Zero-Shot Tokenizer Transfer

Benjamin Minixhofer, Edoardo M. Ponti, Ivan Vulić

Intro: an argument for modular tokenization

Image models are often pretrained on low resolution, then “uptrained” on higher resolution

e.g. PaliGemma (Beyer et al., 2024)

128x128

256x256

512x512

Intro: an argument for modular tokenization

Image models are often pretrained on low resolution, then “uptrained” on higher resolution

e.g. PaliGemma (Beyer et al., 2024)

128x128

256x256

512x512

Why not do the same for text?

```
and then-she found herself at
 last in the beautiful garden
```

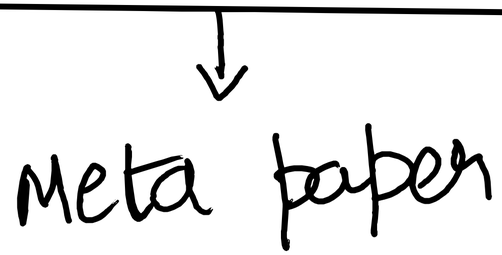
```
and then-she found herself at
 last in the beautiful garden
```

Weaker Tokenization → More granular tokenization.
i.e increasing vocab size

Intro: an argument for modular tokenization

Modular Tokenization methods may *accelerate* the switch to subword-free models

- My (highly speculative) best bet for getting rid of subword tokenization is
  - take a pretrained LLM, let's say Llama 3.1-8B
  - transfer to byte-level (e.g. via ZeTT, we will get to that)
    - or MYTE-level (Limisiewicz et al., 2024)
  - retrofit with Mixture-of-Depths and Multi-Token Prediction



Intro: an argument for modular tokenization

Different tokenizers represent different tradeoffs

Locking in on one tokenizer when pretraining is a big constraint

- Gemma Tokenizer: Every digit is a separate token
  - Pro: allocates higher compute budget to processing numbers
  - Con: Numbers quickly take up a lot of tokens

- XLM-V Tokenizer: Use a large (~900k) vocabulary to cover 100 languages
  - Pro: high compression in many languages
  - Con: Embedding parameters need ~3GB VRAM (93% of the model)

Intro: an argument for modular tokenization

Different tokenizers represent different tradeoffs

Locking in on one tokenizer when pretraining is a big constraint

	Size	NSL (↓) Avg.	Code	Eng.	Mult.
GPT-2 (Radford et al., 2019)	50k	1.13	1.19	0.86	1.33
DeepSeek Coder (DeepSeek AI, 2023)	32k	1.06	1.00	0.98	1.19
Llama (Touvron et al., 2023a)	32k	1.00	1.00	1.00	1.00
CodeGen (Nijkamp et al., 2023)	50k	1.05	0.95	0.86	1.33
CodeT5 (Wang et al., 2021)	32k	1.29	0.94	1.11	1.83
SantaCoder (Allal et al., 2023)	49k	1.04	0.88	1.07	1.17
StarCoder (Li et al., 2023)	49k	0.99	0.87	1.04	1.07
Replit Code (Replit, 2023)	32k	1.00	0.85	1.06	1.10
GPT-4 (OpenAI, 2023)	100k	0.85	0.75	0.84	0.95
InCoder (Fried et al., 2023)	50k	1.03	0.74	1.02	1.31

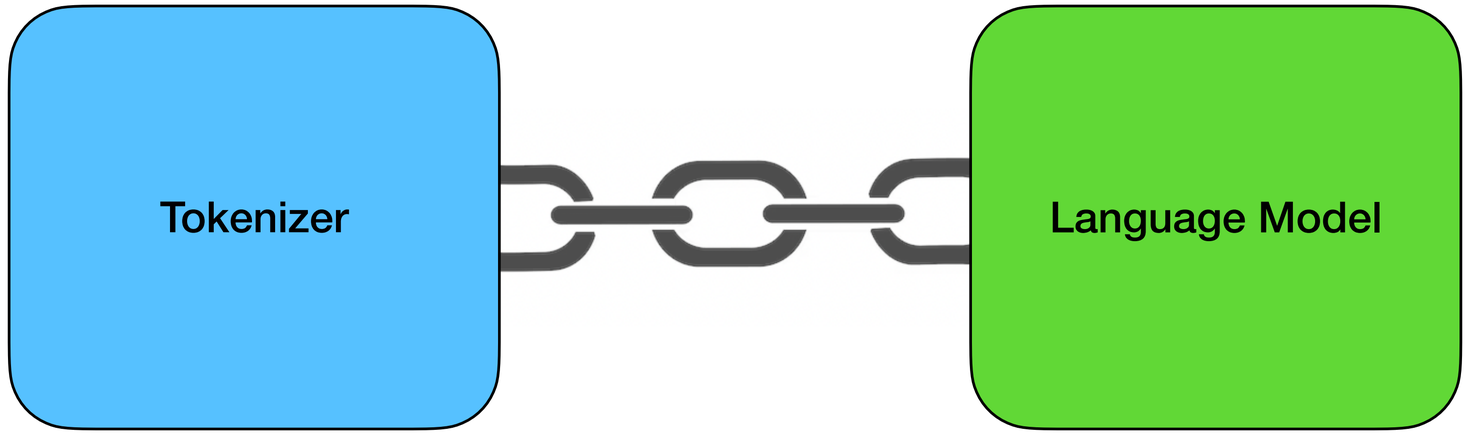
GPT 4-0 200K

from Dagan et al., 2024

Llama - 3.1B 128K

Now, let’s talk about Zero-Shot Tokenizer Transfer

Language Models are bound to their Tokenizer

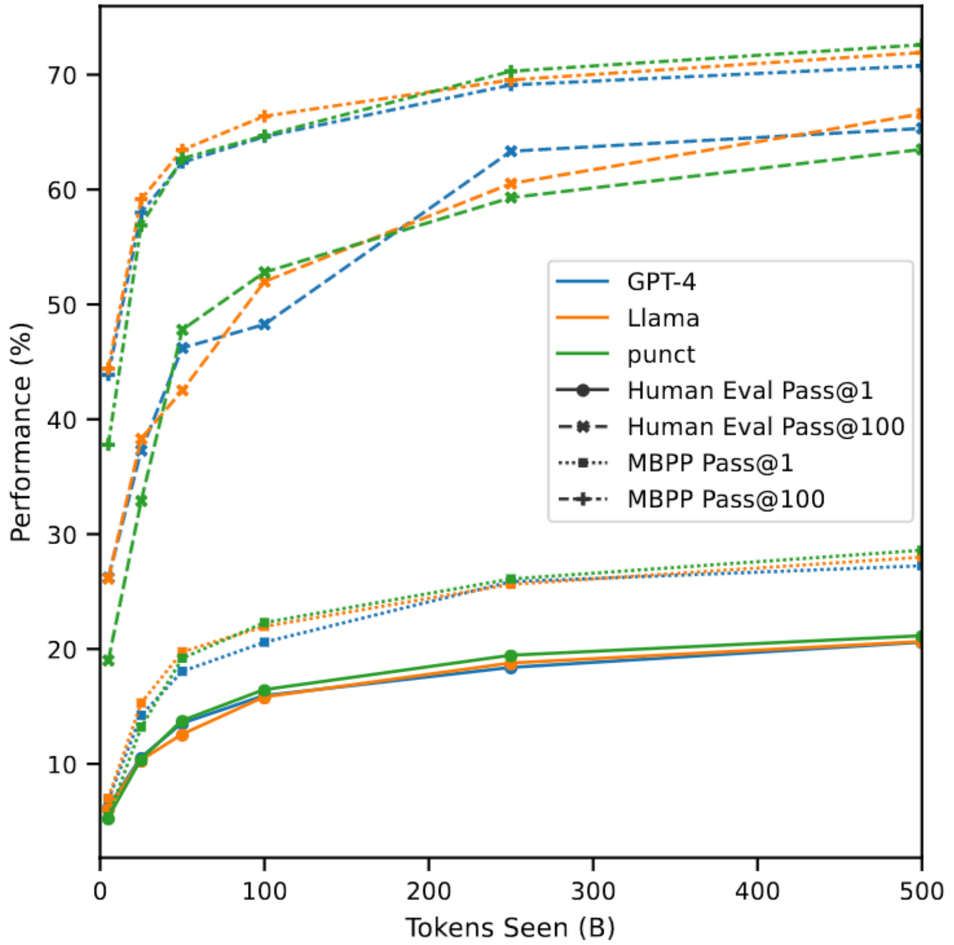


Language Models are bound to their Tokenizer

“Code Llama re-uses the tokenizer that was used by its base model, Llama 2, which uses the tokenizer from the original Llama model. This means that Code Llama, while being a popular model fine-tuned on a domain specific task, is still limited by the decisions taken during the pretraining of the original base model.”
(Dagan et al., 2024)

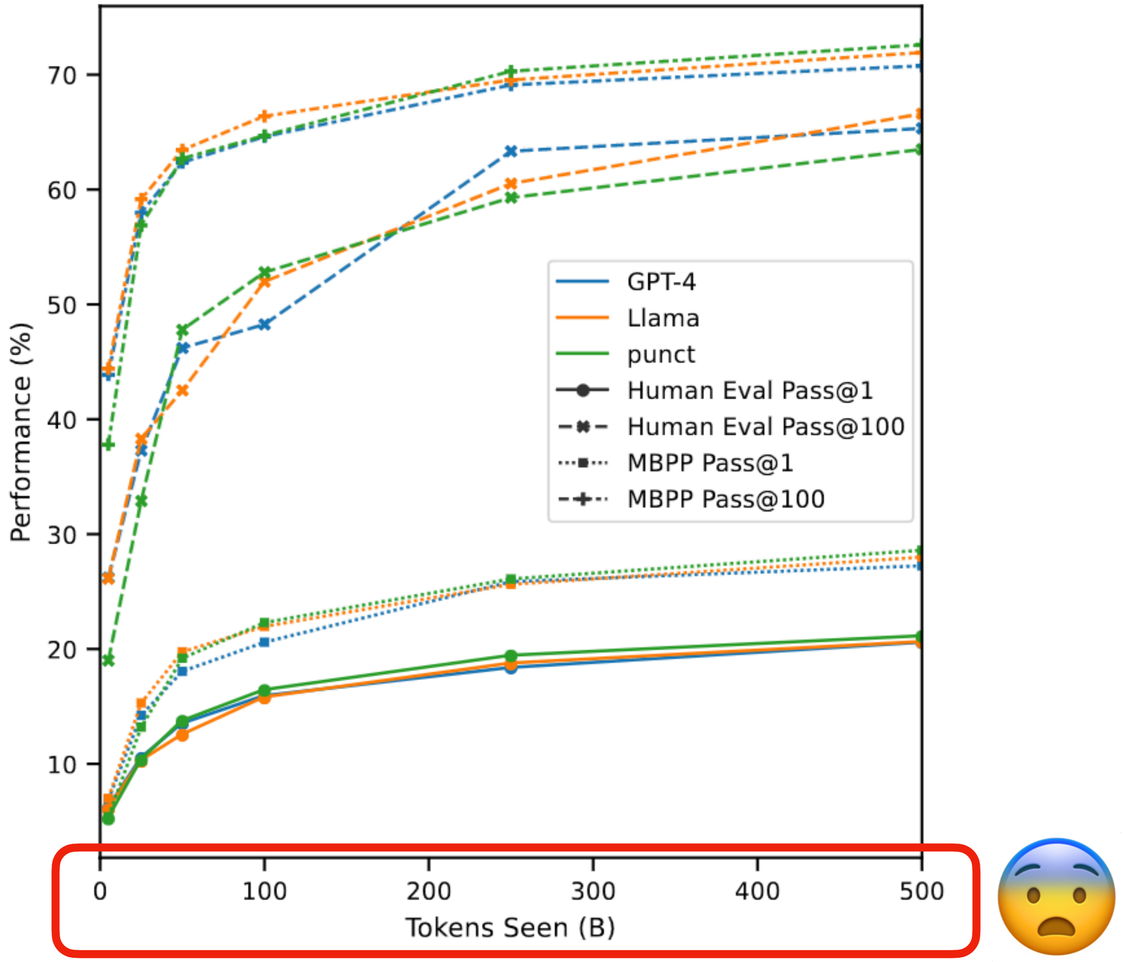
“In multilingual settings, subword tokenizers lead to disproportionate fragmentation rates for different languages and writing scripts. […] This directly increases cost of API usage for certain language speakers, even if they convey the same information as the others.”
(Ahia et al., 2023)

Language Models are bound to their Tokenizer



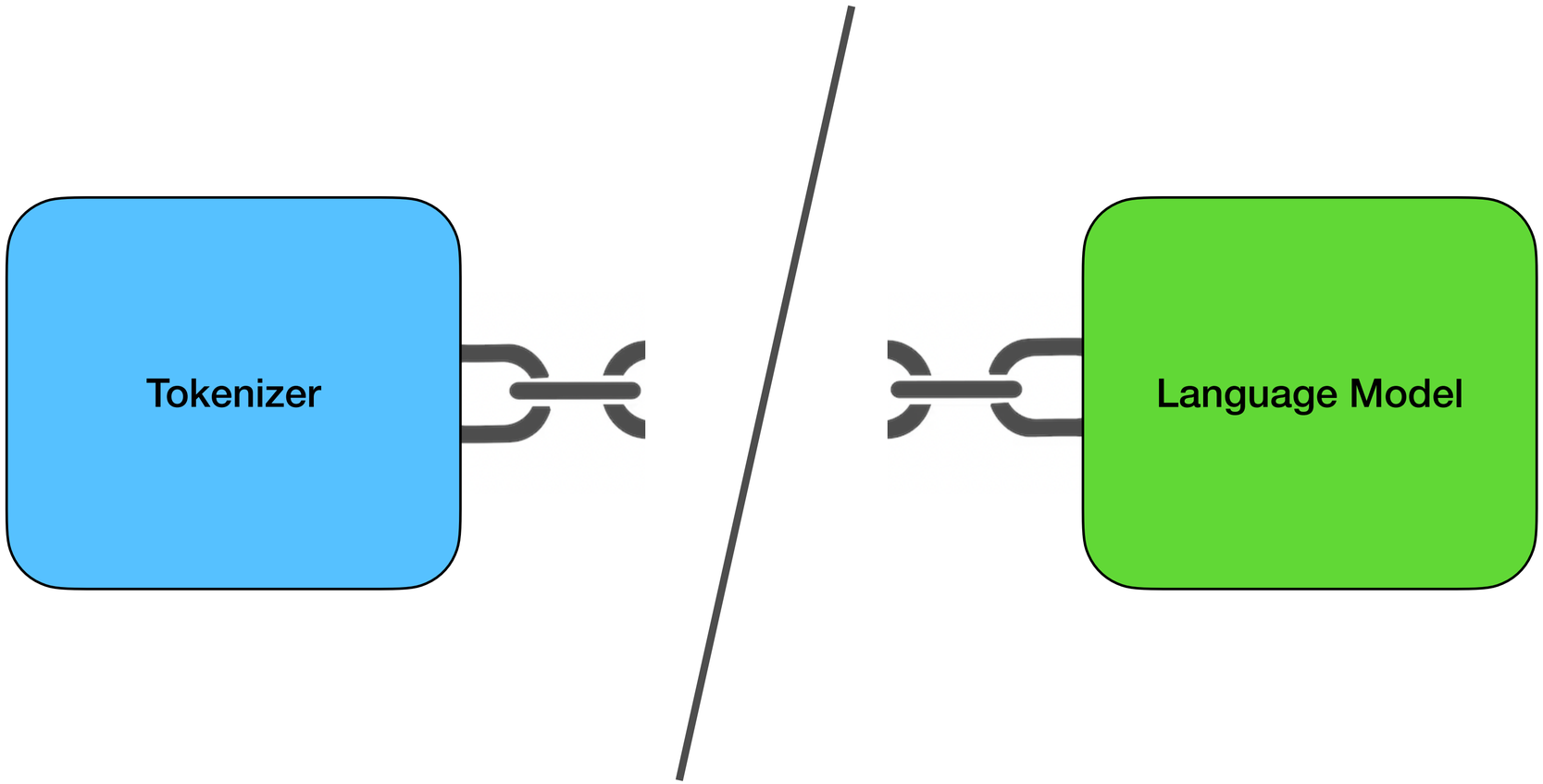
(Dagan et al., 2024)

Language Models are bound to their Tokenizer



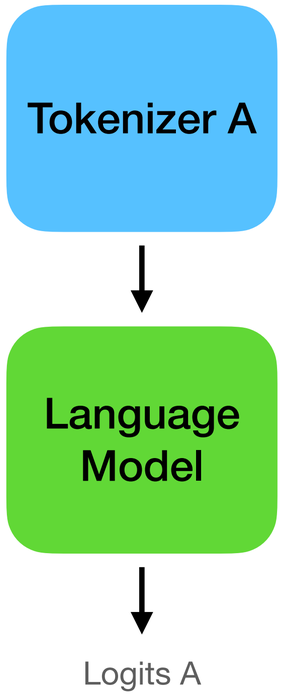
(Dagan et al., 2024)

We want to detach Language Models from their Tokenizers

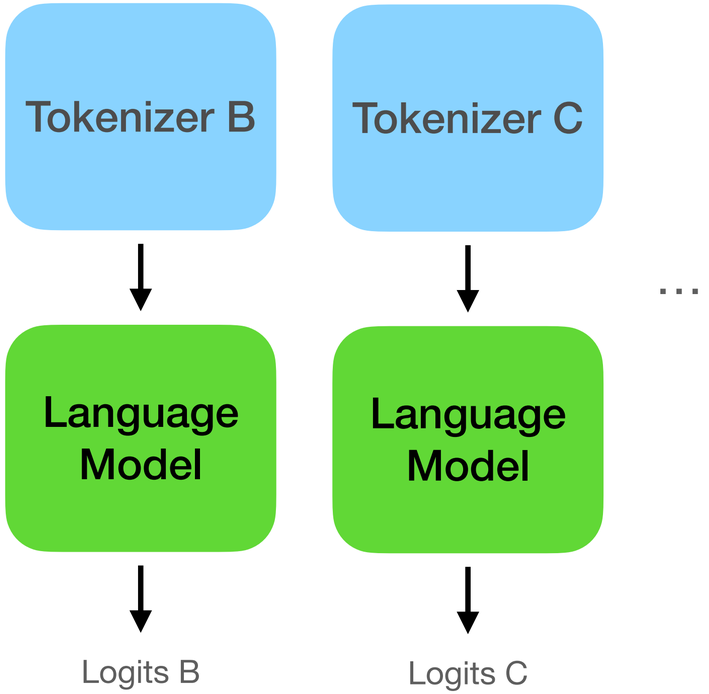


Another View

We have this…



We want this…

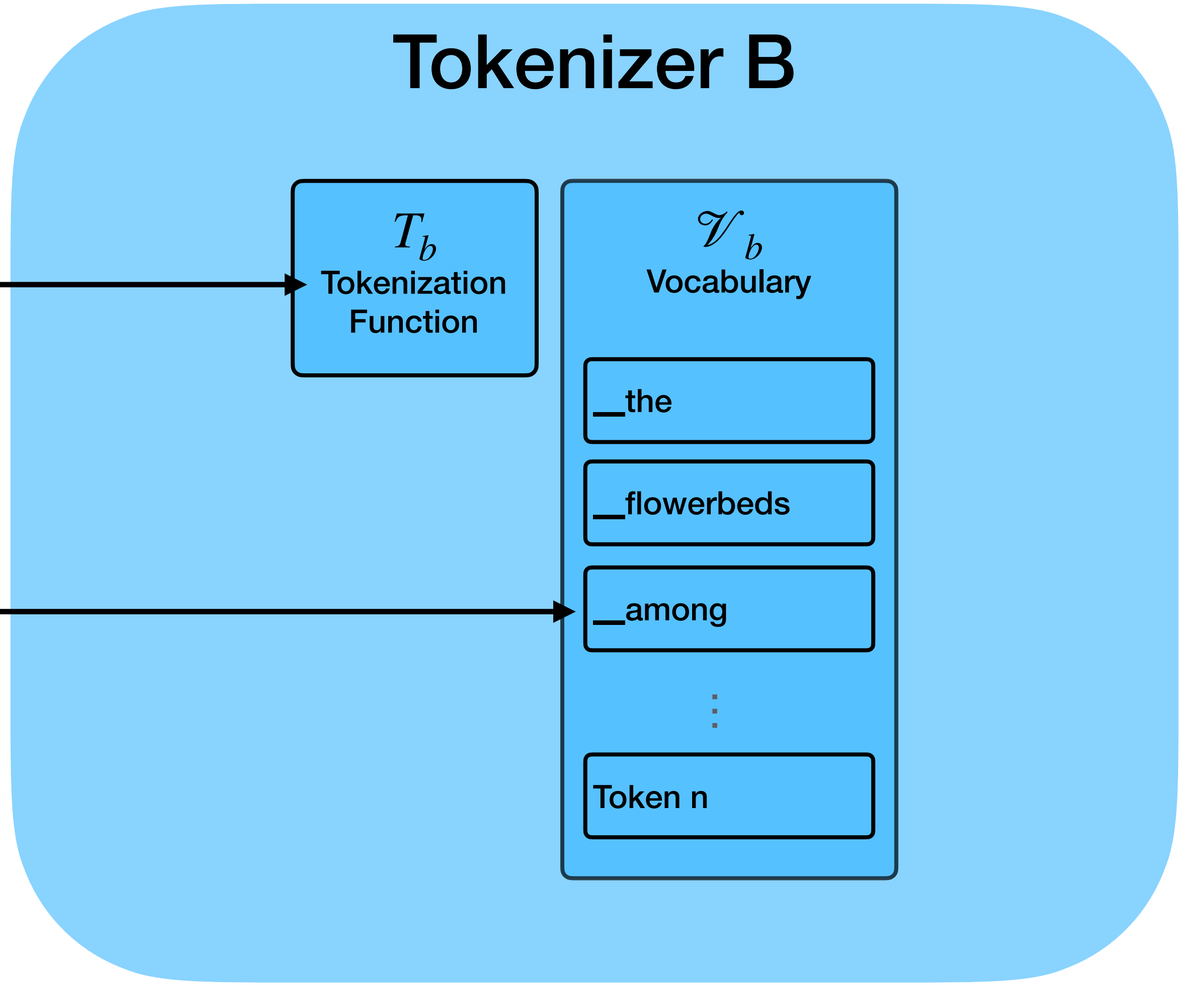


…

A look inside a tokenizer

Maps text to a sequence of elements of $\mathcal{V}_b$

The “universe” of possible tokens



The tokenization function determines how the vocabulary is used

Example: Llama2 vocabulary, BPE vs Longest Prefix

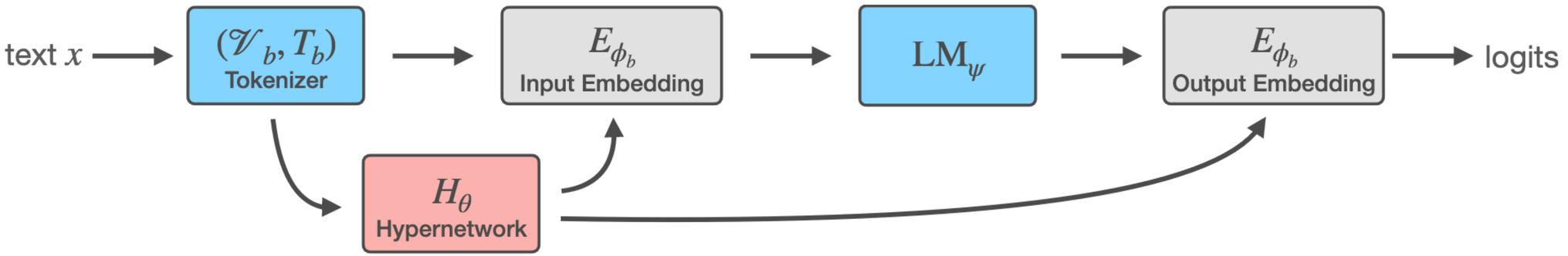
− 20 removals 1 line Copy

```
1 _Did _you _know _you _have _two _little _yellow , _nine - vol t - b atter y - s ized _ad ren al _g lands _in _your _body , _just _ch illing _out , _max in âĢĻ , _relax in âĢĻ _all _cool _on _top _of _your _kid ne ys ? _Some one _told _me _this _and _I _checked _it _out . _Turn s _out _it âĢĻ s _true . Ċ It _seems _as _though _your _ad ren al _g lands _are _kind _of _like _tho se _British _Royal _Gu ards _with _the _big , _black _f uz zy _h ats _who _ stand _like _stat ues _in _front _of _Buck ingham _Palace . _They _just _st and _there _quietly , _not _doing _much _really , _just _enjo ying _the _br own , _sli pp ery _beach _that _is _your _kid ne ys . Ċ However , _if _anyt hing _start ling _should _happen _that _requires _your _attention _âĢĶ _lik e _say _you âĢĻ re _about _to _give _a _speech _at _a _wed ding _or _your _ hear _a _tw ig _crack _outside _your _tent _or _your _door bell _rings _in _the _middle _of _the _night _âĢĶ _then _they _le ap _into _action , _jump ing _out _of _their _peace ful _sl umber _to _s que e ze _out _a _big _do s e _of _ad ren al ine _right _into _your _body , _p ump ing _you _up , _and _turning _you _into _a _prim al , _war rior - like _version _of _yourself . Ċ When _t ension _runs _high _and _ad ren al ine _is _secret ed _into _your _body _some _cra zy _things _can _happen _âĢĶ _sometimes _called _the _figh t - or - fl ight _response : Ċ - _Your _heart _rate _increases . _And _spec ifically , _your _body _starts _sending _blood _to _all _your _big _mus cle s _and _diver ts _it _away _from _âĢĽ non - crit ical âĢĿ _parts _of _your _body , _like _your _brain , _imm une _system , _and _dig est ive _system . _I _guess _someone _figured _you _could _dig est _the _sand wich _after _yo u _killed _the _bear .
```

+ 20 additions 1 line Copy

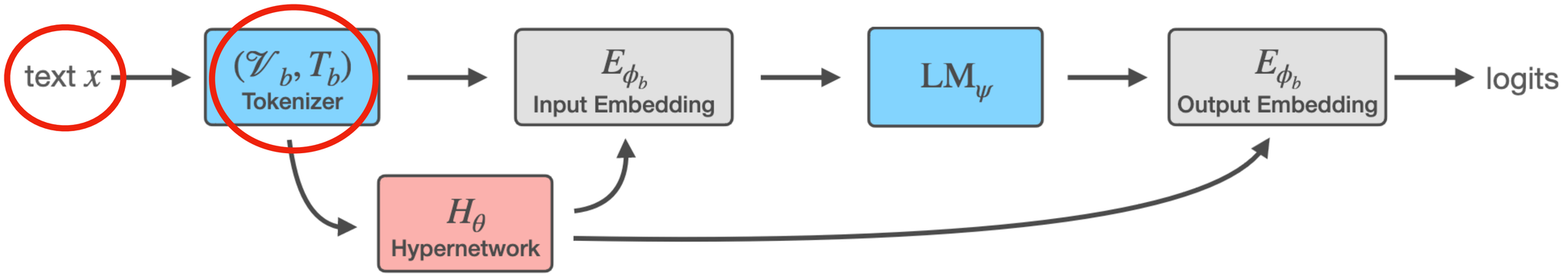
```
1 _Did _you _know _you _have _two _little _yellow , _nine - vol t - bat tery - size d _ad ren al _gla nd s _in _your _body , _just _chi ll ing _out , _m ax in âĢĻ , _relax in âĢĻ _all _cool _on _top _of _your _kid ney s ? _Some one _told _me _this _and _I _checked _it _out . _Turn s _out _it âĢĻ s _tru e . Ċ It _seems _as _though _your _ad ren al _gla nd s _are _kind _of _like _those _British _Royal _Guard s _with _the _big , _black _fu zz y _hat s _w ho _stand _like _statue s _in _front _of _Buck ingham _Palace . _They _just _stand _there _quietly , _not _doing _much _really , _just _enjoy ing _the _brown , _sli pper y _beach _that _is _your _kid ney s . Ċ However , _if _a nything _start ling _should _happen _that _requires _your _attention _âĢĶ _ like _say _you âĢĻ re _about _to _give _a _speech _at _a _wed ding _or _you r _hear _a _tw ig _crack _outside _your _tent _or _your _door bell _rings _ in _the _middle _of _the _night _âĢĶ _then _they _le ap _into _action , _ju mp ing _out _of _their _peace ful _sl umber _to _squ ee ze _out _a _big _do s e _of _ad ren alin e _right _into _your _body , _pu mp ing _you _up , _an d _turning _you _into _a _prima l , _war rior - like _version _of _yourself . Ċ When _tens ion _runs _high _and _ad ren alin e _is _secret ed _into _yo ur _body _some _cra zy _things _can _happen _âĢĶ _sometimes _called _the _f ight - or - fli gh t _response : Ċ - _Your _heart _rate _increases . _And _ specifically , _your _body _starts _sending _blood _to _all _your _big _mus cles _and _diver ts _it _away _from _âĢĽ non - crit ical âĢĿ _parts _of _yo ur _body , _like _your _brain , _imm une _system , _and _dig esti ve _syste m . _I _guess _someone _figured _you _could _dig est _the _sand wich _after _you _killed _the _bear .
```

A Hypernetwork for ZeTT



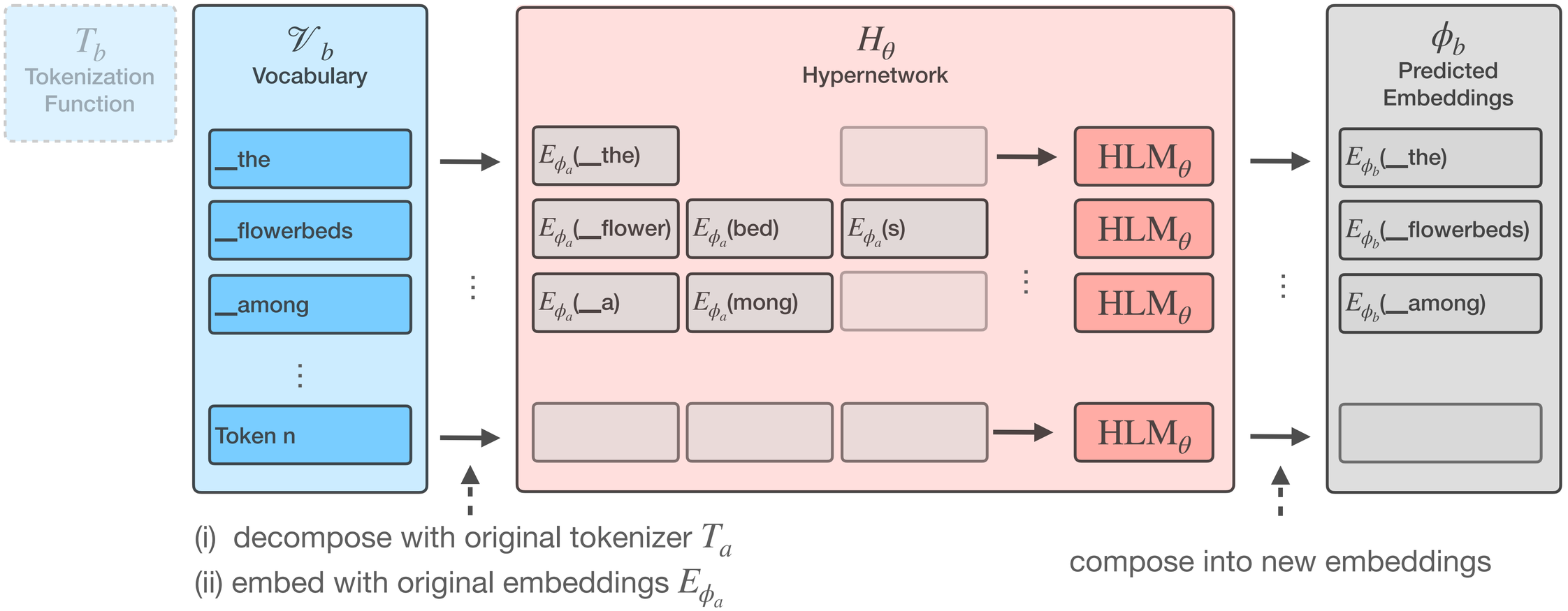
HLM

A Hypernetwork for ZeTT



To train this, we need to sample texts ***and tokenizers.***

Architecture

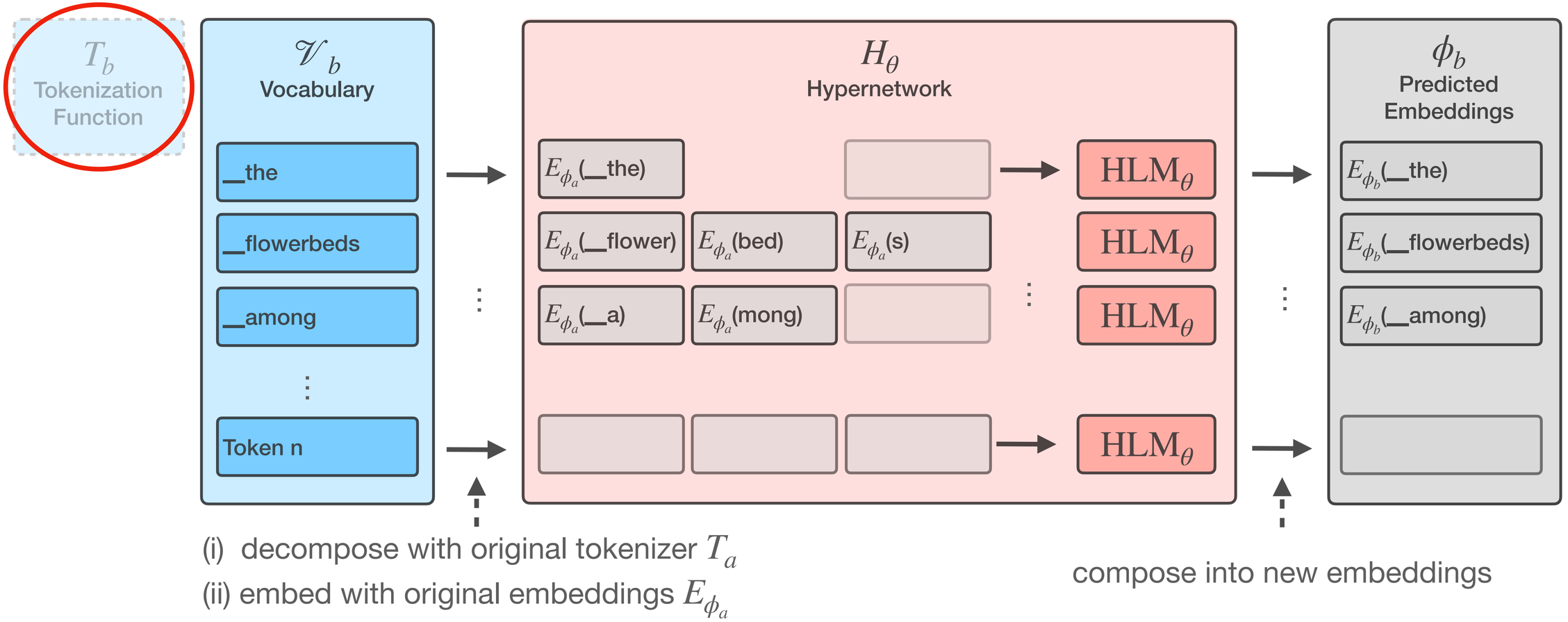


(i) decompose with original tokenizer T_a

(ii) embed with original embeddings E_{ϕ_a}

compose into new embeddings

Architecture



(i) decompose with original tokenizer T_a

(ii) embed with original embeddings E_{ϕ_a}

compose into new embeddings

We ***"amortise"*** over the tokenization function.

Zero-Shot Tokenizer Transfer

- Once we have a trained hypernetwork, we can use it to predict the embeddings for any arbitrary tokenizer, enabling zero-shot tokenizer transfer (ZeTT)
- ZeTT ‘detaches’ language models from the tokenizer they were pretrained with
- We can also keep training the hypernetwork (or the entire model) for some steps with the target tokenizer (n-shot tokenizer transfer).
- The baselines are heuristics with which we can initialise embeddings:
  - OFA (Liu et al, 2023)
  - FOCUS (Dobler & de Melo, 2023)
  - FVT (Gee et al, 2021)

Zero-Shot Tokenizer Transfer

Table 2: Performance of Mistral-7B-v0.1 after zero-shot and n-shot tokenizer transfer (training on 800M tokens). We evaluate transfer to the GPT2 tokenizer on natural language benchmarks and transfer to the StarCoder tokenizer on HumanEvalPack. Note that continued training with the original tokenizer (*original@800M*) does not consistently improve performance.

#shots	Method	**Natural Language (→ GPT2 Tok.)**						**Code (pass@1) (→ StarCoder Tok.)**					
								HumanEvalPack					
		PiQA	HS	ARC	BoolQ	MMLU	Avg.	js	go	py	cpp	java	Avg.
original		80.7	81.0	79.5	83.6	59.6	76.9	28.7	20.1	29.3	29.9	32.3	28.1
original@800M		82.1	82.7	80.6	80.6	57.8	76.8	31.7	19.5	28.7	27.4	26.2	26.7
0-shot	FOCUS	69.2	63.8	45.7	60.4	38.8	55.6	21.9	1.8	0.0	20.1	22.6	13.3
	ours	**79.7**	**77.5**	**73.0**	**81.9**	**53.0**	**73.0**	**23.8**	**17.7**	**18.9**	**28.7**	**26.8**	**23.2**
n-shot	FOCUS@800M	74.8	74.3	72.4	73.3	48.9	68.7	24.4	17.1	22.6	22.6	26.2	22.6
	ours@800M	**80.9**	**80.7**	**77.8**	**80.7**	**54.4**	**74.9**	**28.0**	**25.0**	**26.2**	**29.9**	**28.7**	**27.6**

Zero-Shot Tokenizer Transfer

Table 3: Accuracy of Mistral-7B on XCOPA with language-specific tokenizers zero-shot transferred via FOCUS and our hypernetwork. The standard errors are between 2.1% and 2.3%.

	et	**ht**	**id**	**it**	**qu**	**sw**	**ta**	**tr**	**vi**	**Avg.**
original	46.6	51.6	58.0	65.8	48.4	51.4	54.4	56.4	59.0	54.6
FOCUS	52.0	53.0	51.2	49.2	**51.4**	54.6	54.0	55.2	49.8	52.3
ours	**53.4**	**57.2**	**60.0**	**65.6**	50.0	**57.2**	**55.8**	**57.4**	**57.2**	**57.1**
Δaccuracy	+7%	+6%	+2%	-0%	+1%	+6%	+1%	+1%	-2%	+3%
Δlength	-72%	-42%	-52%	-36%	-54%	-51%	-83%	-57%	-59%	-54%

Table 4: 5-shot accuracy of Mistral-7B on multilingual MMLU with the original tokenizer and language-specific tokenizers zero-shot transferred via FOCUS and our hypernetwork.

	original	**FOCUS**	**ours**	**Δaccuracy**	**Δlength**
German	51.6	26.2	**43.7**	-8%	-37%
Spanish	53.6	26.2	**45.9**	-8%	-32%
French	53.6	27.4	**44.8**	-9%	-30%
Italian	52.5	25.8	**42.7**	-10%	-36%
Russian	49.9	27.2	**35.1**	-15%	-47%

Huge decrease in num. of tokens

Cool, but do we have to train a hypernet for every model?

- No! Embeddings of fine-tunes are usually compatible with the base model.
- But yes, every base model needs a separate hypernetwork (for now)

Can't use HLM of Llama with Deepseek

Table 5: Single model rating results on MT-Bench of transferring Mistral-7B-Instruct-v0.1 to the GPT2 tokenizer using the hypernetwork trained for the base Mistral-7B model. We use `gpt-3.5-turbo-1106` as a judge. *orig.* is the original fine-tuned model, *base* the model with the same tokenizer but embeddings substituted for the base models' embeddings. λ is the scaling factor for the weight differences in Task Arithmetic (Ilharco et al., 2023).

	original		**0-shot**		**n-shot**			
Embeddings	orig.	base	FOCUS	ours	ours@800			
λ	-	-	-	-	0.0	0.3	0.5	0.7
Score (1 to 10)	7.33	7.48	5.03	**6.56**	6.59	6.75	**6.82**	6.77

Cool, but can we close the ZeTT gap?

tbd :) there are some promising directions.

- Incorporate distillation into the hypernetwork training
  - Currently simply uses causal language modelling
  - We could incorporate a lot more signal from the pretrained LM by doing distillation
  - Needs some way to convert probabilities from one tokenization to another (not trivial)

- Alleviate performance loss through amortisation
  - Some of the gap probably stems from having to amortise over the embeddings. Either:
  - Incorporate the tokenization function parameters into the hypernetwork input
  - Decide on a fixed tokenization function (e.g. longest prefix), and find a way to cheaply convert LMs to this tokenization function.

Thank you!

- Code: http://github.com/bminixhofer/zett
- Paper: https://arxiv.org/abs/2405.07883
- Email: bminixhofer@gmail.com