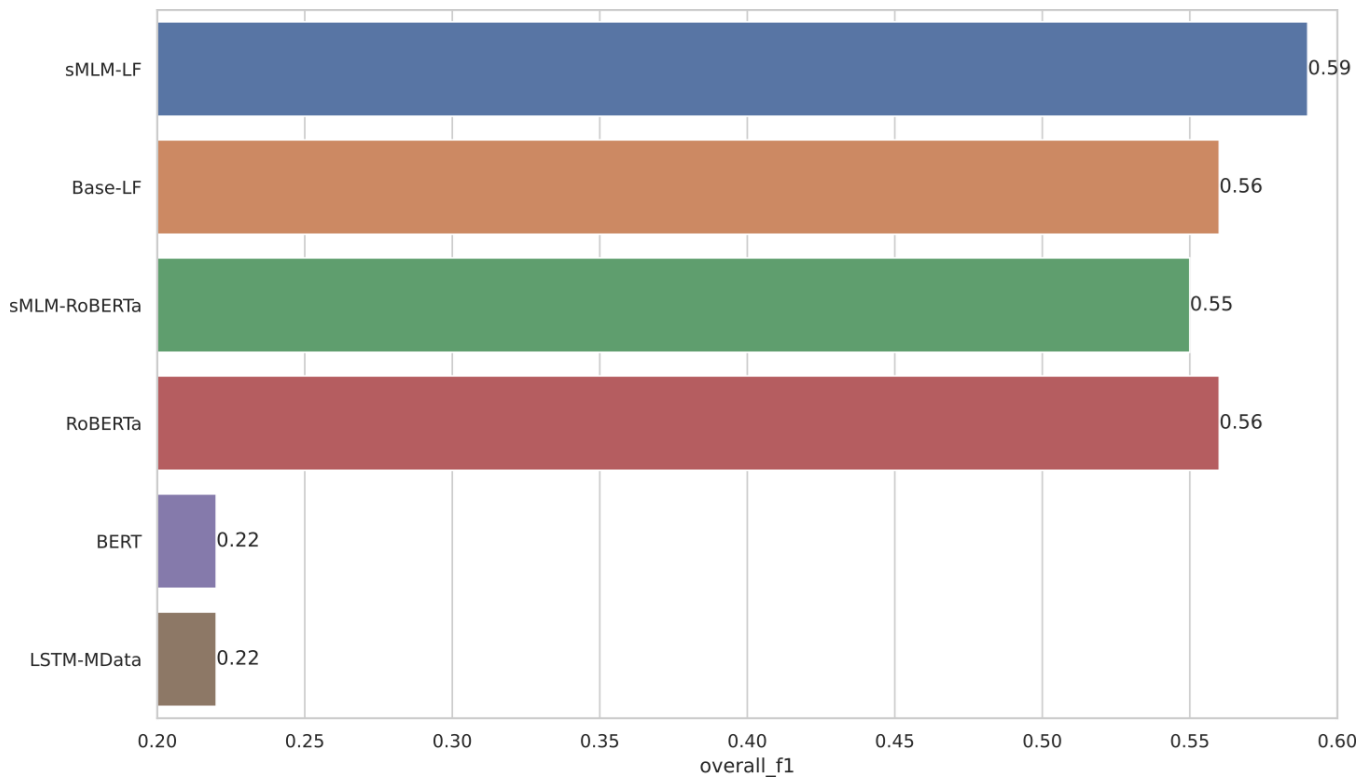
- Finetune a pretrained Transformer-based LM to predict these markers given the context and use for downstream tasks
- Incorporate complete thread context with Longformer
 - Replace user names with special tokens and global attention

Argument Component Identification



Argument Component Identification (ACT)

Performance on Reddit
CMV-modes dataset for
token-level component
prediction



Inter-component Relation Type Prediction

Prompt-based method

- Leverages upon the MLM pretraining/finetuning of Transformer LMs

Inter-component Relation Type Prediction

Prompt-based method

- Leverages upon the MLM pretraining/finetuning of Transformer LMs
- sMLM tuned the LM to predict 'relation signaling' tokens among spans
 - Naturally aligns to Relation Type prediction

Inter-component Relation Type Prediction

Prompt-based method

- Leverages upon the MLM pretraining/finetuning of Transformer LMs
- sMLM tuned the LM to predict 'relation signaling' tokens among spans
 - Naturally aligns to Relation type prediction

USER-1 CMV: I feel skill is largely determined by experience. Compliments on skill are almost meaningless. In high school, I thought I was "good at math" as I'm the son of a math teacher and electrical engineer. In college, I learned that math was not something you're "good at" but something you have to put hard work into and is almost the sole determiner in the level of skill you obtain.

So then isn't almost any compliment almost to be expected? I've spent a lot of time with similar problems -- how could I not know all the details and little tricks of these problems? I feel a compliment recognizes something given: I feel everyone is passionate about something, whether it be math or psychology or medicine. I don't hear "you're so good at biology" but I think I should.

USER-2 Then wouldn't a complement be just an acknowledgement of the time and effort you put into something that most people see as hard or worthwhile? This implies the complement is meaningful.

(Most people don't do this - either they don't put the time and effort into something generally hard or worthwhile or the time and effort isn't hard or worthwhile .)

Inter-component Relation Type Prediction

Prompt-based method

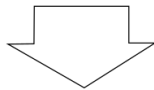
- Leverages upon the MLM pretraining/finetuning of Transformer LMs
- sMLM tuned the LM to predict 'relation signaling' tokens among spans
 - Naturally aligns to Relation Type Prediction

USER-1 CMV: I feel skill is largely determined by experience. Compliments on skill are almost meaningless. In high school, I thought I was "good at math" as I'm the son of a math teacher and electrical engineer. In college, I learned that math was not something you're "good at" but something you have to put hard work into and is almost the sole determiner in the level of skill you obtain.

So then isn't almost any compliment almost to be expected? I've spent a lot of time with similar problems -- how could I not know all the details and little tricks of these problems? I feel a compliment recognizes something given: I feel everyone is passionate about something, whether it be math or psychology or medicine. I don't hear "you're so good at biology" but I think I should.

USER-2 Then wouldn't a complement be just an acknowledgement of the time and effort you put into something that most people see as hard or worthwhile? This implies the complement is meaningful.

(Most people don't do this - either they don't put the time and effort into something generally hard or worthwhile or the time and effort isn't hard or worthwhile .)



Create prompt from
thread

<thread token sequence> **USER-1** said <component-1> [MASK][MASK][MASK] **USER-2** said <component-2>

Inter-component Relation Type Prediction

Prompt-based method

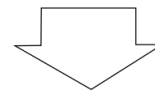
- Leverages upon the MLM pretraining/finetuning of Transformer LMs
- sMLM tuned the LM to predict 'relation signaling' tokens among spans
 - Naturally aligns to relation-type prediction

USER-1 CMV: I feel skill is largely determined by experience. Compliments on skill are almost meaningless. In high school, I thought I was "good at math" as I'm the son of a math teacher and electrical engineer. In college, I learned that math was not something you're "good at" but something you have to put hard work into and is almost the sole determiner in the level of skill you obtain.

So then isn't almost any compliment almost to be expected? I've spent a lot of time with similar problems -- how could I not know all the details and little tricks of these problems? I feel a compliment recognizes something given: I feel everyone is passionate about something, whether it be math or psychology or medicine. I don't hear "you're so good at biology" but I think I should.

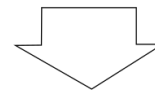
USER-2 Then wouldn't a complement be just an acknowledgement of the time and effort you put into something that most people see as hard or worthwhile? This implies the complement is meaningful.

(Most people don't do this - either they don't put the time and effort into something generally hard or worthwhile or the time and effort isn't hard or worthwhile .)



Create prompt from thread

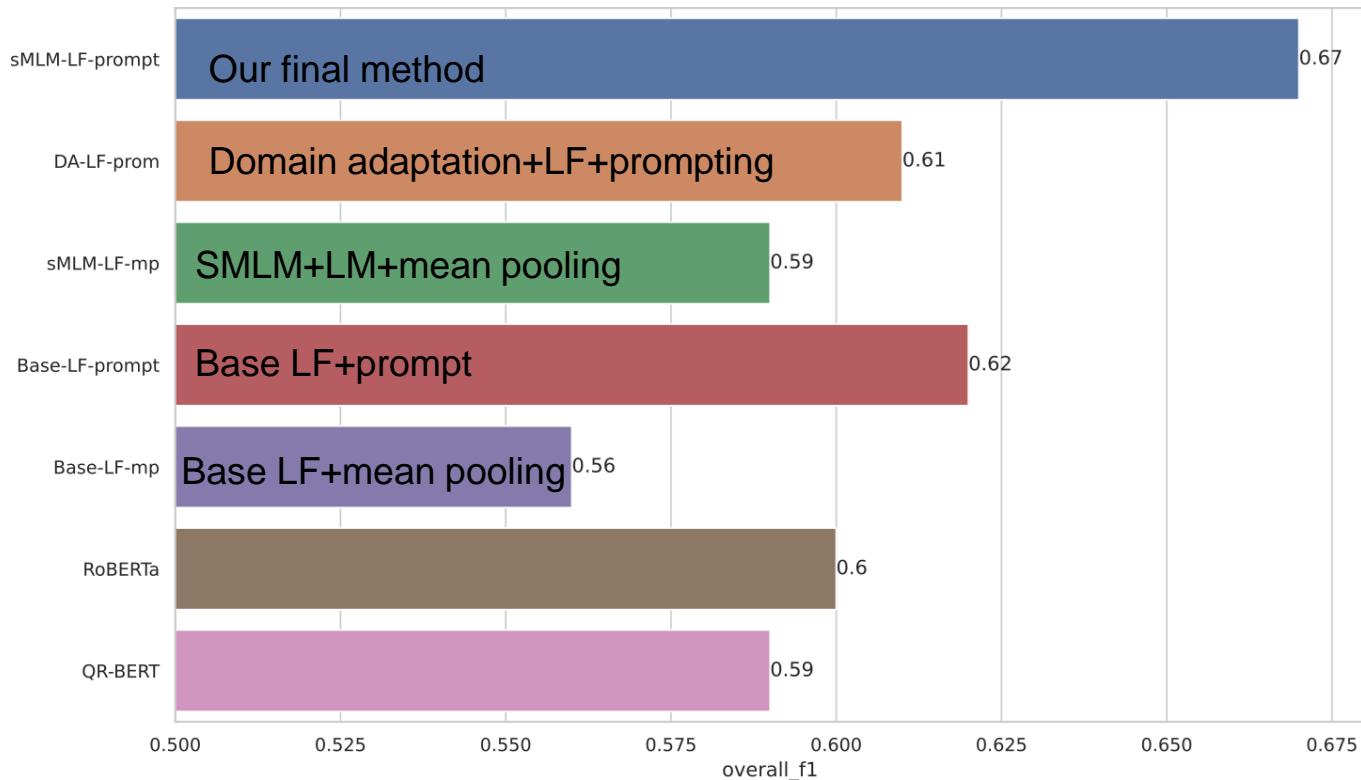
<thread token sequence> **USER-1** said <component-1> [MASK][MASK][MASK] **USER-2** said <component-2>



sMLM-finetuned LM encodes the prompt and takes concatenated output at [MASK] positions

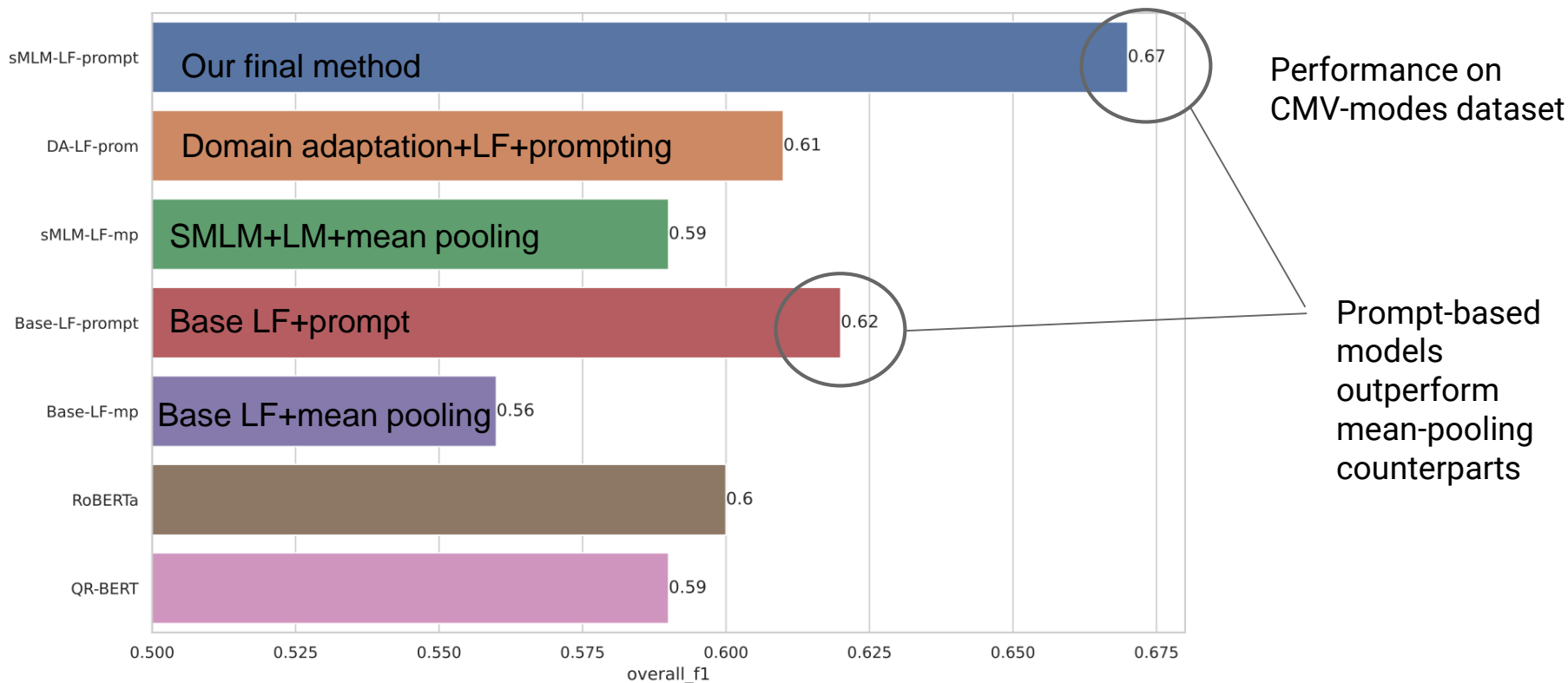
Classify relation between <component-1> and <component-2>

Inter-component Relation Type Prediction

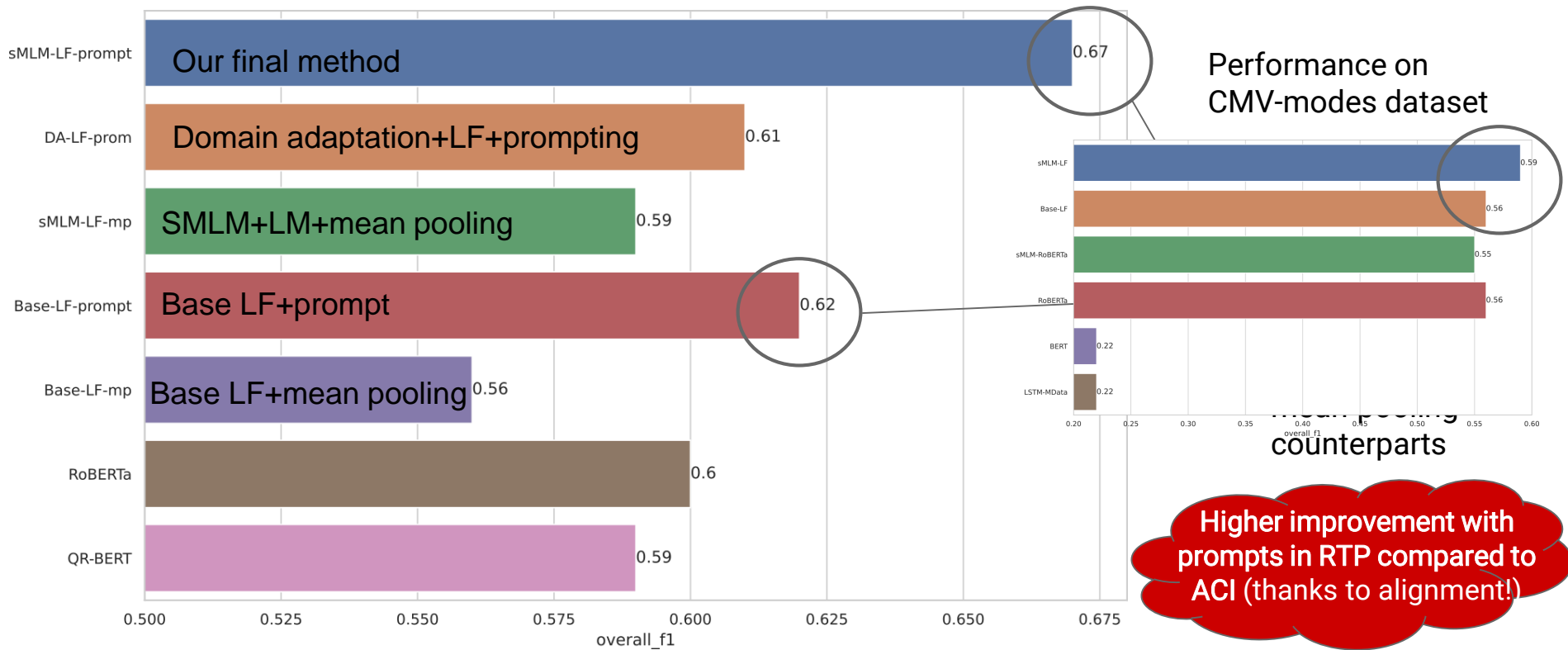


Performance on
CMV-modes dataset

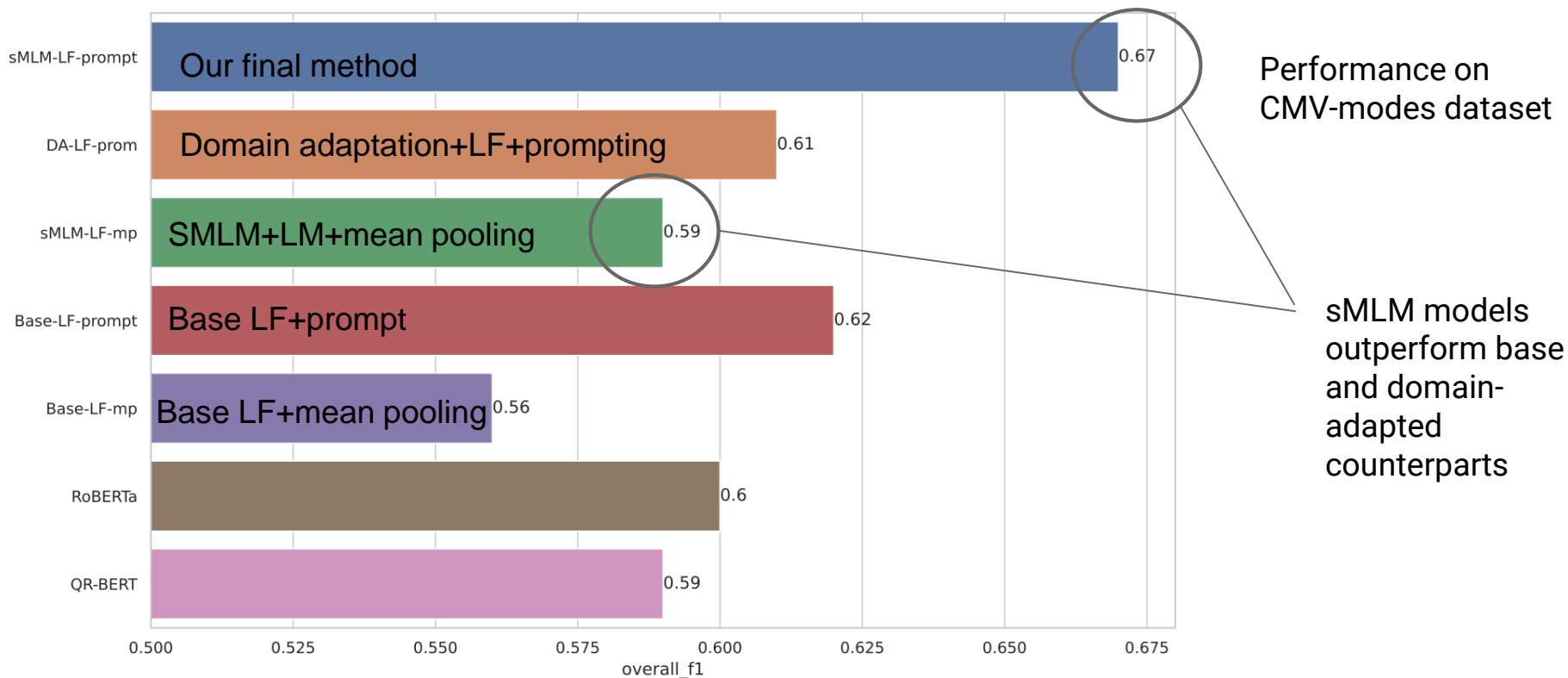
Inter-component Relation Type Prediction



Inter-component Relation Type Prediction



Inter-component Relation Type Prediction (RTP)



Prompting as an alignment problem

Our hypothesis

Prompting works better when the task is similar to the pretraining objective/data

- MLM-based models are better suited to predict relation between pair of sentences
- CLM-based models trained on codes (e.g., Codex) are superior to solve reasoning problems by generating codes

Our findings

Special unsupervised finetuning followed by prompt tuning specific to the task

Dutta et al., ACL 2022

Aligning cross-lingual prompt-labels both in input and label space for in-context learning

Ongoing

In-context learning (ICL)

In-context learning pipeline

1. **Pretraining documents** are conditioned on a **latent concept** (e.g., biographical text)



2. Create **independent examples** from a **shared concept**. If we focus on full names, wiki bios tend to relate them to nationalities.



Input (x)	Output (y)	Delimiter
Albert Einstein was	German	\n
Mahatma Gandhi was	Indian	\n
Marie Curie was	?	...brilliant? ...Polish?

3. **Concatenate examples into a prompt** and predict next word(s). **Language model (LM)** implicitly **infers the shared concept** across examples despite the unnatural concatenation

Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was

LM

Polish

In-context learning (ICL)

Shared concept:

1. Context and test input are 'similar' in some sense
2. The notion of the task should be shared between the two

1. Pretraining documents are conditioned on a **latent concept** (e.g., biographical text)



2. Create independent examples from a **shared concept**. If we focus on full names, wiki bios tend to relate them to nationalities.



Albert Einstein was a German theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is best known for developing the theory of relativity, but he also

Input (x)	Output (y)	Delimiter
Albert Einstein was	German	\n
Mahatma Gandhi was	Indian	\n
Marie Curie was	?	...brilliant? ...Polish?

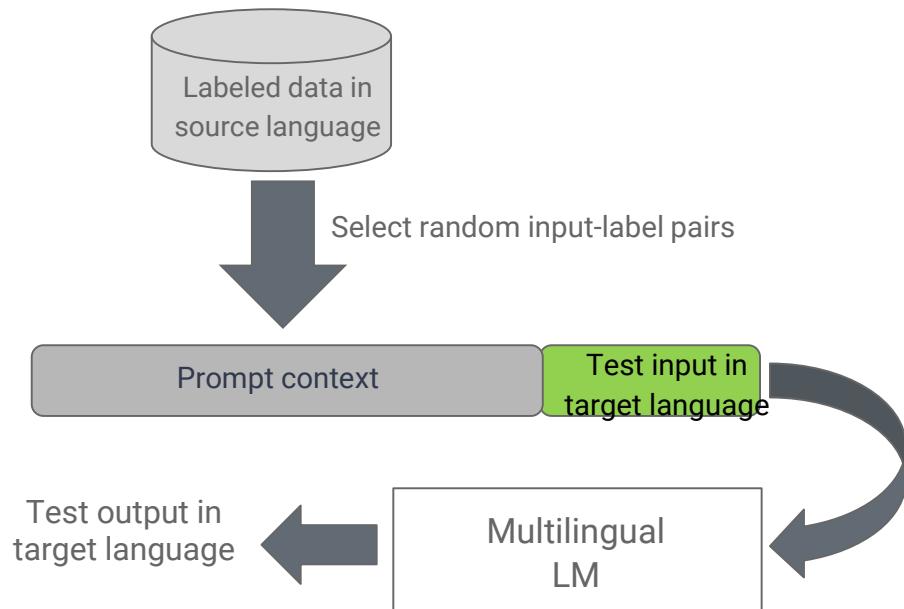
3. Concatenate examples into a prompt and predict next word(s). **Language model (LM)** implicitly infers the **shared concept** across examples despite the unnatural concatenation

Albert Einstein was German \n Mahatma Gandhi was Indian \n Marie Curie was

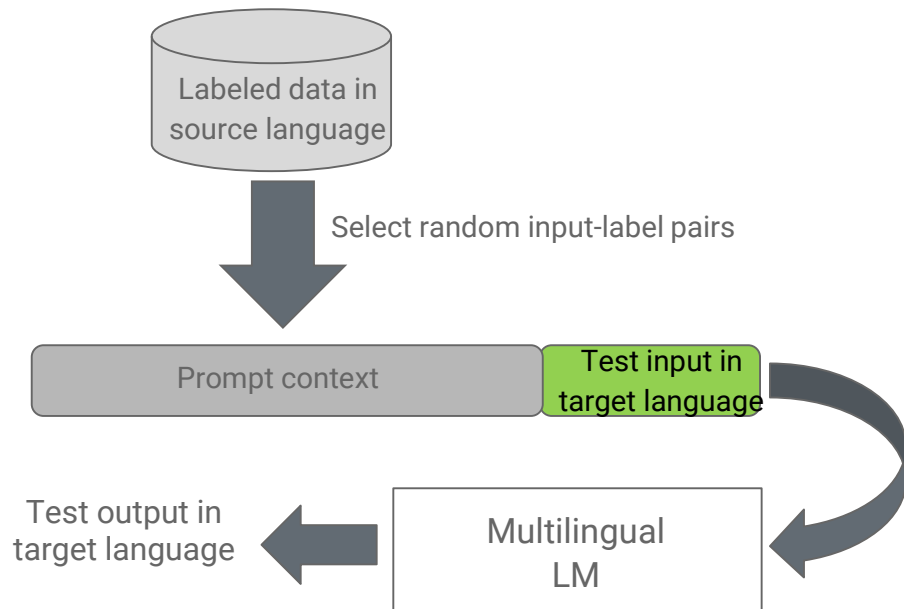
LM

Polish

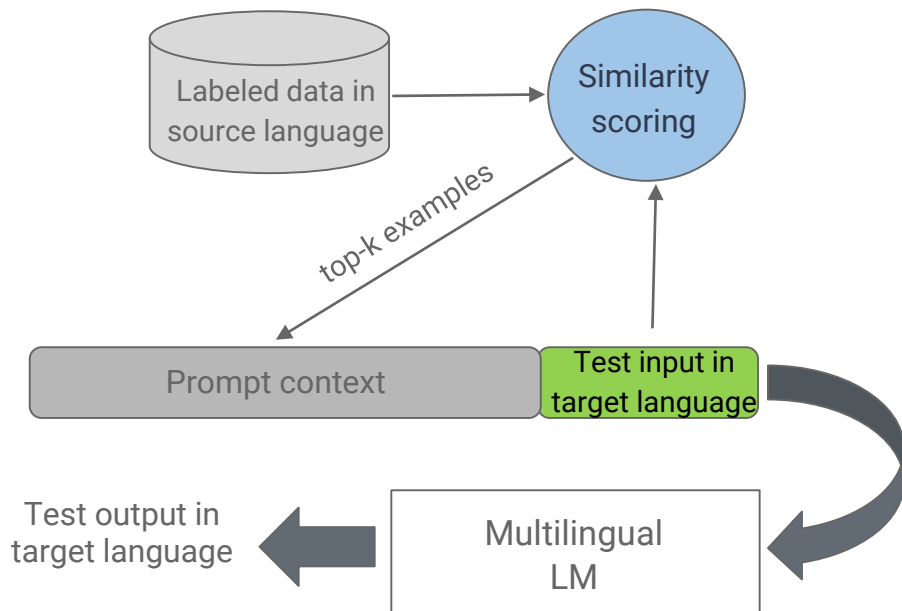
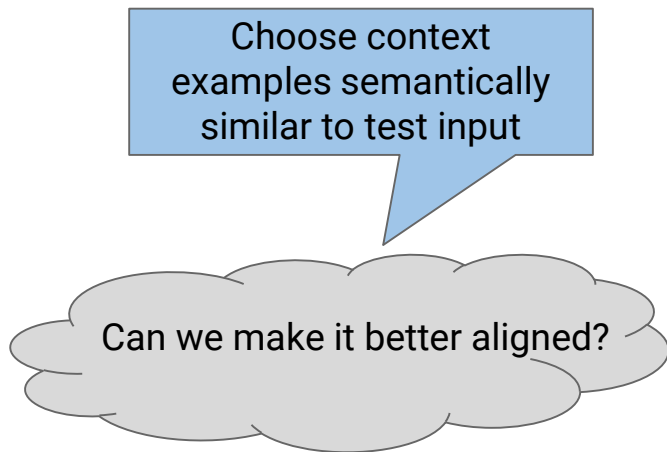
Cross-lingual ICL: random prompts



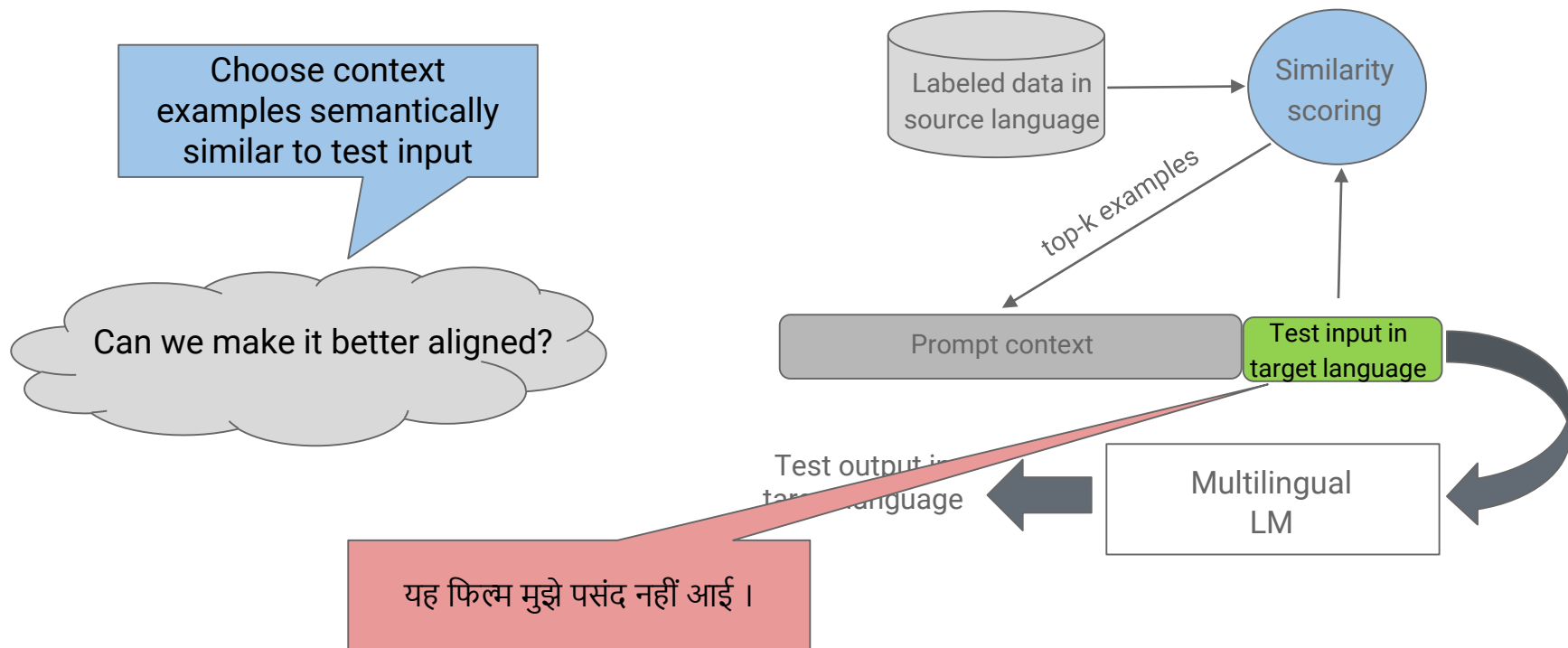
Cross-lingual ICL: random prompts



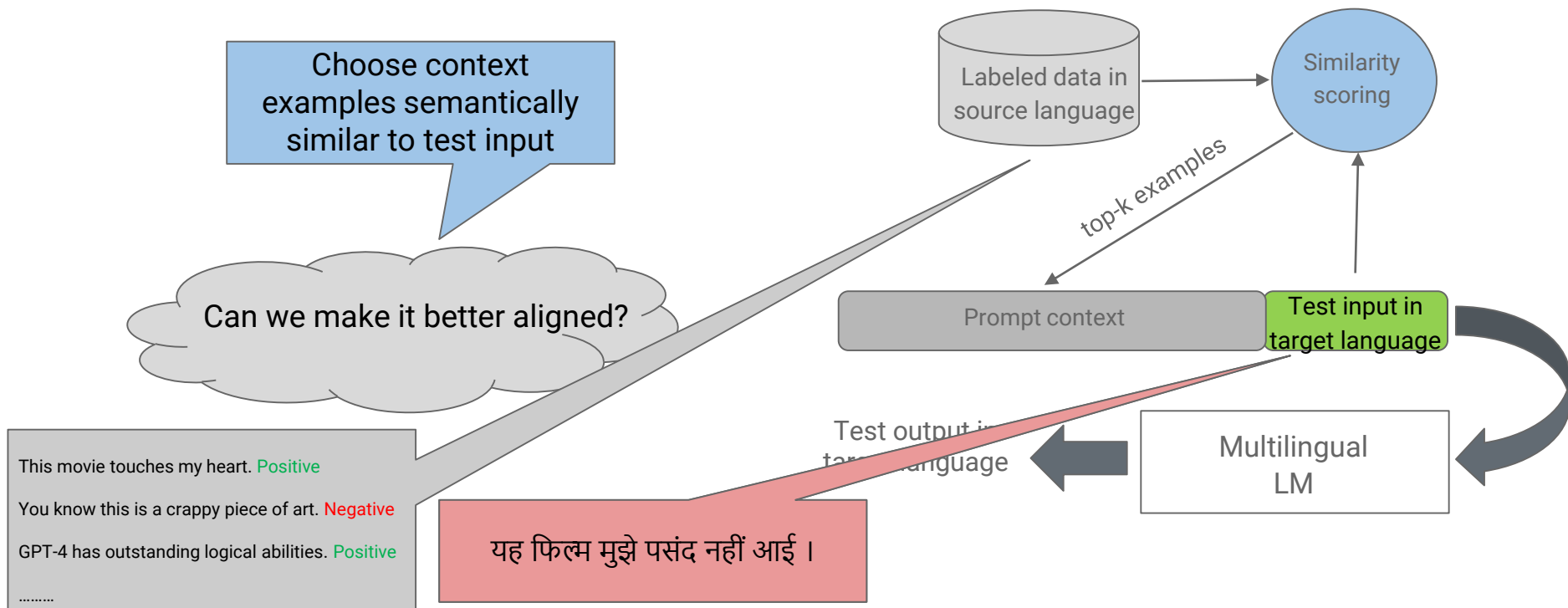
Cross-lingual ICL: semantic alignment



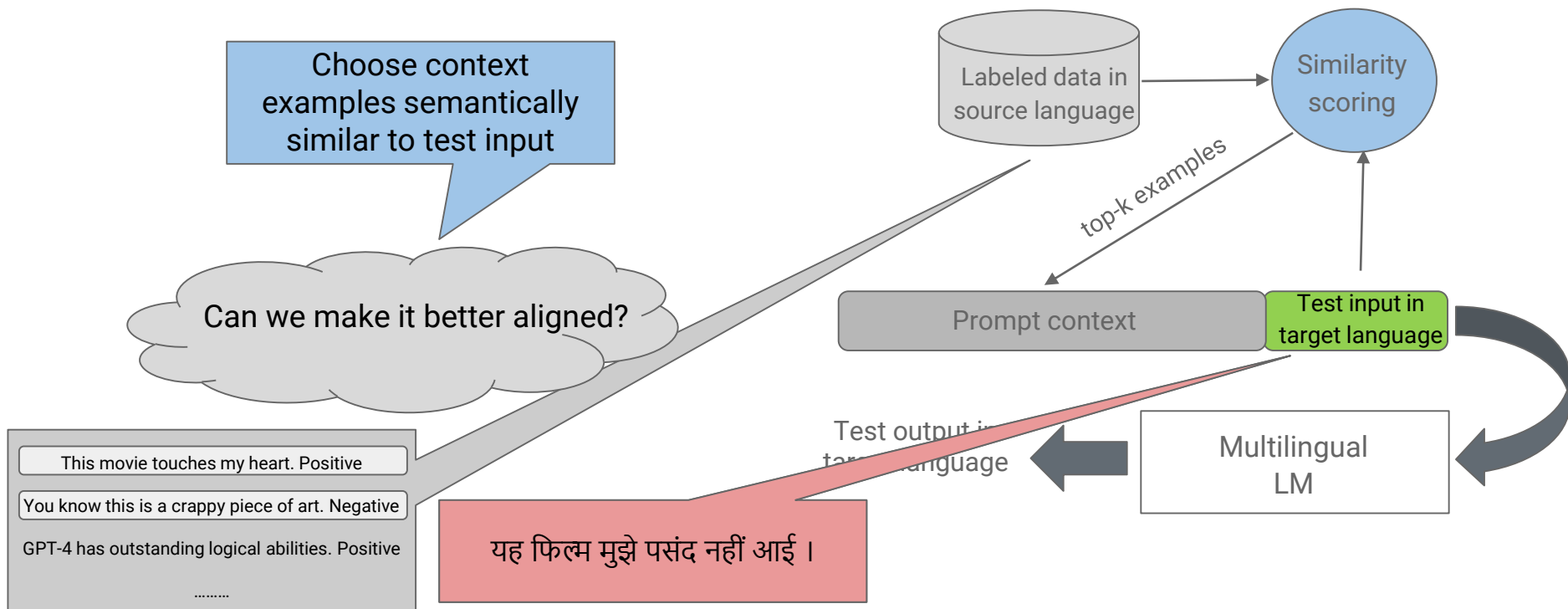
Cross-lingual ICL: semantic alignment



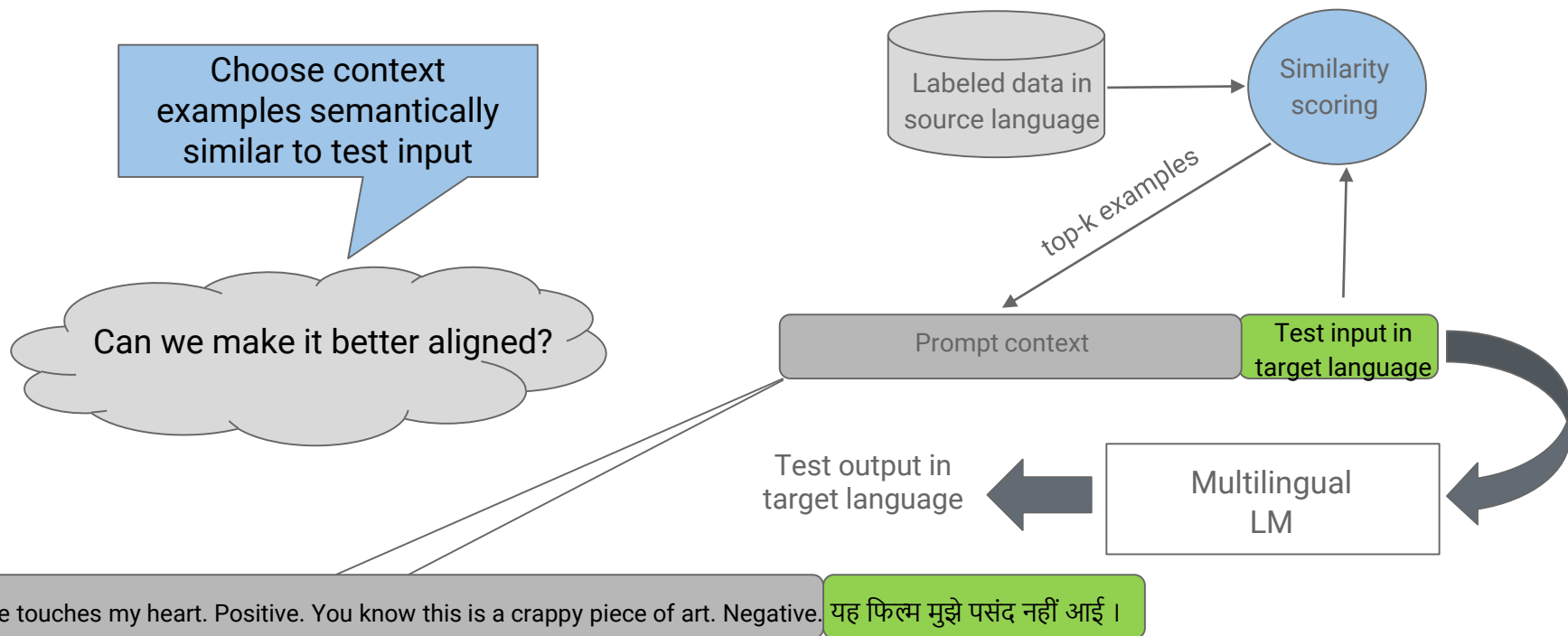
Cross-lingual ICL: semantic alignment



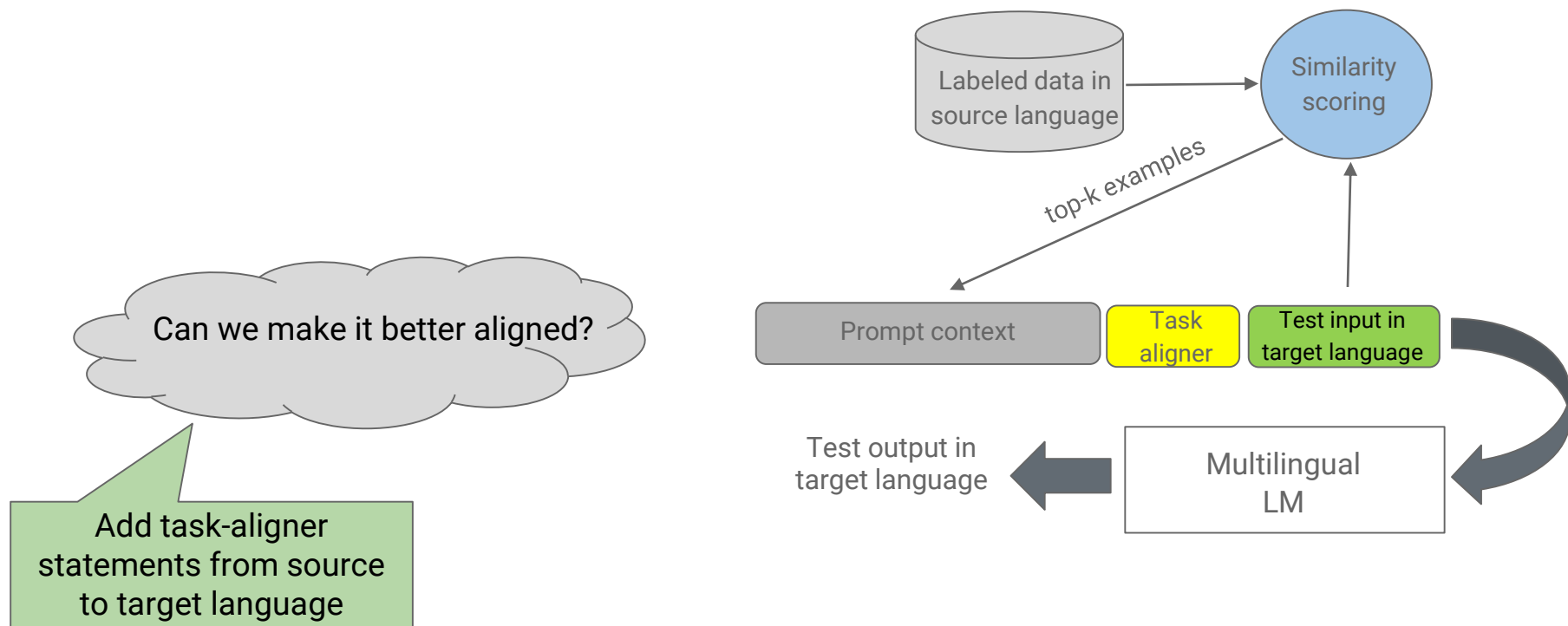
Cross-lingual ICL: semantic alignment



Cross-lingual ICL: semantic alignment



Cross-lingual ICL: task alignment

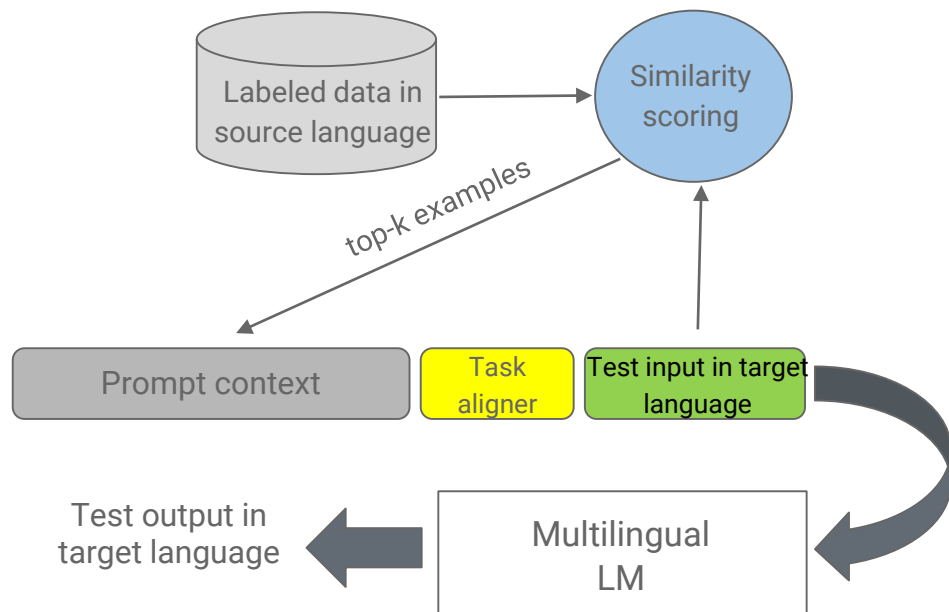


Cross-lingual ICL: task alignment

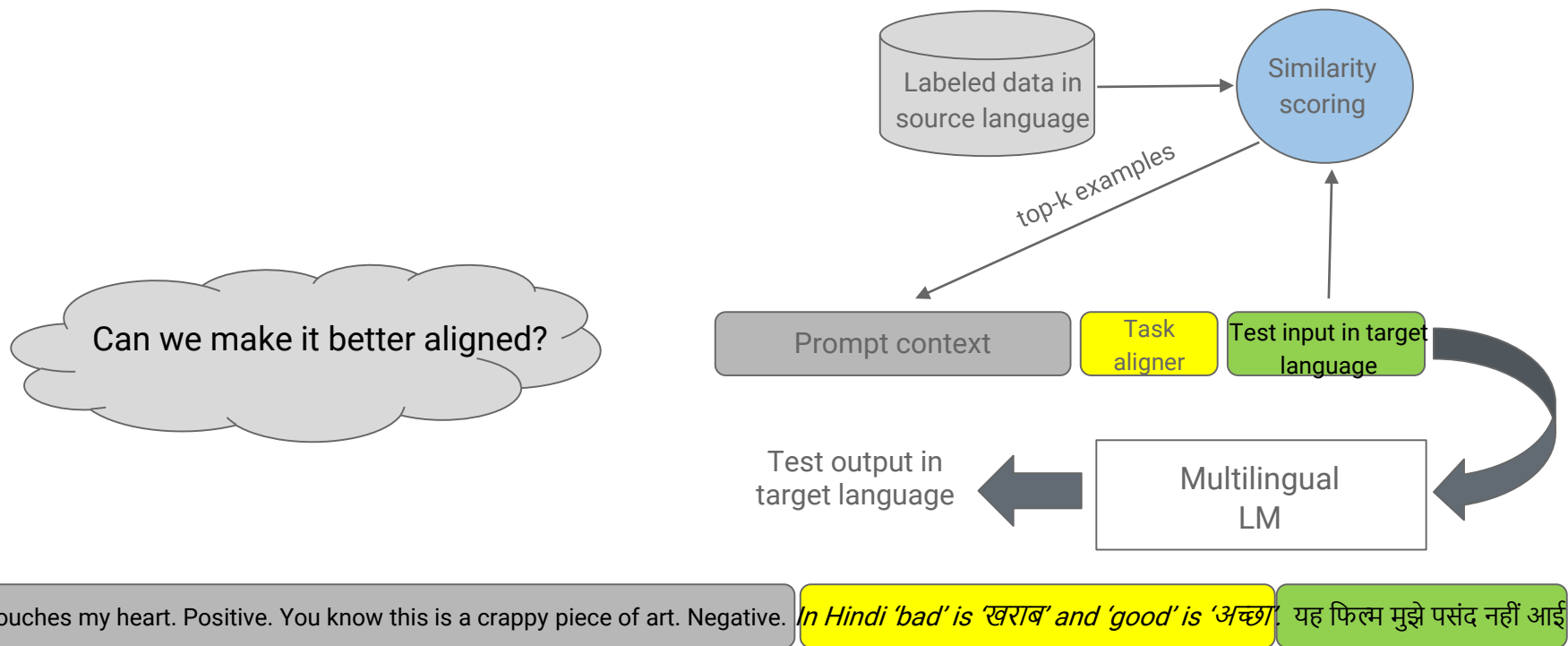
Task alignment example	
Task: Sentiment classification	
Source: English	Target: Hindi
Aligner: In Hindi 'bad' is 'खराब' and 'good' is 'अच्छा'	

Can we make it better aligned?

Add task-aligner statements from source to target language



Cross-lingual ICL: task alignment



Alignment improves ICL!

LM used: XGLM (7.5B params)

Datasets used:

- Multilingual Amazon Reviews Corpus (MARC)
- Cross-language sentiment classification (CLS)
- HatEval

Languages:

En -> English, De -> German, Es -> Spanish, Fr -> French, Ja -> Japanese, Zh -> Mandarin

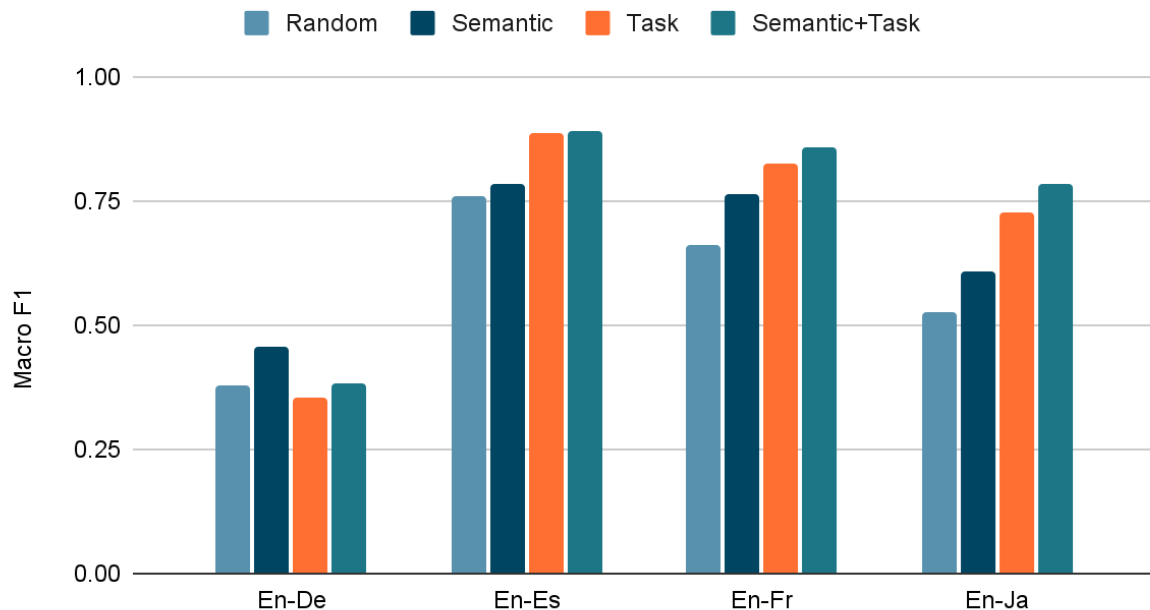
MARC dataset



Alignment improves ICL!

In most language pairs, both types of alignments bring improvements

MARC dataset

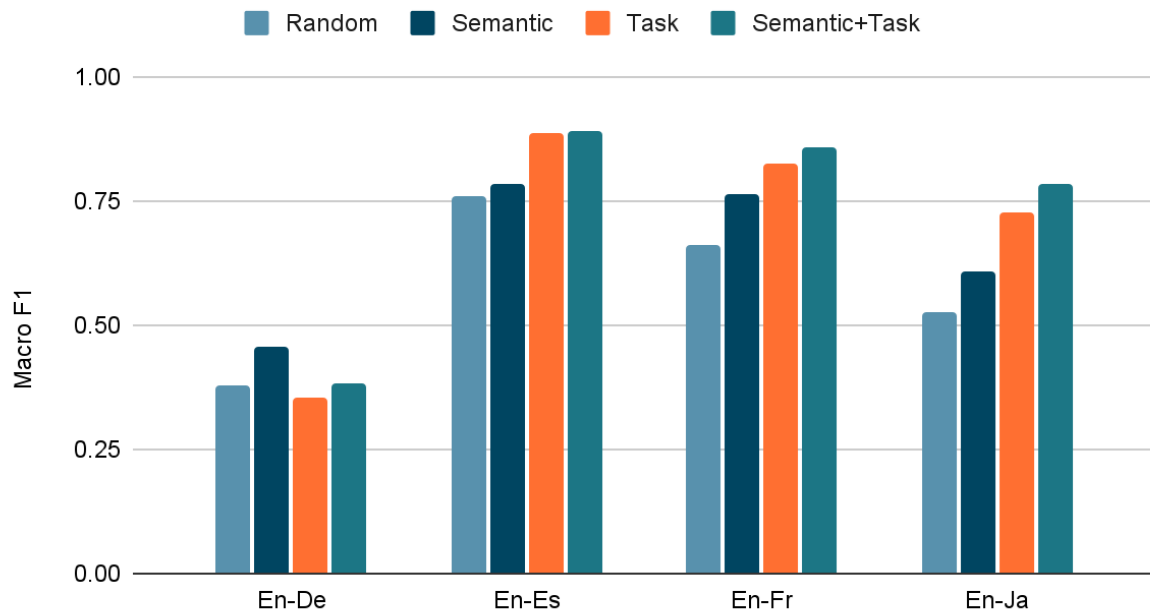


Alignment improves ICL!

In most language pairs, both types of alignments bring improvements

- Exception: when target language is German or Mandarin

MARC dataset

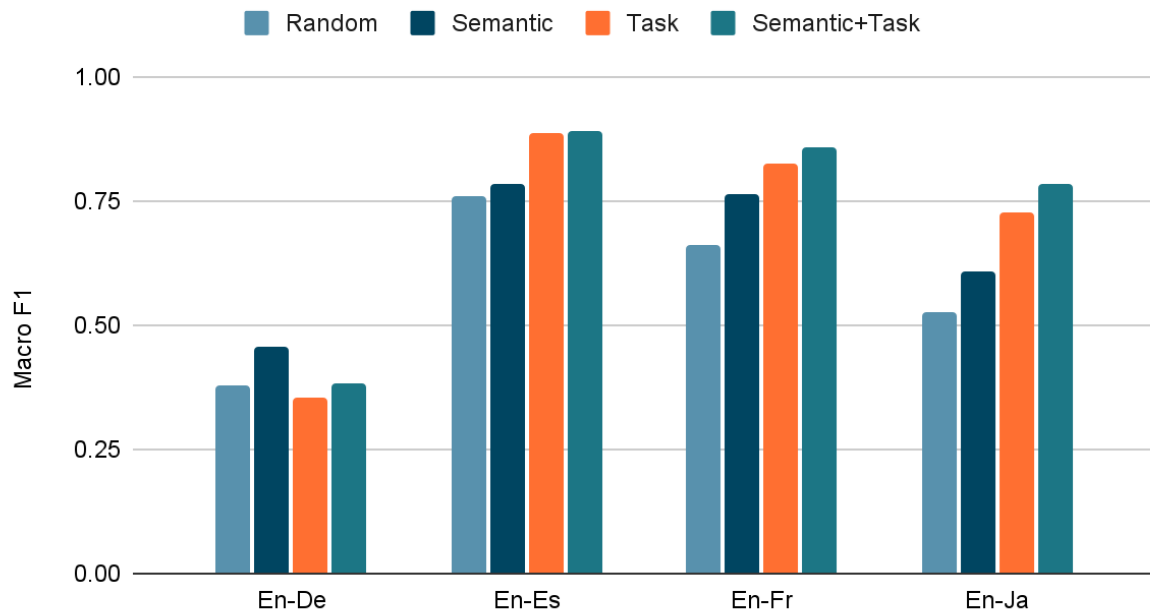


Alignment improves ICL!

In most language pairs, both types of alignments bring improvements

- Exception: when target language is German or Mandarin
- However, with XGLM 2.7B parameter model, these two languages show improvement with alignment

MARC dataset



Why does Task Alignment Work?

Setup \ Target language	de	en	es	fr	ja	zh
Random prompt	0.345	0.633	0.731	0.557	0.499	0.462
Uniform label space	0.441	0.570	0.493	0.414	0.483	0.594
Task alignment by language information only	0.346	0.645	0.733	0.575	0.543	0.508
Task alignment via third language	0.345	0.687	0.755	0.673	0.601	0.423
Incorrect task alignment	0.338	0.665	0.787	0.647	0.544	0.339
Task Alignment	0.338	0.722	0.830	0.758	0.730	0.335

Understanding how task alignment works. Average F1-Macro across all source-target pairs on MARC.

Automated aligner generation using mT5

Setup \ Target	MARC					CLS			HatEval
	de	es	fr	ja	zh	de	fr	ja	es
Random prompting	0.380	0.761	0.663	0.526	0.362	0.682	0.412	0.609	0.435
Semantic alignment	0.458	0.783	0.762	0.608	0.450	0.677	0.505	0.691	0.493
Task-based alignment	0.355	0.888	0.826	0.727	0.333	0.620	0.696	0.752	0.499
Automated aligner	0.531	0.792	0.699	0.599	0.350	0.721	0.430	0.610	0.438

Misalignment between
pretraining distribution of
XGLM and mT5

- Automated aligner is better than random prompting.
- It is competitive to semantic prompting in some languages.
- It fails to incorporate any task-specific signals, therefore failing to beat Task-based alignment.

Takeaways

- The efficiency of prompt design predominantly depends on pretraining objectives and downstream task
- Alignment via simple measures like semantic similarity can boost performance significantly
- Specially in cross-lingual setting, notion of task needs to be transferred from context to target

Our team working on LLMs



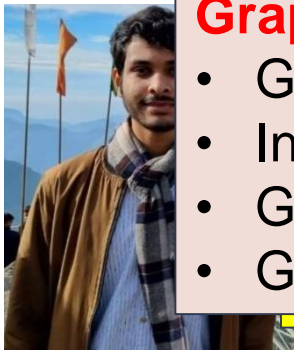
Subhabrata

Foundational models

- Physics-inspired LLMs
- Brain-inspired LLMs
- Increasing the reasoning ability of LLMs
- Layer-editing

Funding Agencies

Facebook,
DRDO

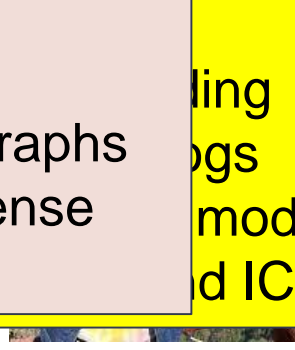


Manish

Graph Neural Networks

- Graph transformer
- Inducting learning on graphs
- Graph attacks and defense
- Graph applications

Karish



Eshaan

Applications

- Cyber safety (hate speech, misinformation)
- Mental health
- Text summarization
- Dialog systems

Gurusna

Joykirat

Laboratory for Computational Social Systems (LCS2)



Hiring MS, PhD, Postdoc, RA
(contact: tanchak@iitd.ac.in)

 <http://lcs2.in>

 @lcs2lab

Flipkart

Thanks to our sponsors



accenture



hike

