LangSAMP: Language-Script Aware Multilingual Pretraining

Yihong Liu[1,2,*], Haotian Ye[1,2,*],
Chunlan Ma[1,2], Mingyang Wang[1,2,3], Hinrich Schütze[1,2]

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CIS

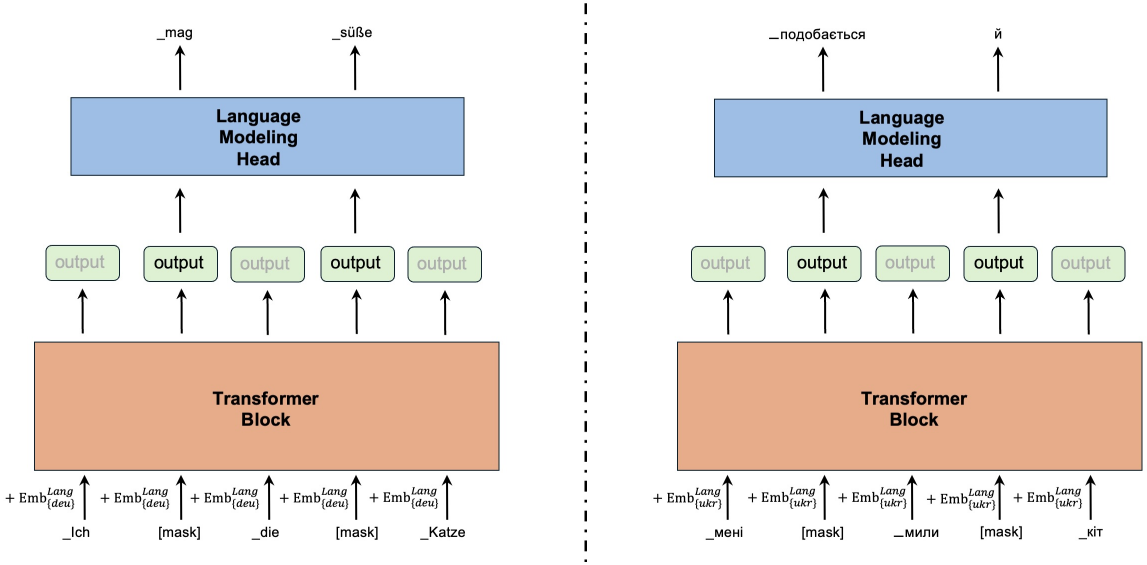
mcml Munich Center for Machine Learning

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]Bosch Center for Artificial Intelligence

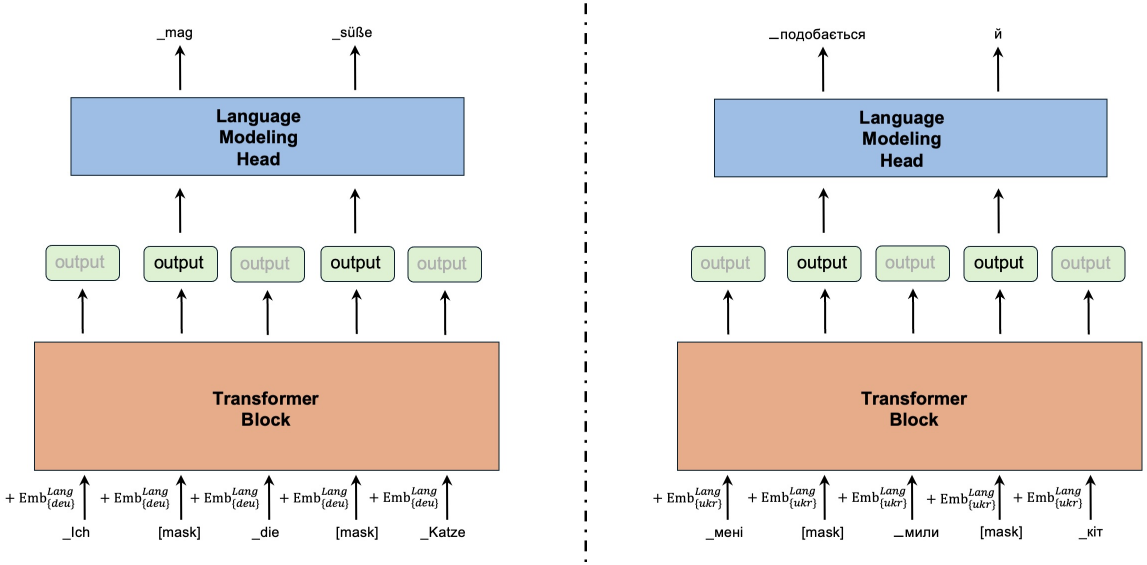
September 24, 2025

Outline

LangSAMP- Motivation & Background

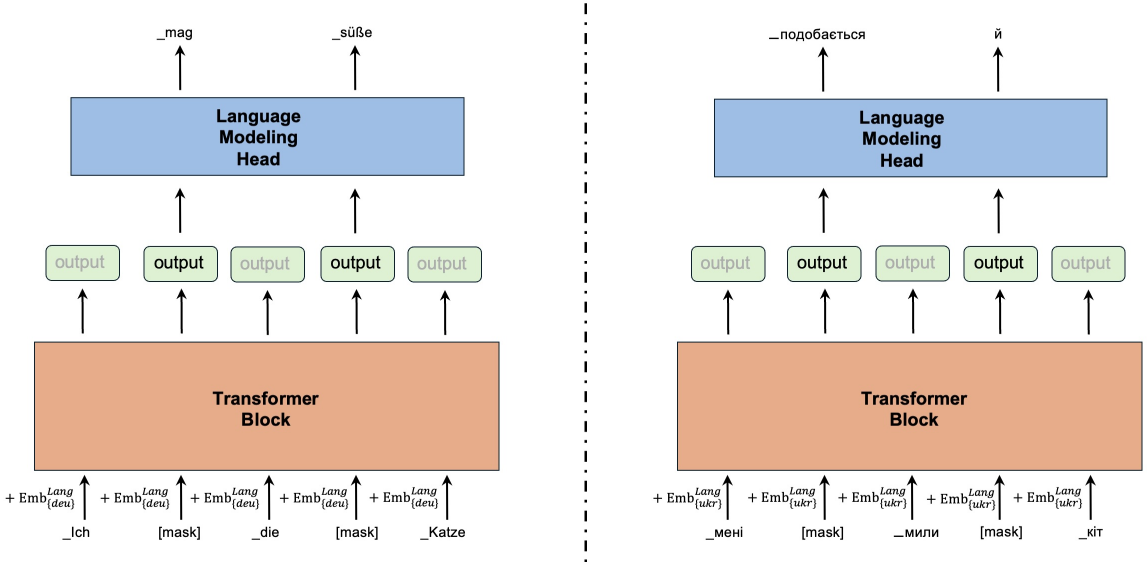


- Early crosslingual models like XLM leverage **language embeddings**.



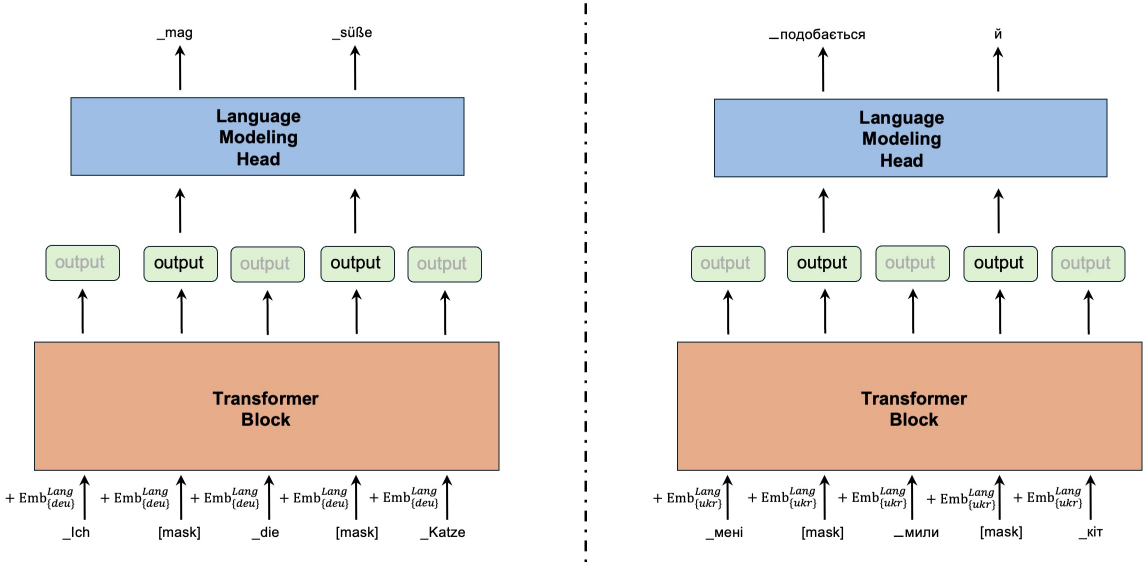
- Early crosslingual models like XLM leverage **language embeddings**.
  - learnable vectors

LangSAMP- Motivation & Background



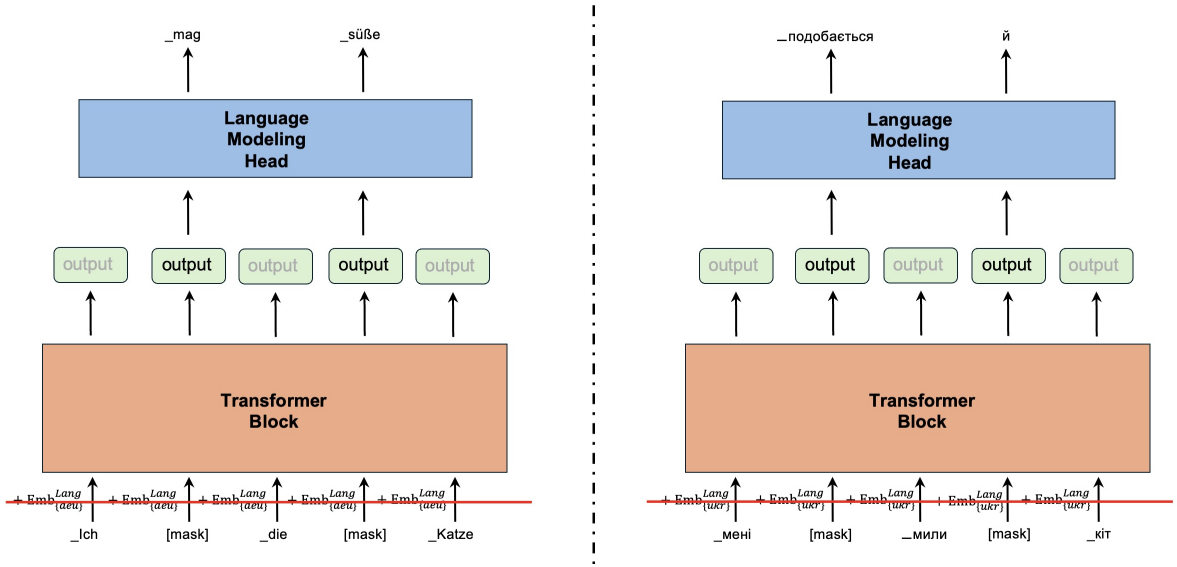
- Early crosslingual models like XLM leverage **language embeddings**.
  - learnable vectors
  - capture language-specific information

LangSAMP- Motivation & Background



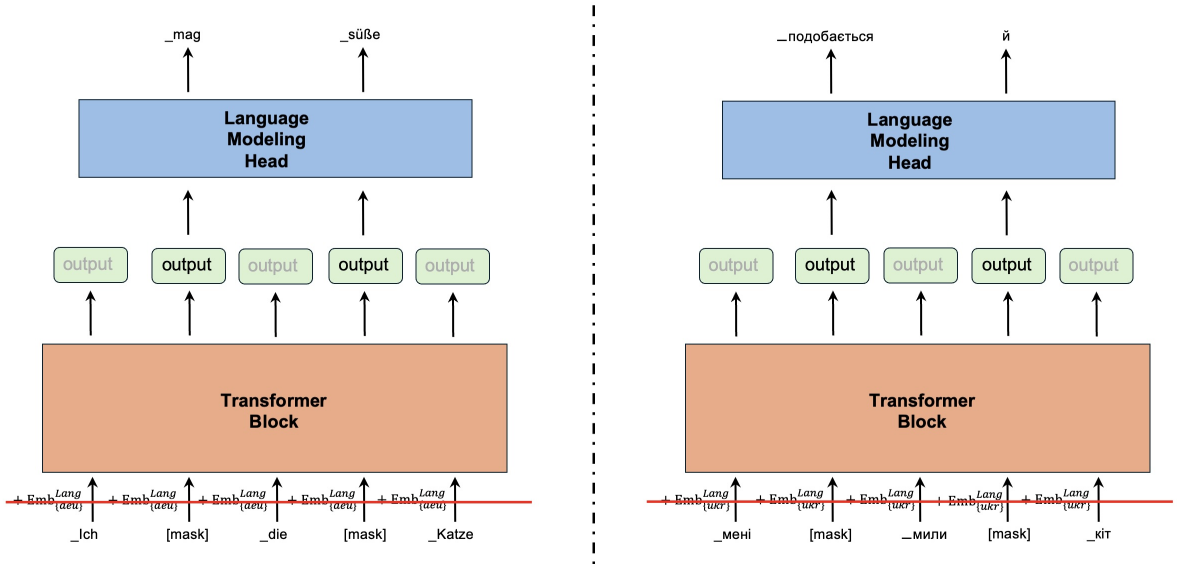
- Early crosslingual models like XLM leverage **language embeddings**.
  - learnable vectors
  - capture language-specific information
  - useful for guiding generation, e.g., MT

LangSAMP- Motivation & Background



Recent **mPLMs** like XLM-R **remove** such language embeddings.

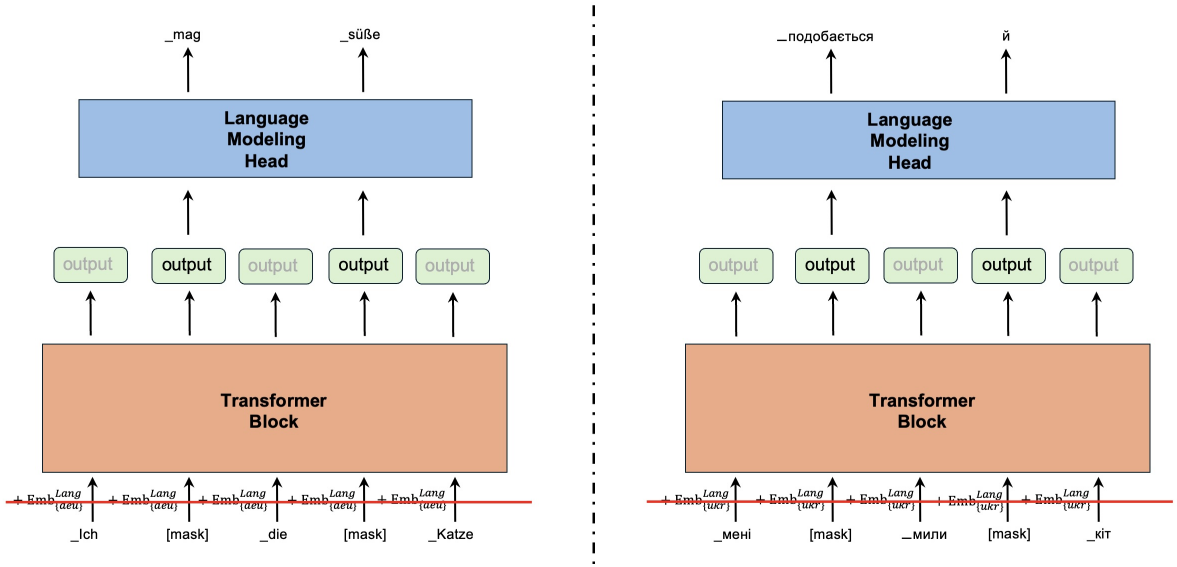
LangSAMP- Motivation & Background



Recent **mPLMs** like XLM-R **remove** such language embeddings.

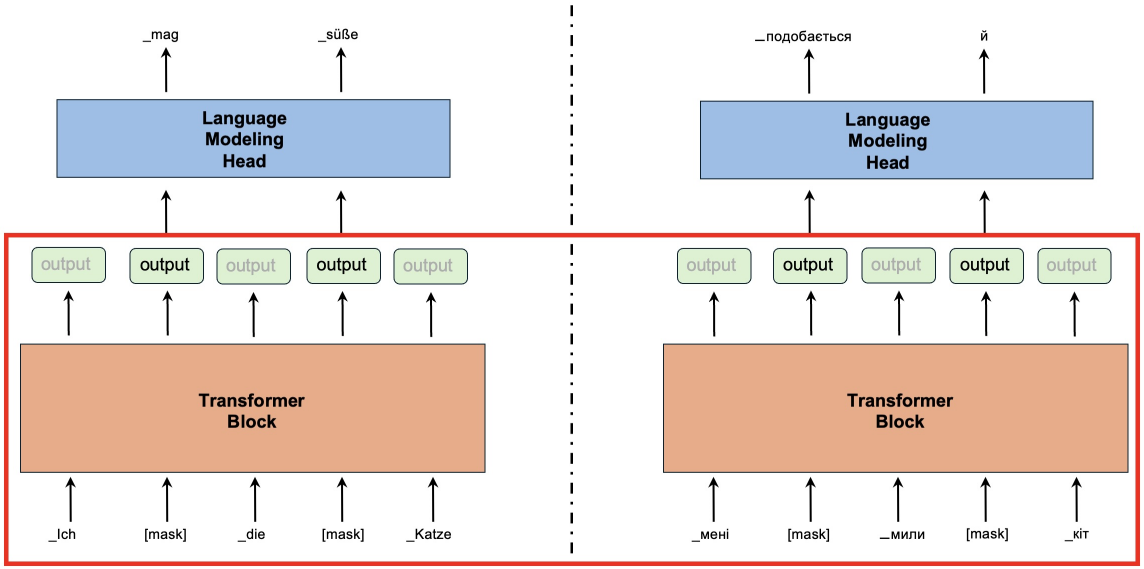
- What's **bad** about such removal?

LangSAMP- Motivation & Background

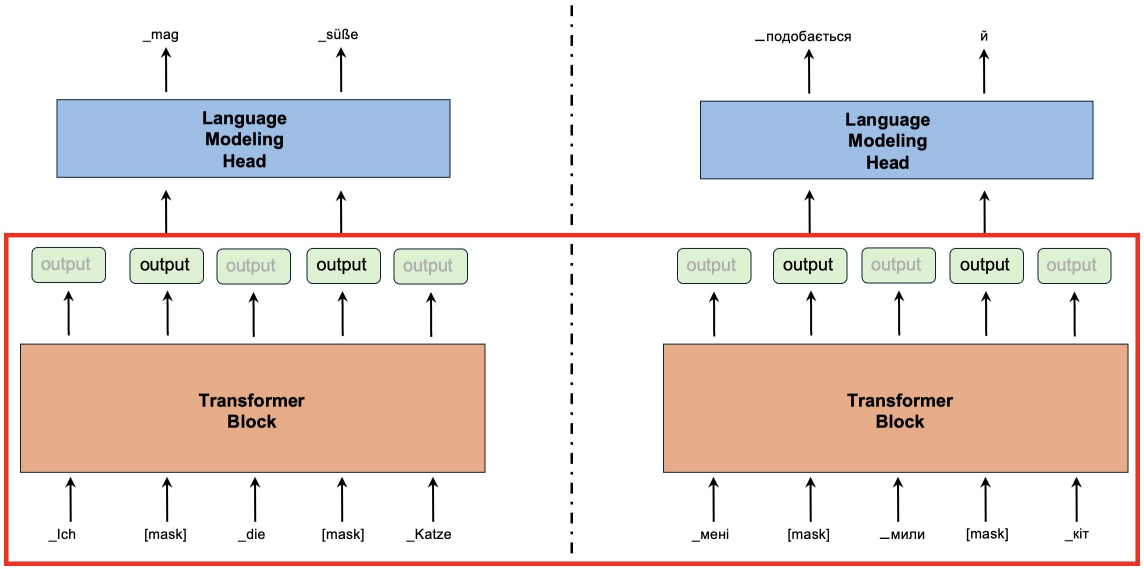


Recent **mPLMs** like XLM-R **remove** such language embeddings.

- What's **bad** about such removal?
  - The contextual token embeddings have to encode all **language-specific** information
  - This may hinder **language neutrality**
  - Language neutrality is usually important for multilinguality, e.g., (zero-shot) **crosslingual transfer**.



- What's **good** about such removal?

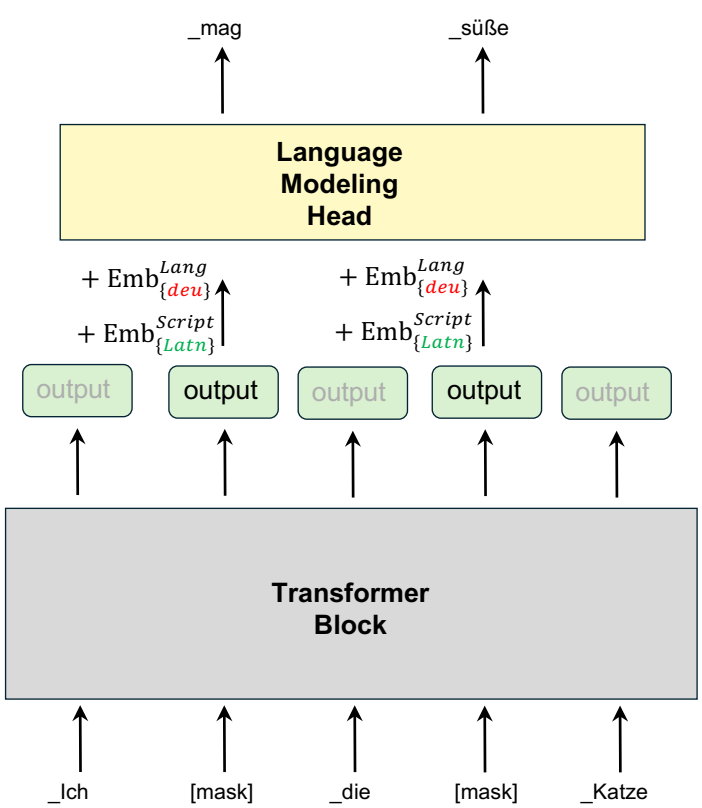


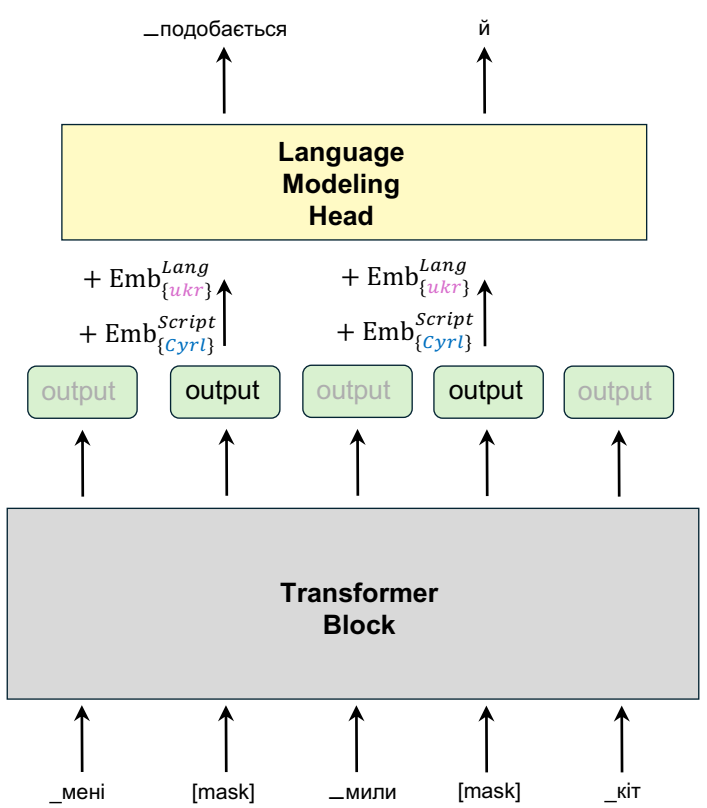
- What's **good** about such removal?
  - Universal text encoder without requiring language IDs as input
  - The backbone model can be used effectively for downstream tasks

- Why not integrate language/script embeddings **wisely** so that
  - **representations' language-neutrality is improved**
  - **the backbone remains the same as common mPLMs**

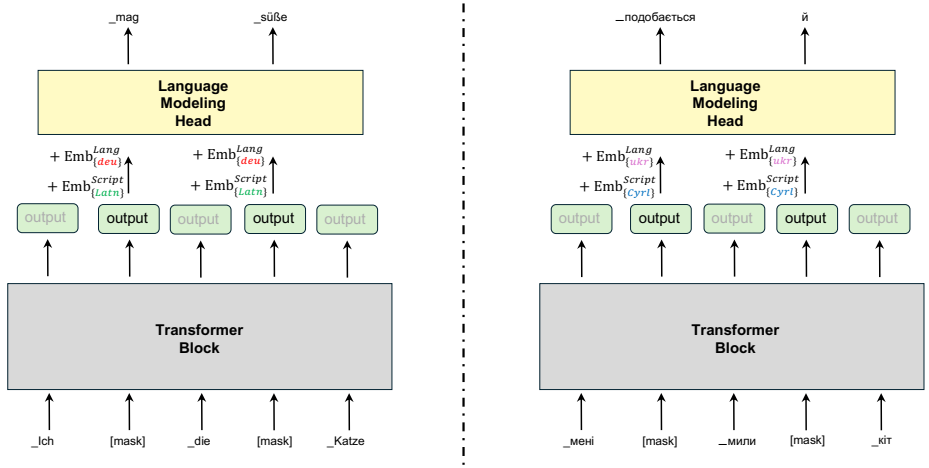
?

We add the language and script embedding to **the outputs of the Transformer blocks** at token position i: $\boldsymbol{o}_i = \boldsymbol{h}_i + \boldsymbol{E}_l^{Lang} + \boldsymbol{E}_s^{Script}$.

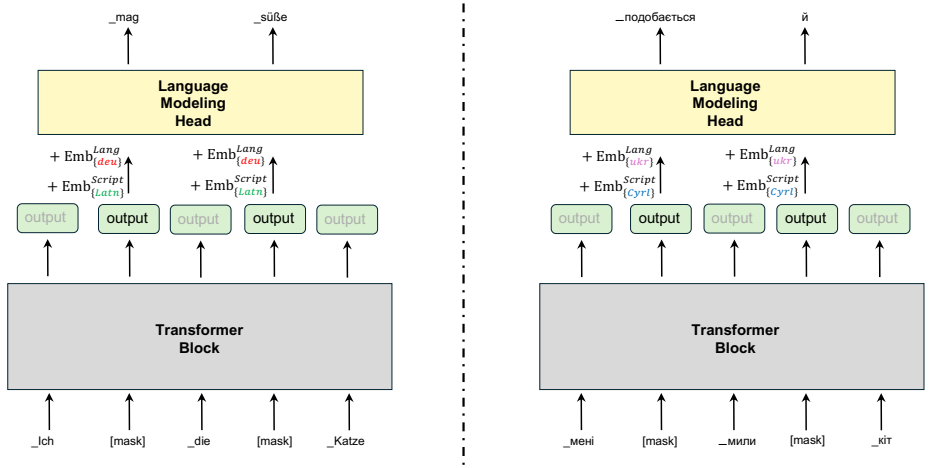




Note: language/script embeddings are **not required** to obtain the final Transformer output, i.e., the final contextual token embeddings.

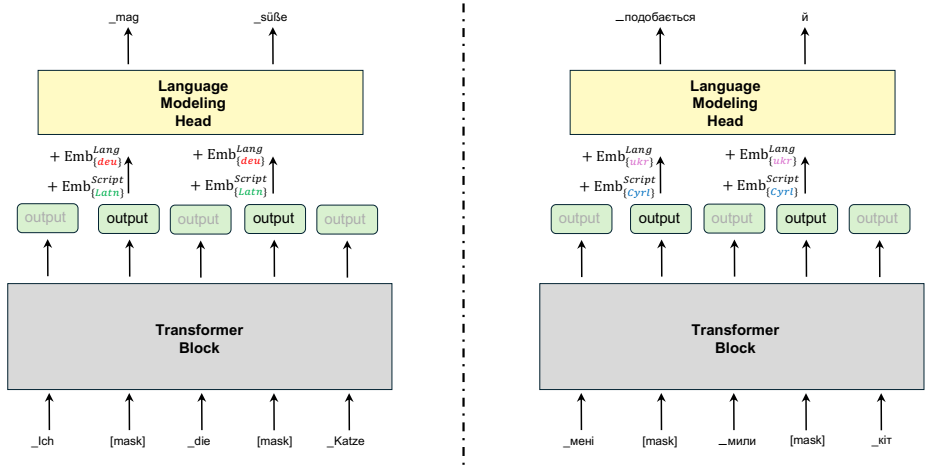


Note: language/script embeddings are **not required** to obtain the final Transformer output, i.e., the final contextual token embeddings.



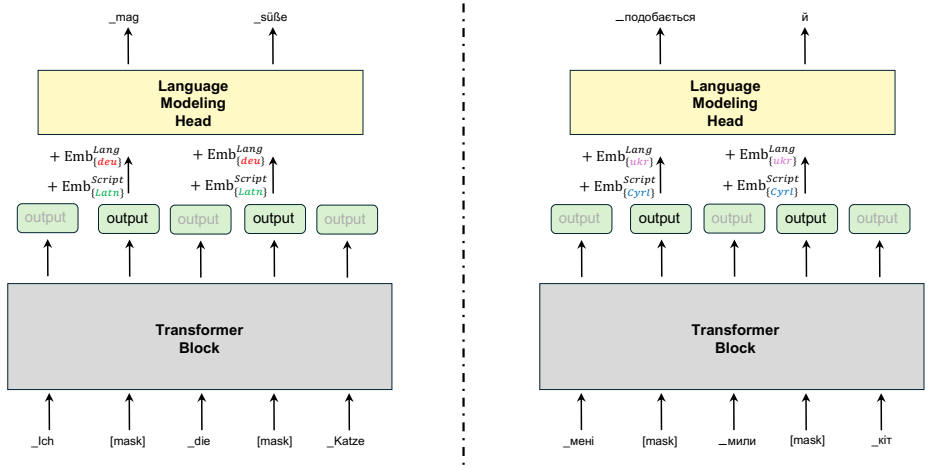
- **No language or script IDs** as input needed in fine-tuning.

Note: language/script embeddings are **not required** to obtain the final Transformer output, i.e., the final contextual token embeddings.



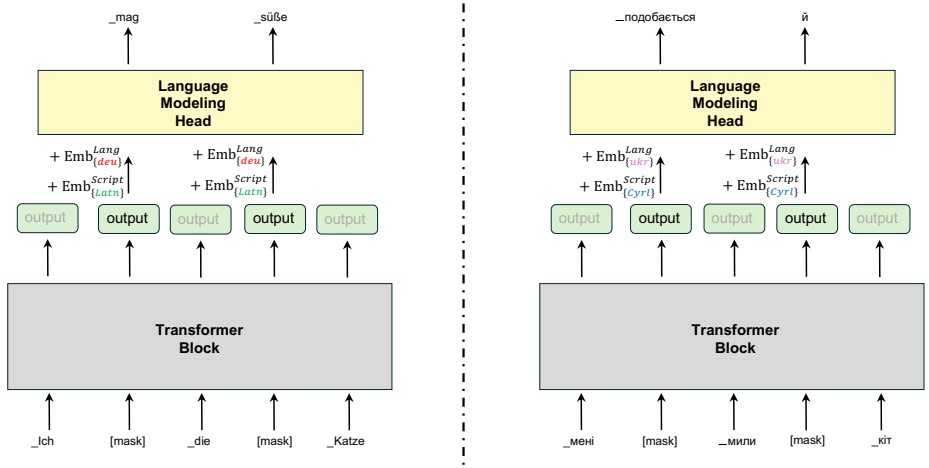
- **No language or script IDs** as input needed in fine-tuning.
- The backbone **remains the same as most mPLMs**.

Note: language/script embeddings are **not required** to obtain the final Transformer output, i.e., the final contextual token embeddings.



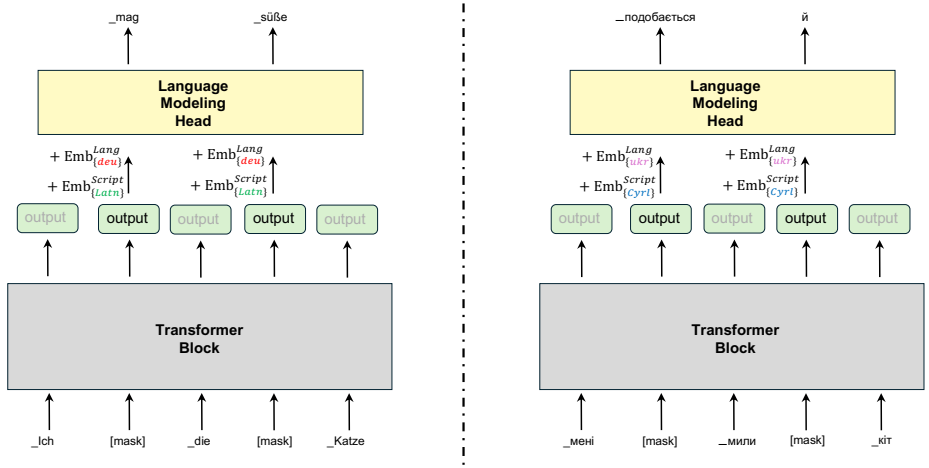
- **No language or script IDs** as input needed in fine-tuning.
- The backbone **remains the same as most mPLMs**.
- It can be fine-tuned in the **standard way in NLP pipelines**.

Note: language/script embeddings are **not required** to obtain the final Transformer output, i.e., the final contextual token embeddings.



- **No language or script IDs** as input needed in fine-tuning.
- The backbone **remains the same as most mPLMs**.
- It can be fine-tuned in the **standard way in NLP pipelines**.
  - Token (and position) embeddings + Transformer blocks as **backbone**.

Note: language/script embeddings are **not required** to obtain the final Transformer output, i.e., the final contextual token embeddings.



- **No language or script IDs** as input needed in fine-tuning.
- The backbone **remains the same as most mPLMs**.
- It can be fine-tuned in the **standard way in NLP pipelines**.
  - Token (and position) embeddings + Transformer blocks as **backbone**.
  - A **task-specific classifier** is added on top of it.

Outline

Main Results

- Continued pretraining XLM-R on Glot500-c
- **head**: the language covered by XLM-R.
- **tail**: the language not covered by XLM-R

	tail		head		Latn		non-Latn		all	
	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP
SR-B	36.9 (0.0)	**39.5** (0.0)	60.6 (0.0)	**61.3** (0.0)	40.7 (0.0)	**42.8** (0.0)	51.2 (0.0)	**53.5** (0.0)	42.9 (0.0)	**45.1** (0.0)
SR-T	56.9 (0.0)	**58.6** (0.0)	74.8 (0.0)	**76.1** (0.0)	67.5 (0.0)	**68.7** (0.0)	73.7 (0.0)	**75.6** (0.0)	69.7 (0.0)	**71.1** (0.0)
Taxi1500	47.1 (4.8)	**50.8** (2.4)	59.9 (2.9)	**61.2** (1.2)	48.2 (4.6)	**51.7** (2.1)	58.8 (3.1)	**60.1** (1.7)	50.3 (4.2)	**53.4** (2.0)
SIB200	69.0 (1.4)	**70.2** (1.9)	82.2 (1.4)	**82.6** (1.2)	72.1 (1.3)	**73.1** (1.8)	81.1 (1.5)	**81.7** (1.2)	75.0 (1.3)	**75.9** (1.6)
NER	60.1 (0.6)	**60.8** (0.8)	64.0 (0.6)	**64.1** (0.6)	67.0 (0.5)	**67.6** (0.6)	**53.9** (0.7)	**53.9** (0.5)	62.2 (0.5)	**62.6** (0.6)
POS	61.3 (1.0)	**61.4** (0.9)	76.0 (0.4)	**76.2** (0.4)	**74.6** (0.5)	74.5 (0.4)	66.2 (1.0)	**66.8** (0.8)	71.5 (0.6)	**71.6** (0.5)

Main Results

- Continued pretraining XLM-R on Glot500-c
- **head**: the language covered by XLM-R.
- **tail**: the language not covered by XLM-R

	tail		head		Latn		non-Latn		all	
	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP
SR-B	36.9 (0.0)	**39.5** (0.0)	60.6 (0.0)	**61.3** (0.0)	40.7 (0.0)	**42.8** (0.0)	51.2 (0.0)	**53.5** (0.0)	42.9 (0.0)	**45.1** (0.0)
SR-T	56.9 (0.0)	**58.6** (0.0)	74.8 (0.0)	**76.1** (0.0)	67.5 (0.0)	**68.7** (0.0)	73.7 (0.0)	**75.6** (0.0)	69.7 (0.0)	**71.1** (0.0)
Taxi1500	47.1 (4.8)	**50.8** (2.4)	59.9 (2.9)	**61.2** (1.2)	48.2 (4.6)	**51.7** (2.1)	58.8 (3.1)	**60.1** (1.7)	50.3 (4.2)	**53.4** (2.0)
SIB200	69.0 (1.4)	**70.2** (1.9)	82.2 (1.4)	**82.6** (1.2)	72.1 (1.3)	**73.1** (1.8)	81.1 (1.5)	**81.7** (1.2)	75.0 (1.3)	**75.9** (1.6)
NER	60.1 (0.6)	**60.8** (0.8)	64.0 (0.6)	**64.1** (0.6)	67.0 (0.5)	**67.6** (0.6)	**53.9** (0.7)	**53.9** (0.5)	62.2 (0.5)	**62.6** (0.6)
POS	61.3 (1.0)	**61.4** (0.9)	76.0 (0.4)	**76.2** (0.4)	**74.6** (0.5)	74.5 (0.4)	66.2 (1.0)	**66.8** (0.8)	71.5 (0.6)	**71.6** (0.5)

- Consistently better performance on both head and tail languages.

Main Results

- Continued pretraining XLM-R on Glot500-c
- **head**: the language covered by XLM-R.
- **tail**: the language not covered by XLM-R

	tail		head		Latn		non-Latn		all	
	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP
SR-B	36.9 (0.0)	**39.5** (0.0)	60.6 (0.0)	**61.3** (0.0)	40.7 (0.0)	**42.8** (0.0)	51.2 (0.0)	**53.5** (0.0)	42.9 (0.0)	**45.1** (0.0)
SR-T	56.9 (0.0)	**58.6** (0.0)	74.8 (0.0)	**76.1** (0.0)	67.5 (0.0)	**68.7** (0.0)	73.7 (0.0)	**75.6** (0.0)	69.7 (0.0)	**71.1** (0.0)
Taxi1500	47.1 (4.8)	**50.8** (2.4)	59.9 (2.9)	**61.2** (1.2)	48.2 (4.6)	**51.7** (2.1)	58.8 (3.1)	**60.1** (1.7)	50.3 (4.2)	**53.4** (2.0)
SIB200	69.0 (1.4)	**70.2** (1.9)	82.2 (1.4)	**82.6** (1.2)	72.1 (1.3)	**73.1** (1.8)	81.1 (1.5)	**81.7** (1.2)	75.0 (1.3)	**75.9** (1.6)
NER	60.1 (0.6)	**60.8** (0.8)	64.0 (0.6)	**64.1** (0.6)	67.0 (0.5)	**67.6** (0.6)	**53.9** (0.7)	**53.9** (0.5)	62.2 (0.5)	**62.6** (0.6)
POS	61.3 (1.0)	**61.4** (0.9)	76.0 (0.4)	**76.2** (0.4)	**74.6** (0.5)	74.5 (0.4)	66.2 (1.0)	**66.8** (0.8)	71.5 (0.6)	**71.6** (0.5)

- Consistently better performance on both head and tail languages.
- Both non-Latin and Latin languages benefit from LangSAMP.

Main Results

- Continued pretraining XLM-R on Glot500-c
- **head**: the language covered by XLM-R.
- **tail**: the language not covered by XLM-R

	tail		head		Latn		non-Latn		all	
	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP	Baseline	LangSAMP
SR-B	36.9 (0.0)	**39.5** (0.0)	60.6 (0.0)	**61.3** (0.0)	40.7 (0.0)	**42.8** (0.0)	51.2 (0.0)	**53.5** (0.0)	42.9 (0.0)	**45.1** (0.0)
SR-T	56.9 (0.0)	**58.6** (0.0)	74.8 (0.0)	**76.1** (0.0)	67.5 (0.0)	**68.7** (0.0)	73.7 (0.0)	**75.6** (0.0)	69.7 (0.0)	**71.1** (0.0)
Taxi1500	47.1 (4.8)	**50.8** (2.4)	59.9 (2.9)	**61.2** (1.2)	48.2 (4.6)	**51.7** (2.1)	58.8 (3.1)	**60.1** (1.7)	50.3 (4.2)	**53.4** (2.0)
SIB200	69.0 (1.4)	**70.2** (1.9)	82.2 (1.4)	**82.6** (1.2)	72.1 (1.3)	**73.1** (1.8)	81.1 (1.5)	**81.7** (1.2)	75.0 (1.3)	**75.9** (1.6)
NER	60.1 (0.6)	**60.8** (0.8)	64.0 (0.6)	**64.1** (0.6)	67.0 (0.5)	**67.6** (0.6)	**53.9** (0.7)	**53.9** (0.5)	62.2 (0.5)	**62.6** (0.6)
POS	61.3 (1.0)	**61.4** (0.9)	76.0 (0.4)	**76.2** (0.4)	**74.6** (0.5)	74.5 (0.4)	66.2 (1.0)	**66.8** (0.8)	71.5 (0.6)	**71.6** (0.5)

- Consistently better performance on both head and tail languages.
- Both non-Latin and Latin languages benefit from LangSAMP.
- Improvements can vary slightly across different task types.

Outline

Ablation Study

	SR-B			SR-T			Taxi1500			SIB200			NER			POS		
	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all
vanilla model	11.9	56.4	23.2	46.0	77.7	68.6	18.1	58.6	28.4	56.1	**83.0**	68.3	55.1	62.8	59.3	49.9	75.7	67.8
w/ $\boldsymbol{E}^{Lang}$	13.1	57.9	24.5	**49.1**	79.0	70.5	18.3	58.5	28.5	57.2	82.7	68.8	**55.2**	**63.0**	**59.5**	49.9	75.8	67.8
w/ $\boldsymbol{E}^{Script}$	12.5	57.4	23.9	48.3	78.4	69.8	18.5	57.0	28.2	56.6	82.1	68.2	55.1	62.4	59.0	**50.8**	**76.2**	**68.4**
w/ $\boldsymbol{E}^{Lang}$ and $\boldsymbol{E}^{Script}$	**13.4**	**58.7**	**24.9**	**49.1**	**79.5**	**70.8**	**20.6**	**58.8**	**30.3**	**57.9**	**83.0**	**69.3**	54.9	61.6	58.6	49.7	75.6	67.6

Ablation Study

	SR-B			SR-T			Taxi1500			SIB200			NER			POS		
	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all
vanilla model	11.9	56.4	23.2	46.0	77.7	68.6	18.1	58.6	28.4	56.1	**83.0**	68.3	55.1	62.8	59.3	49.9	75.7	67.8
w/ $\boldsymbol{E}^{Lang}$	13.1	57.9	24.5	**49.1**	79.0	70.5	18.3	58.5	28.5	57.2	82.7	68.8	**55.2**	**63.0**	**59.5**	49.9	75.8	67.8
w/ $\boldsymbol{E}^{Script}$	12.5	57.4	23.9	48.3	78.4	69.8	18.5	57.0	28.2	56.6	82.1	68.2	55.1	62.4	59.0	**50.8**	**76.2**	**68.4**
w/ $\boldsymbol{E}^{Lang}$ and $\boldsymbol{E}^{Script}$	**13.4**	**58.7**	**24.9**	**49.1**	**79.5**	**70.8**	**20.6**	**58.8**	**30.3**	**57.9**	**83.0**	**69.3**	54.9	61.6	58.6	49.7	75.6	67.6

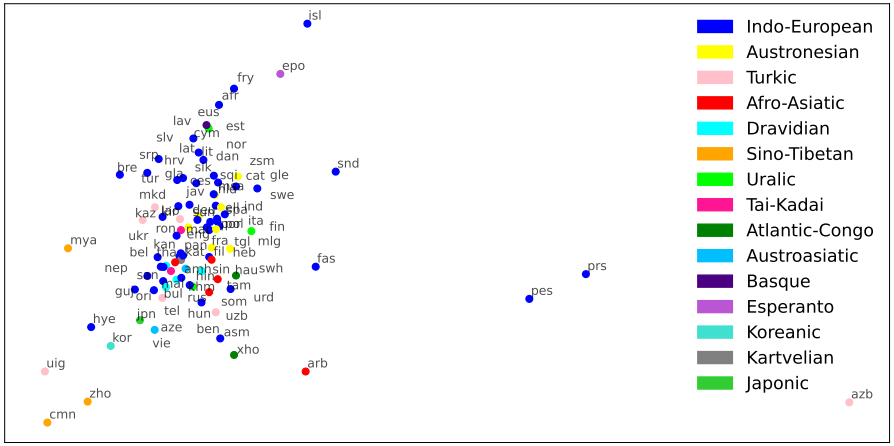
- Both language and script embeddings help.

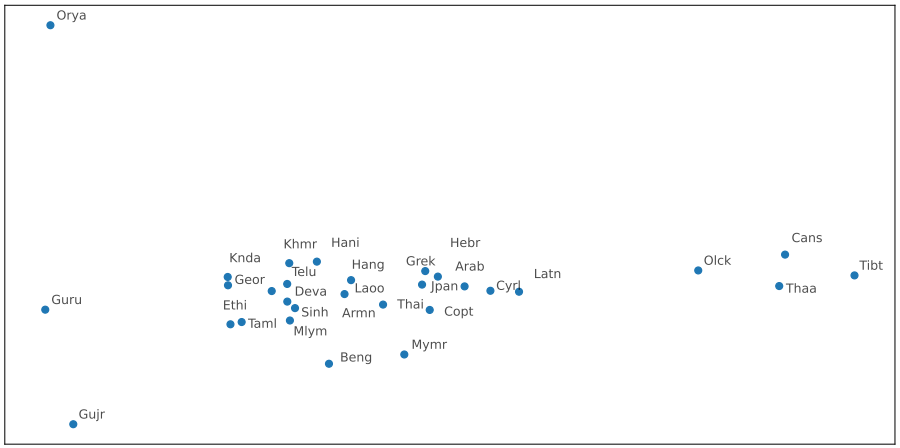
Ablation Study

	SR-B			SR-T			Taxi1500			SIB200			NER			POS		
	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all
vanilla model	11.9	56.4	23.2	46.0	77.7	68.6	18.1	58.6	28.4	56.1	**83.0**	68.3	55.1	62.8	59.3	49.9	75.7	67.8
w/ $\boldsymbol{E}^{Lang}$	13.1	57.9	24.5	**49.1**	79.0	70.5	18.3	58.5	28.5	57.2	82.7	68.8	**55.2**	**63.0**	**59.5**	49.9	75.8	67.8
w/ $\boldsymbol{E}^{Script}$	12.5	57.4	23.9	48.3	78.4	69.8	18.5	57.0	28.2	56.6	82.1	68.2	55.1	62.4	59.0	**50.8**	**76.2**	**68.4**
w/ $\boldsymbol{E}^{Lang}$ and $\boldsymbol{E}^{Script}$	**13.4**	**58.7**	**24.9**	**49.1**	**79.5**	**70.8**	**20.6**	**58.8**	**30.3**	**57.9**	**83.0**	**69.3**	54.9	61.6	58.6	49.7	75.6	67.6

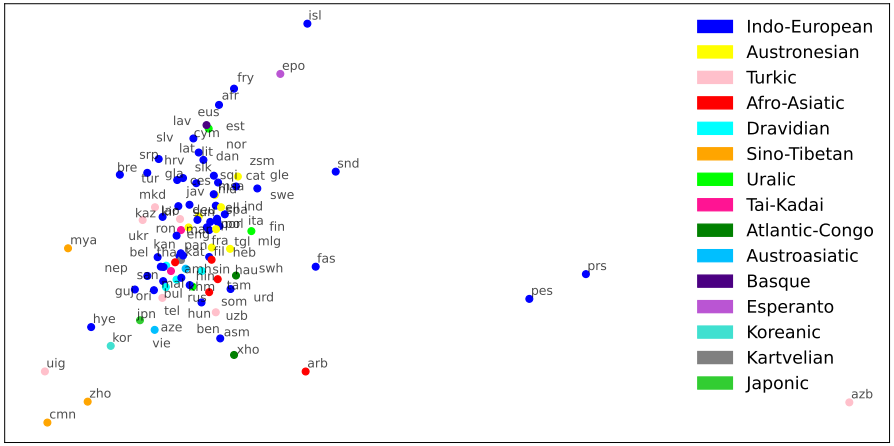
- Both language and script embeddings help.
- Improvement varies across task types.

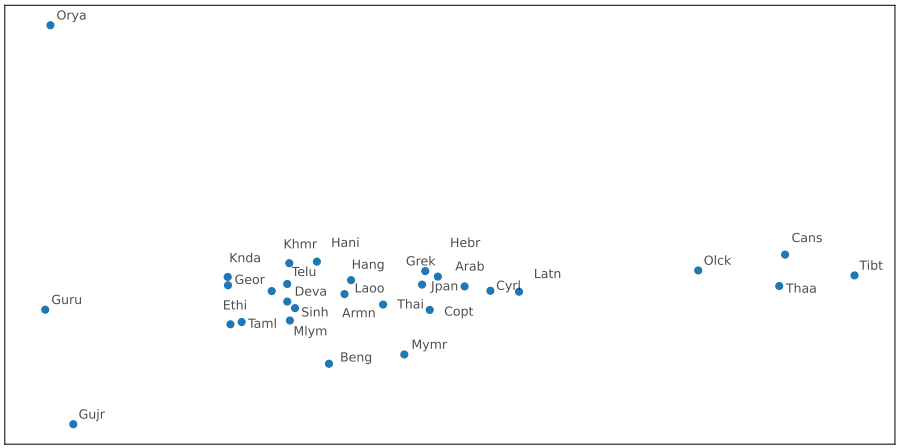
Visualization of Auxiliary Embeddings with PCA





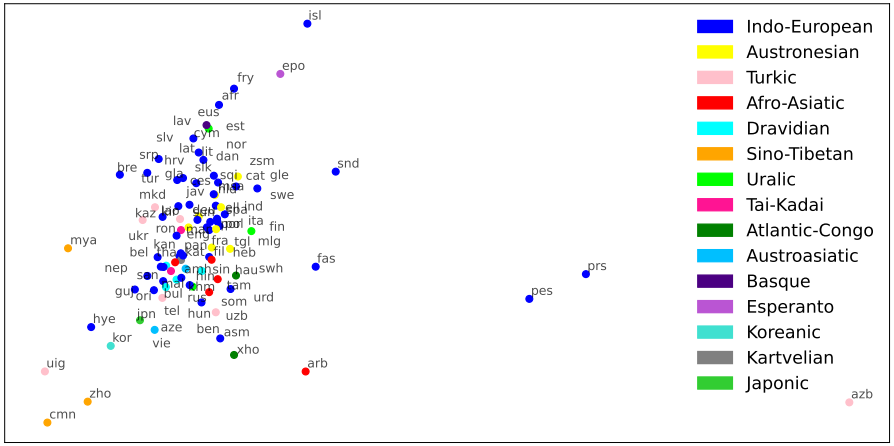
Visualization of Auxiliary Embeddings with PCA

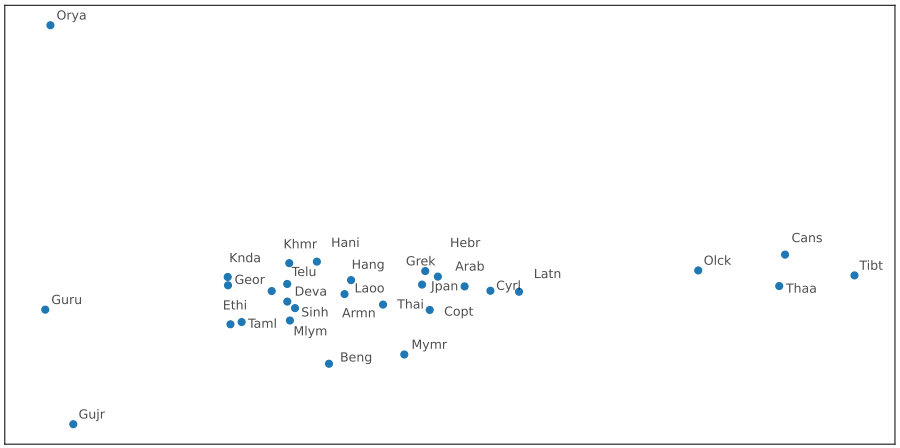




- Similar or related languages/scripts are located close to each other.

Visualization of Auxiliary Embeddings with PCA

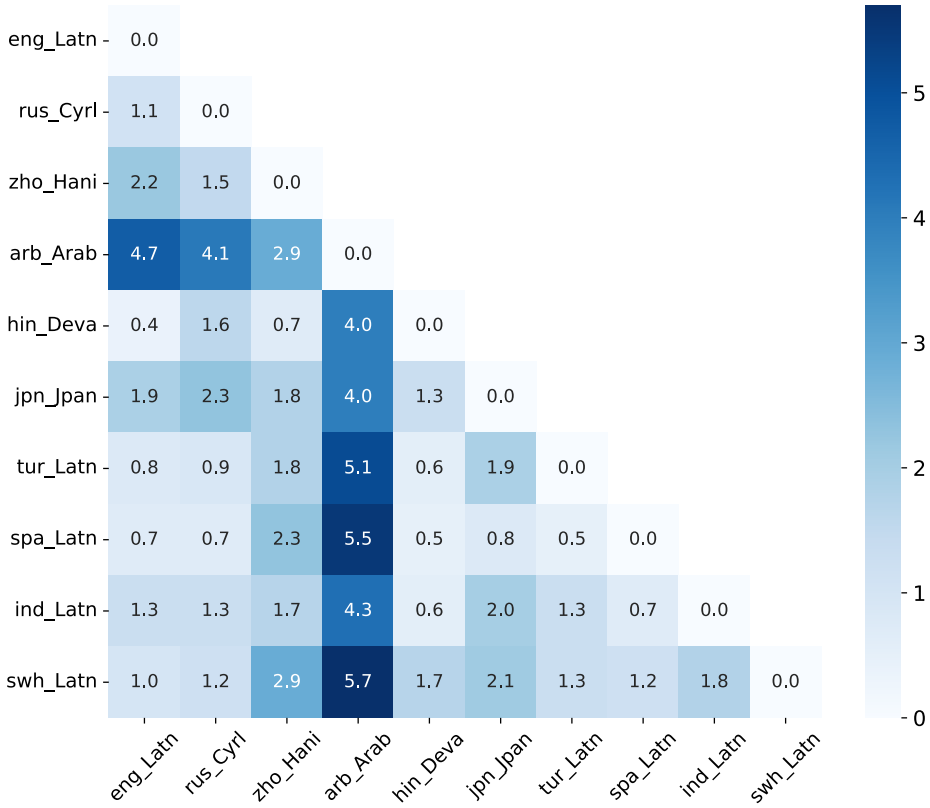




- Similar or related languages/scripts are located close to each other.
- This shows that the auxiliary embeddings capture language- and script-specific information.

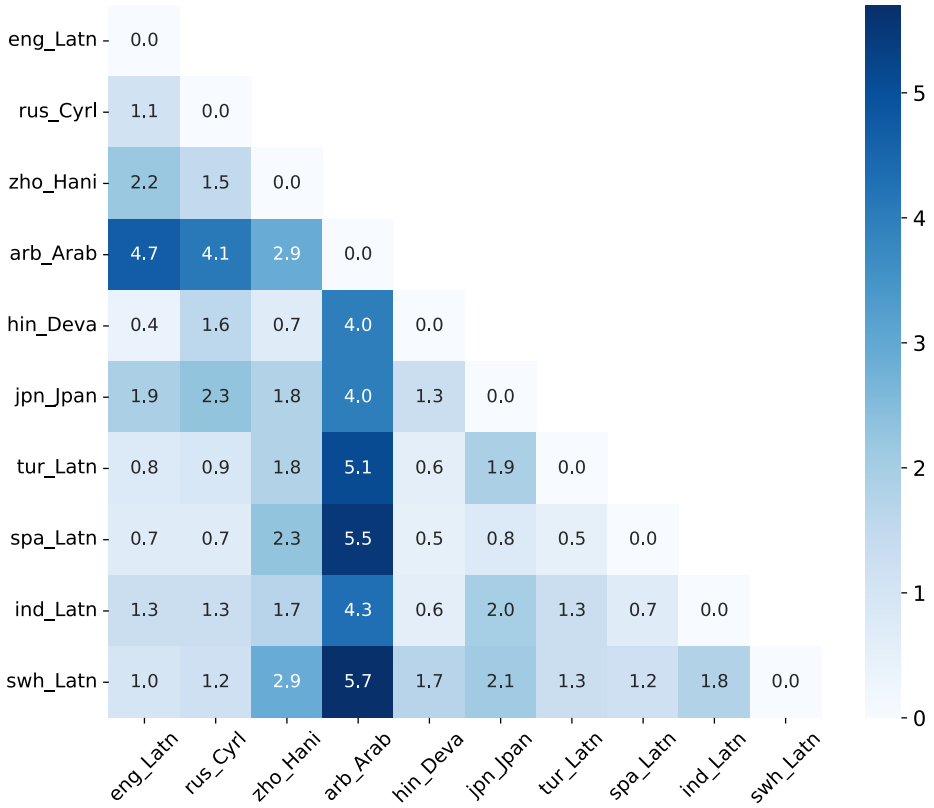
Similarity of Parallel Sentences Across Languages

We selected **10** languages: **eng_Latn**, **rus_Cyrl**, **zho_Hani**, **arb_Arab**, **hin_Deva**, **jpn_Jpan**, **tur_Latn**, **spa_Latn**, **ind_Latn**, and **swa_Latn**. Pairwise cosine similarity of sentence representations is computed (baseline and LangSAMP). Improvement (by percentage) is shown.



Similarity of Parallel Sentences Across Languages

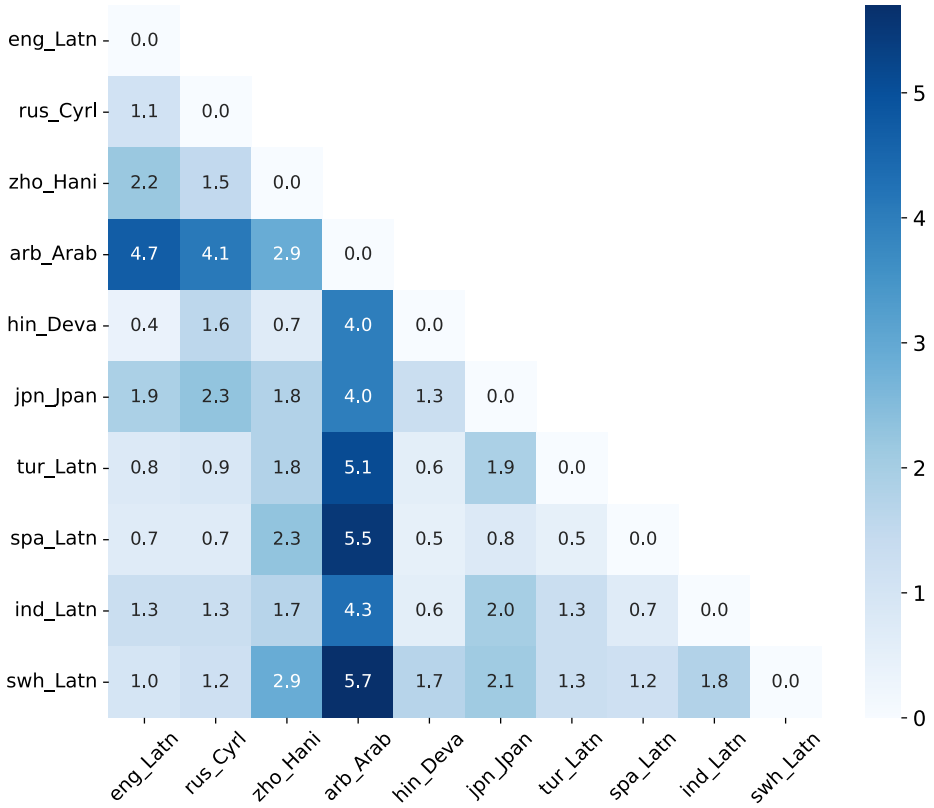
We selected **10** languages: **eng_Latn**, **rus_Cyrl**, **zho_Hani**, **arb_Arab**, **hin_Deva**, **jpn_Jpan**, **tur_Latn**, **spa_Latn**, **ind_Latn**, and **swa_Latn**. Pairwise cosine similarity of sentence representations is computed (baseline and LangSAMP). Improvement (by percentage) is shown.



- Similarity between any two languages is improved in LangSAMP.

Similarity of Parallel Sentences Across Languages

We selected **10** languages: **eng_Latn**, **rus_Cyrl**, **zho_Hani**, **arb_Arab**, **hin_Deva**, **jpn_Jpan**, **tur_Latn**, **spa_Latn**, **ind_Latn**, and **swa_Latn**. Pairwise cosine similarity of sentence representations is computed (baseline and LangSAMP). Improvement (by percentage) is shown.



- Similarity between any two languages is improved in LangSAMP.
- The enhancement is especially noticeable for typologically distinct languages.

Case Study: Source Language Selection

Instead of always using English, we select the **language** as source language whose embedding is most cosine-similar to the target language for *zero-shot crosslingual transfer*.

	tail		head		Latn		non-Latn		all	
	English	Source	English	Source	English	Source	English	Source	English	Source
Taxi1500	47.3	**48.3**	59.1	**60.3**	48.4	**49.0**	58.1	**60.5**	50.2	**51.2**
SIB200	**67.9**	**67.9**	81.2	**81.6**	71.0	**71.1**	80.3	**80.6**	74.0	**74.2**
NER	61.2	**61.7**	64.1	**65.6**	**67.5**	66.9	54.6	**58.5**	62.8	**63.8**
POS	**63.2**	53.8	**77.0**	72.3	**75.5**	68.4	**68.1**	63.6	**72.8**	66.6

Case Study: Source Language Selection

Instead of always using English, we select the **language** as source language whose embedding is most cosine-similar to the target language for *zero-shot crosslingual transfer*.

	tail		head		Latn		non-Latn		all	
	English	Source	English	Source	English	Source	English	Source	English	Source
Taxi1500	47.3	**48.3**	59.1	**60.3**	48.4	**49.0**	58.1	**60.5**	50.2	**51.2**
SIB200	**67.9**	**67.9**	81.2	**81.6**	71.0	**71.1**	80.3	**80.6**	74.0	**74.2**
NER	61.2	**61.7**	64.1	**65.6**	**67.5**	66.9	54.6	**58.5**	62.8	**63.8**
POS	**63.2**	53.8	**77.0**	72.3	**75.5**	68.4	**68.1**	63.6	**72.8**	66.6

- This shows that language embeddings facilitate the selection of optimal source languages for more effective crosslingual transfer.

Conclusion and Future Work

LangSAMP: a multilingual pretraining approach that adds auxiliary language and script embeddings *after* transformer layers to encourage language-neutral representations.

- **Simple yet effective** for improving language-neutrality.
- **Consistent gains** over baselines across downstream tasks.
- **Auxiliary embeddings** may be leveraged as useful byproducts.

Possible Future Directions:

- Further steer *middle-layer* representations toward neutrality using parallel data with contrastive objectives.
- Leverage embeddings for **controlled multilingual generation** in decoder-only or encoder-decoder models.

Thank you for your attention!