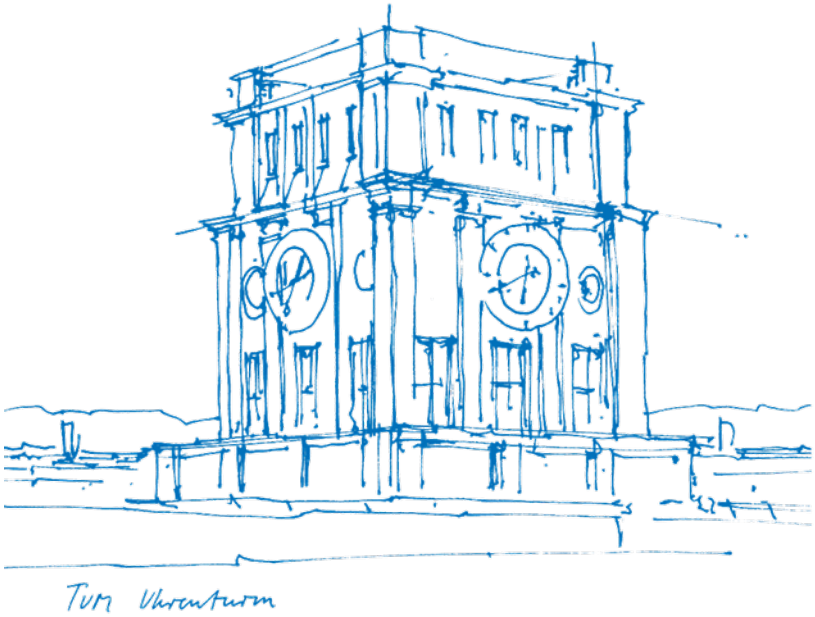
Adaptation to Data Sparsity in Machine Translation and Large Language Models

Wen Lai
Technische Universität München

Supervisor:	Prof. Dr. Alexander Fraser
Mentor:	Prof. Dr. Ivan Titov
Committee:	Prof. Dr. Alexander Fraser, Prof. Dr. Chunyang Chen Prof. Dr. Rico Sennrich, Prof. Dr. Ivan Vulić
Date:	28th November 2025



NLP Group
Chair for Data Analytics & Statistics
TUM School of Computation, Information and Technology

Technical University of Munich

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Munich Center for Machine Learning

Outlines

- Introduction
- Adaptation to Distribution Diversity (Domains, Languages and Styles)
- Adaptation to Data Scarcity (Task-Agnostic Adaptation and Dataset Construction)
- Summary
- Q&A

Introduction

Background: Scaling and Its Limits

- Scaling (model size / compute / data) drove recent NLP progress.
- But scaling presumes abundant, representative data — a hidden assumption.
- When coverage fails, scaling returns diminish quickly.

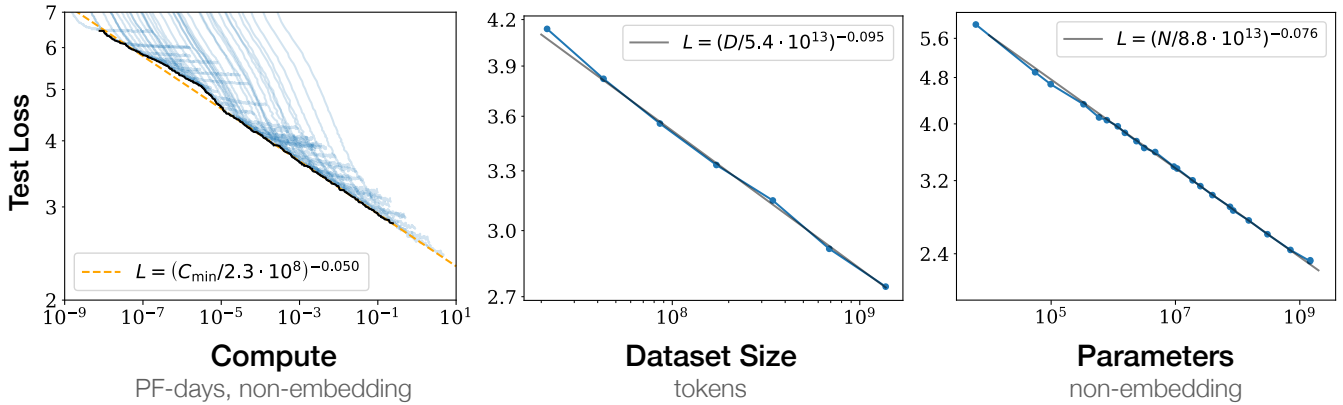


Figure 1: Scaling laws conditional on data richness (Kaplan et al., 2020).

Background: Scaling and Its Limits

- Scaling (model size / compute / data) drove recent NLP progress.
- But scaling presumes abundant, representative data — a hidden assumption.
- When coverage fails, scaling returns diminish quickly.

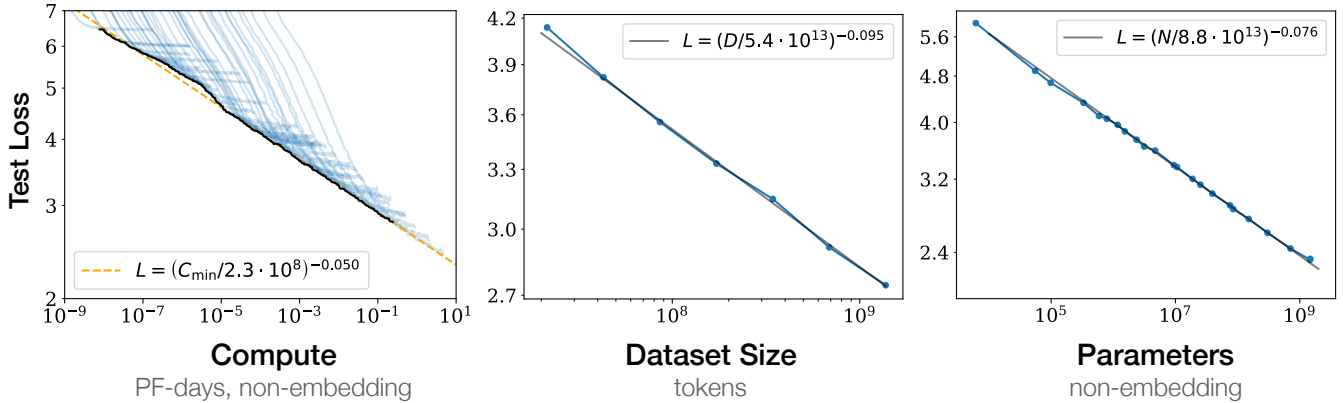


Figure 1: Scaling laws conditional on data richness (Kaplan et al., 2020).

Key insight

Scaling alone is insufficient under poor data coverage.

Problem: Data Sparsity

- **Data sparsity:** insufficient or unrepresentative data across domains, languages, or styles.
  - Pretrained models are brittle and biased on underrepresented varieties.

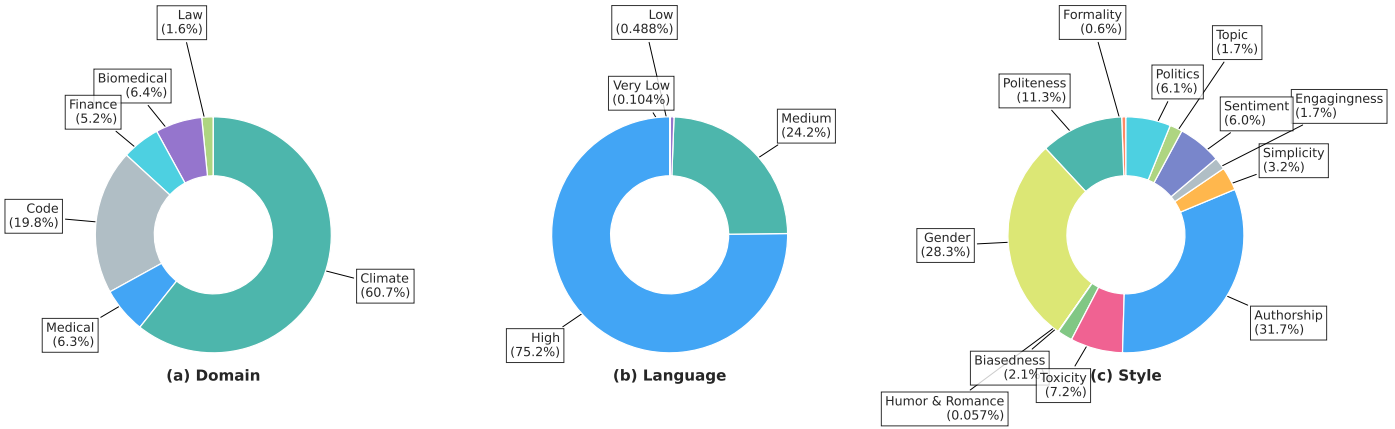


Figure 2: Example of uneven corpora across languages/domains/styles.

How Data Sparsity Manifests in NLP

- **Example 1 – Domain Scarcity:** models fail on spoken/informal inputs.

Figure 3: Example from Google Translate (26th September 2025).

Observation

High-performing systems still fail on underrepresented linguistic varieties.

How Data Sparsity Manifests in NLP

- **Example 2 – Low-Resource Languages:** narrow coverage leads to degraded translation and understanding.

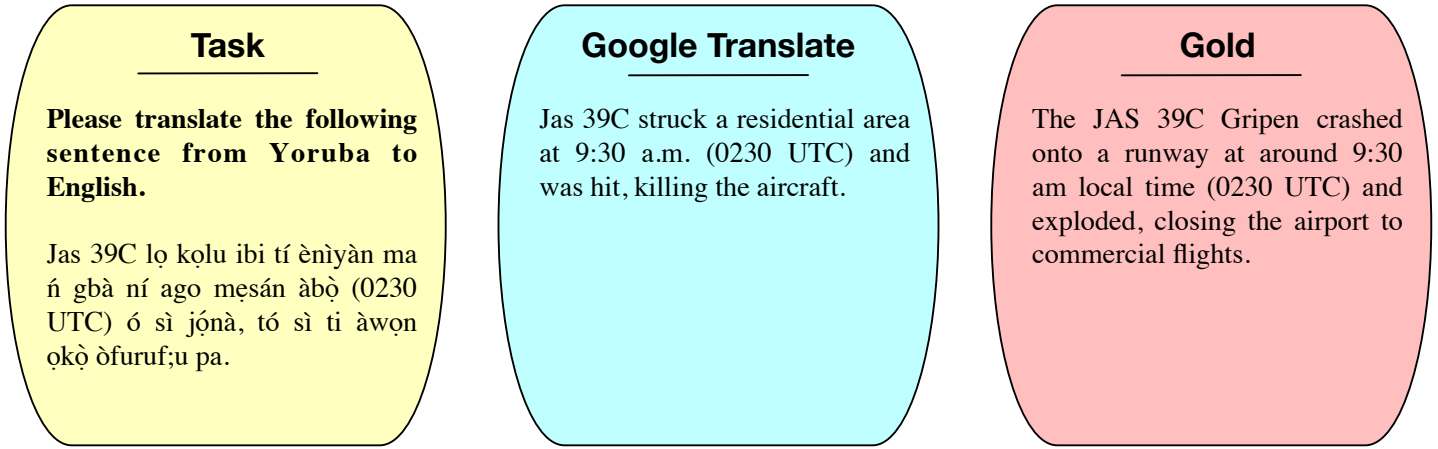


Figure 4: Example from Google Translate (26th September 2025).

Observation

Data imbalance creates systematic inequality in model capability across languages.

Dimensions of Data Sparsity

- **Distributional diversity:** uneven coverage across domains, languages, styles.
- **Data scarcity:** shortage of labeled, domain-specific data; high labeling cost for annotation.

Dimensions of Data Sparsity

- **Distributional diversity:** uneven coverage across domains, languages, styles.
- **Data scarcity:** shortage of labeled, domain-specific data; high labeling cost for annotation.

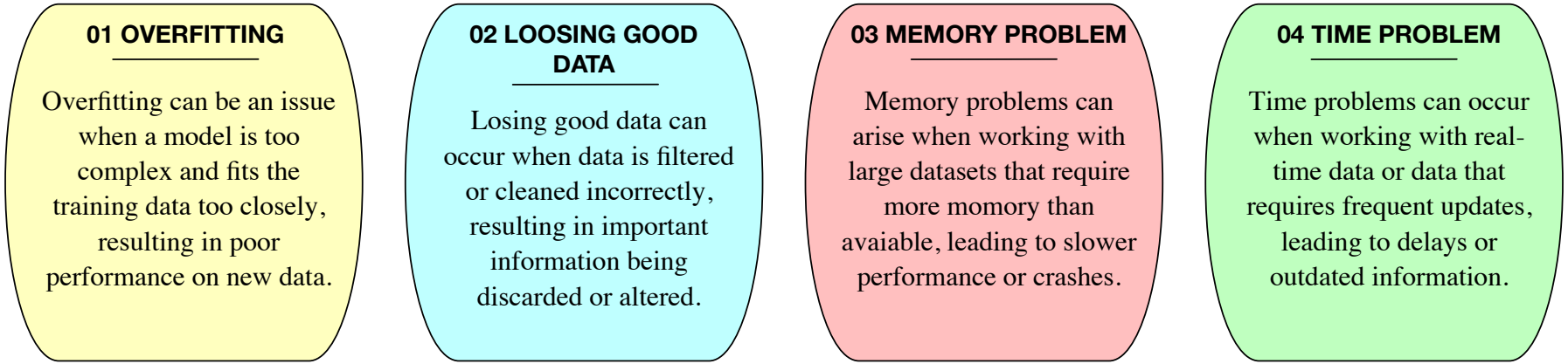


Figure 5: Trade-offs between performance, coverage, cost, efficiency.

Observation

Optimizing all four (performance, coverage, cost, efficiency) is infeasible without adaptation.

Research Focus

(I) Machine Translation (MT)

- **Goal:** reliable source → target translation across languages & domains.
- **Evolution:** rule-based → SMT → NMT → pretraining → LLMs.
- **Challenge:** low-resource languages & domain shift remain major obstacles.

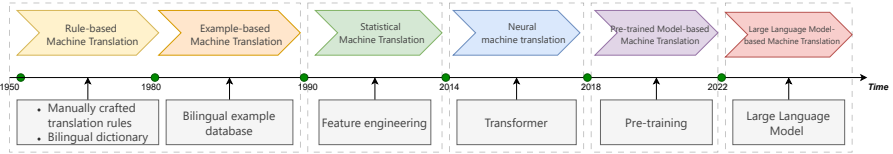


Figure 6: Evolution of machine translation.

Research Focus

(I) Machine Translation (MT)

- **Goal:** reliable source → target translation across languages & domains.
- **Evolution:** rule-based → SMT → NMT → pretraining → LLMs.
- **Challenge:** low-resource languages & domain shift remain major obstacles.

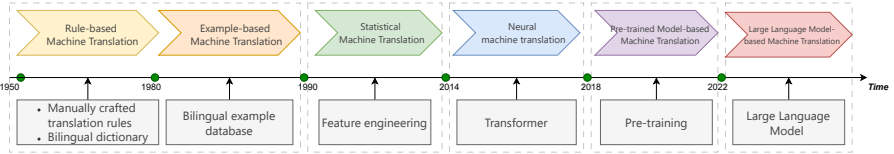


Figure 6: Evolution of machine translation.

(II) Large Language Models (LLMs)

- **Goal:** single backbone for many tasks and languages.
- **Strength:** broad generalization from massive pretraining.
- **Weakness:** performance degrades on rare styles / low-resource inputs.

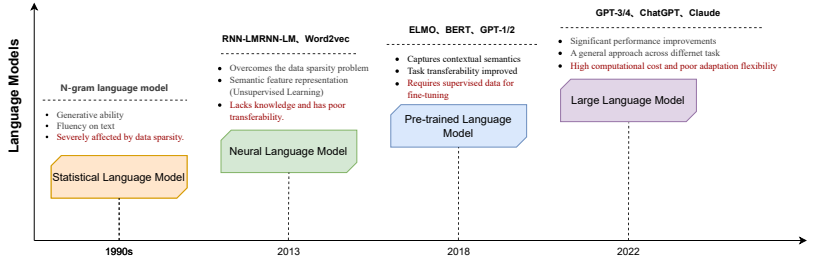


Figure 7: Evolution of large language models.

Why Adaptation (not Training from Scratch)?

- **Cost:** training large models from scratch requires massive compute & data.
- **Practical limits:** privacy, regulation, and labeling scarcity prevent data collection.
- **Stability:** small target datasets cause overfitting if training from scratch.

Why Adaptation (not Training from Scratch)?

- **Cost:** training large models from scratch requires massive compute & data.
- **Practical limits:** privacy, regulation, and labeling scarcity prevent data collection.
- **Stability:** small target datasets cause overfitting if training from scratch.

Core Idea

Adaptation is the key:
Leverage pretrained models and apply targeted techniques to handle data sparsity.

Figure 8: *"If I have seen further than others, it is by standing upon the shoulders of giants."* (Isaac Newton)

Research Goals & Thesis Theme

- **Goal:** Develop adaptation strategies for MT and LLMs under data sparsity.
  - Methods to improve **robustness** to domain & language shift.
  - Techniques to increase **efficiency** (lower resource adaptation).
  - Approaches that boost **generalization** across unseen styles/languages.

Research Goals & Thesis Theme

- **Goal:** Develop adaptation strategies for MT and LLMs under data sparsity.
  - Methods to improve **robustness** to domain & language shift.
  - Techniques to increase **efficiency** (lower resource adaptation).
  - Approaches that boost **generalization** across unseen styles/languages.
- **Theme:** ADaS — Adaptation to Data Sparsity.

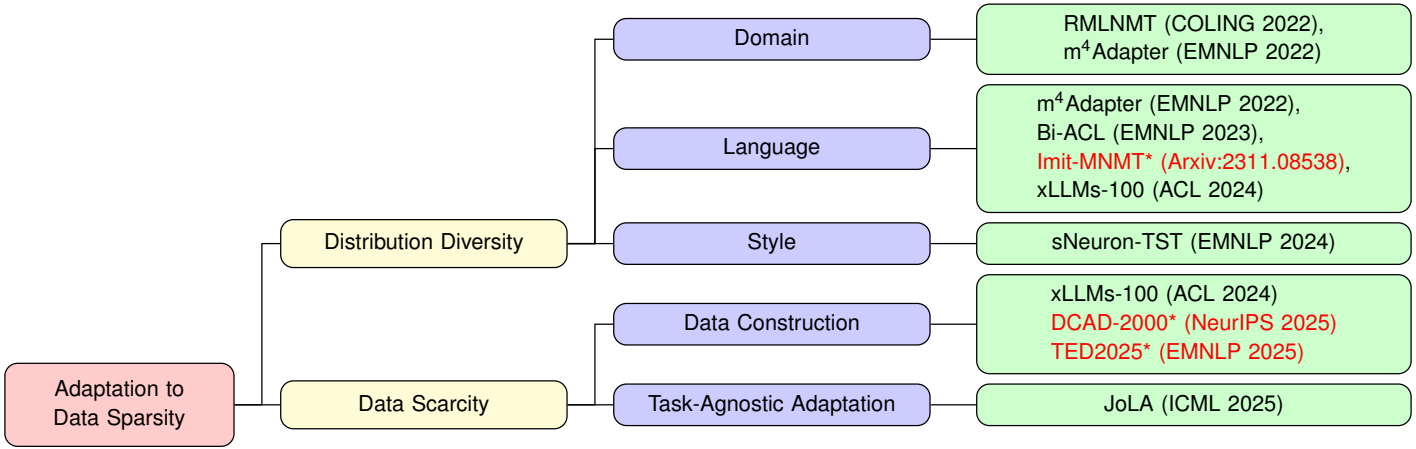


Figure 9: A taxonomy of studies conducted in this dissertation on ADaS.

Adaptation to Distribution Diversity (Part I: Domains)

- **Research Papers:**
  - **Wen Lai**, Jindřich Libovický, and Alexander Fraser. 2022. Improving Both Domain Robustness and Domain Adaptability in Machine Translation. (COLING 2022)
  - **Wen Lai**, Alexandra Chronopoulou, and Alexander Fraser. 2022. m4Adapter: Multilingual Multi-Domain Adaptation for Machine Translation with a Meta-Adapter. (EMNLP 2022)
- **Research Focuses:**
  - Machine Translation + Domain Adaptation (COLING 2022).
  - Multilingual Machine Translation + Domain Adaptation (EMNLP 2022). ⇐ **Present Today!**

m^4*Adapter*: Multilingual Multi-Domain Adaptation for Machine Translation with a Meta-Adapter

(Presented at EMNLP 2022; Abu Dhabi, United Arab Emirates)

Motivation & Research Question

- **Problem:** Real-world NMT requires adaptation to both:
  - New languages (e.g., low-resource or unseen language pairs)
  - New domains (e.g., specialized technical content, informal text)

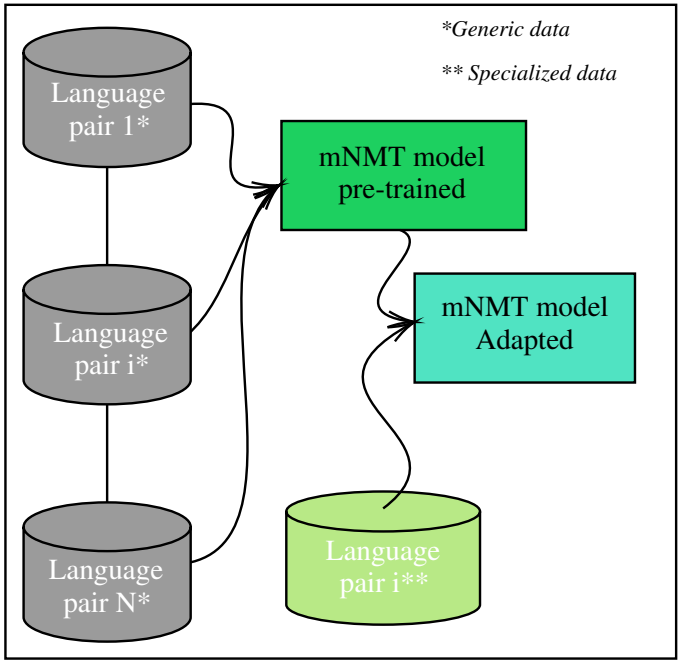


Figure 13: Multilingual Multi-Domain Adaptation.

Motivation & Research Question

- **Problem:** Real-world NMT requires adaptation to both:
  - New languages (e.g., low-resource or unseen language pairs)
  - New domains (e.g., specialized technical content, informal text)
- **Existing approaches:**
  - Fine-tuning: suffers from catastrophic forgetting and inefficiency
  - Adapters: require separate units for each language and domain
  - Both treat language and domain adaptation separately

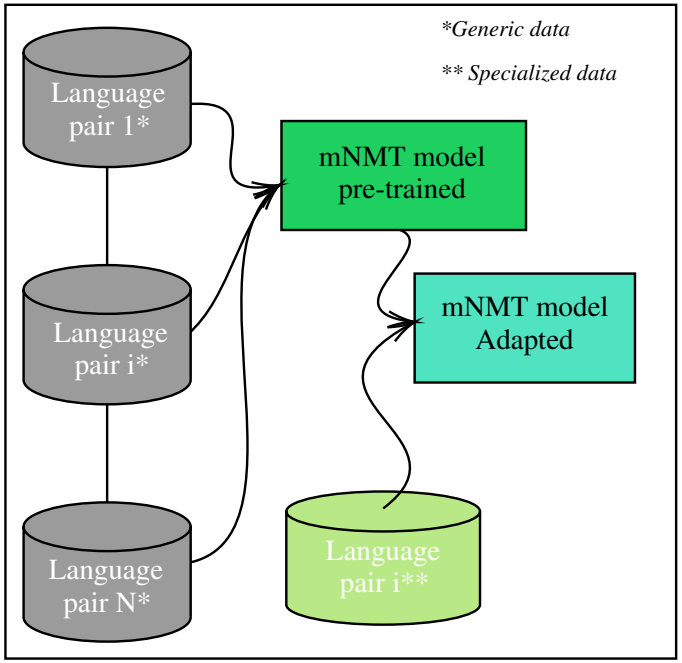


Figure 13: Multilingual Multi-Domain Adaptation.

Motivation & Research Question

- **Problem:** Real-world NMT requires adaptation to both:
  - New languages (e.g., low-resource or unseen language pairs)
  - New domains (e.g., specialized technical content, informal text)
- **Existing approaches:**
  - Fine-tuning: suffers from catastrophic forgetting and inefficiency
  - Adapters: require separate units for each language and domain
  - Both treat language and domain adaptation separately

Research Question

Can we simultaneously adapt to new languages and domains while transferring knowledge across both dimensions?

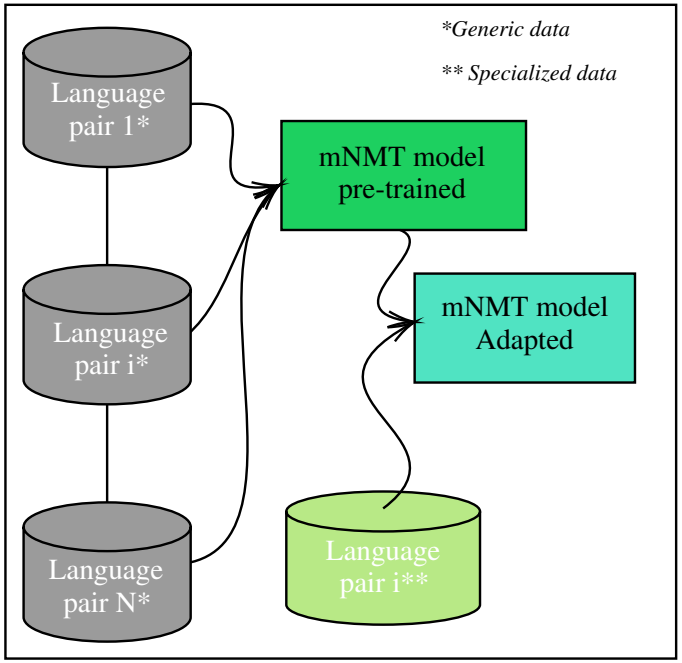


Figure 13: Multilingual Multi-Domain Adaptation.

Our Solution: *m*^{4}*Adapter*

- **Novel framework:**
  - *m*^{4}*Adappter*: Multilingual Multi-Domain Machine Translation with Meta-Adapter.
  - Combines meta-learning with parameter-efficient adapters.

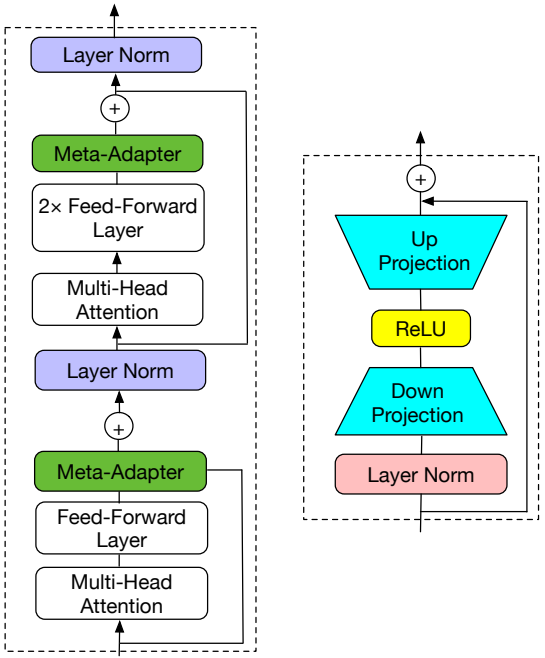


Figure 14: Meta-Adapter architecture

Our Solution: *m*^{4}*Adapter*

- **Novel framework:**
  - *m*^{4}*Adappter*: Multilingual Multi-Domain Machine Translation with Meta-Adapter.
  - Combines meta-learning with parameter-efficient adapters.
- **Key insight:** Language and domain knowledge can be encoded in a shared representation space.

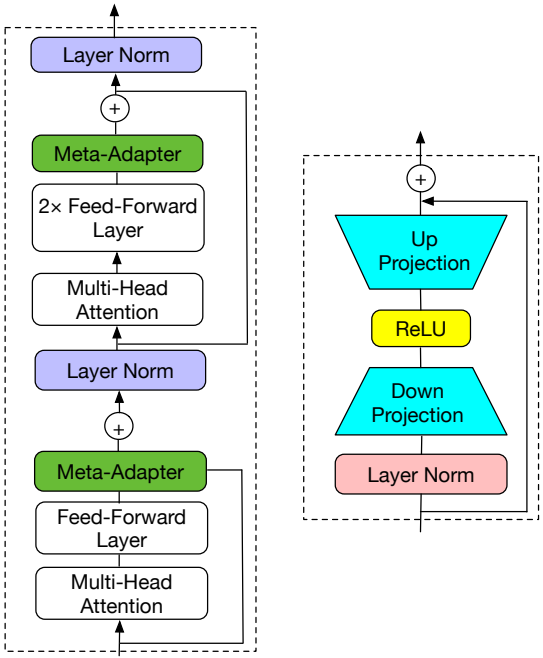


Figure 14: Meta-Adapter architecture

Our Solution: *m*^{4}*Adapter*

- **Novel framework:**
  - *m*^{4}*Adappter*: Multilingual Multi-Domain Machine Translation with Meta-Adapter.
  - Combines meta-learning with parameter-efficient adapters.
- **Key insight:** Language and domain knowledge can be encoded in a shared representation space.
- **Two-phase approach:**
  - **Meta-Training:** Learn generalizable parameters using diverse Domain-Language Pairs (DLPs)
  - **Meta-Adaptation:** Efficiently adapt to new DLPs with minimal data

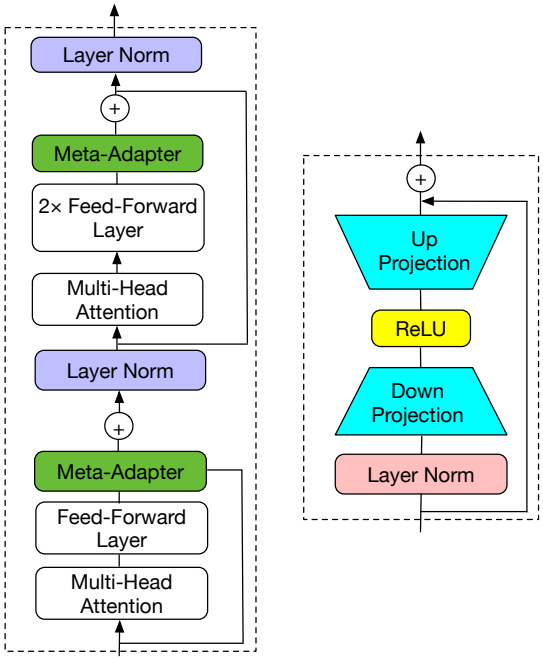


Figure 14: Meta-Adapter architecture

Our Solution: *m*^{4}*Adapter*

- **Novel framework:**
  - *m*^{4}*Adappter*: Multilingual Multi-Domain Machine Translation with Meta-Adapter.
  - Combines meta-learning with parameter-efficient adapters.
- **Key insight:** Language and domain knowledge can be encoded in a shared representation space.
- **Two-phase approach:**
  - **Meta-Training:** Learn generalizable parameters using diverse Domain-Language Pairs (DLPs)
  - **Meta-Adaptation:** Efficiently adapt to new DLPs with minimal data

Key Advantage

m^4*Adapter* uses only 0.75% of parameters compared to full fine-tuning, enabling efficient cross-task knowledge transfer.

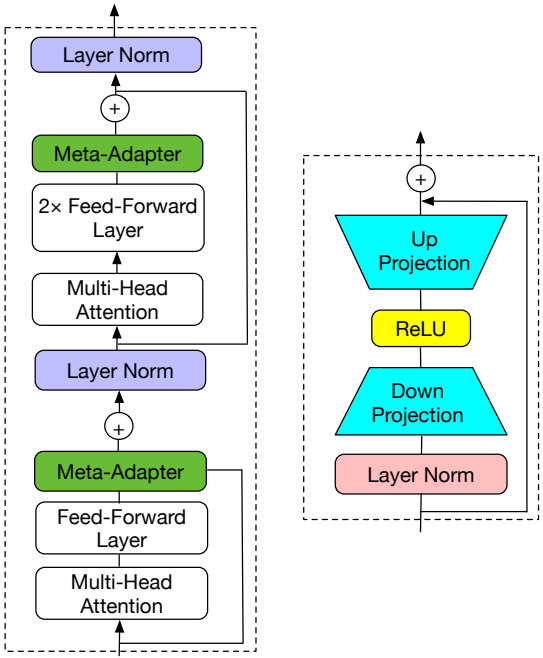


Figure 14: Meta-Adapter architecture

Experimental Setup

- **Datasets:**
  - Two groups: *meta-training* and *meta-adapting*.
  - Domains from OPUS: *EUbookshop, KDE, OpenSubtitles, QED, TED, Ubuntu, Bible, UN, Tanzil, Infopankki*.
  - Languages (ISO 639-1): *en, de, fr, mk, sr, et, hr, hu, fi, uk, is, lt, ar, es, ru, zh*.

Experimental Setup

- **Datasets:**
  - Two groups: *meta-training* and *meta-adapting*.
  - Domains from OPUS: *EUbookshop, KDE, OpenSubtitles, QED, TED, Ubuntu, Bible, UN, Tanzil, Infopankki*.
  - Languages (ISO 639-1): *en, de, fr, mk, sr, et, hr, hu, fi, uk, is, lt, ar, es, ru, zh*.

- **Baselines:**
  - **m2m**: Original m2m model (Fan et al., 2021).
  - **m2m + FT**: m2m fine-tuned on all DLPs.
  - **m2m + tag**: m2m with domain tags (Kobus et al., 2017).
  - **agnostic-adapter**: Mixed-DLP adapters (Cooper Stickland et al., 2021a).
  - **stack-adapter**: Stacked language and domain adapters (Cooper Stickland et al., 2021b).
  - **meta-learning**: MAML algorithm on all DLPs (Sharaf et al., 2020).

Experimental Setup

- **Datasets:**
  - Two groups: *meta-training* and *meta-adapting*.
  - Domains from OPUS: *EUbookshop, KDE, OpenSubtitles, QED, TED, Ubuntu, Bible, UN, Tanzil, Infopankki*.
  - Languages (ISO 639-1): *en, de, fr, mk, sr, et, hr, hu, fi, uk, is, lt, ar, es, ru, zh*.
- **Evaluation Metrics:**
  - BLEU Score for domain robustness and adaptation.

- **Baselines:**
  - **m2m**: Original m2m model (Fan et al., 2021).
  - **m2m + FT**: m2m fine-tuned on all DLPs.
  - **m2m + tag**: m2m with domain tags (Kobus et al., 2017).
  - **agnostic-adapter**: Mixed-DLP adapters (Cooper Stickland et al., 2021a).
  - **stack-adapter**: Stacked language and domain adapters (Cooper Stickland et al., 2021b).
  - **meta-learning**: MAML algorithm on all DLPs (Sharaf et al., 2020).

Results: Meta-Training Stage

- Motivated by Lai et al. (2022), domain robustness can be evaluated during meta-training stage.

	BLEU	specific domain		
		TED	Ubuntu	KDE
m2m	18.18	16.20	20.61	22.04
m2m + FT	20.84	17.53	28.81	29.19
m2m + tag	22.70	18.70	**31.86**	31.53
agnostic-adapter	23.70	**19.82**	31.07	32.74
stack-adapter	21.06	18.34	29.17	30.26
meta-learning	20.01	17.57	28.11	28.59
m*^{4}*Adapter	**23.89**	19.77	31.46	**32.91**

Table 4: Domain robustness during meta-training.

Findings

- *m*^{4}*Adapter* matches or exceeds *agnostic-adapter* in domain robustness.
- Outperforms *m2m + tag*, previously shown to be most robust.

Results (Meta-Testing Stage)

- Similarly, domain adaptability can be evaluated during meta-testing stage.

	DLP (meta-adaptation domain)			specific DLP					
	UN	Tanzil	Infopankki	UN-ar-en	Tanzil-ar-en	Infopankki-ar-en	UN-ar-ru	Tanzil-ar-ru	Infopankki-ar-ru
m2m	32.28	8.72	17.40	38.94	6.44	22.57	22.96	3.64	15.05
m2m + FT	29.93	8.26	15.88	35.11	6.85	21.33	19.10	3.05	14.19
m2m + tag	29.88	8.06	15.93	34.39	6.63	20.12	19.37	2.65	13.68
agnostic-adapter	30.56	8.42	17.36	36.13	6.12	23.08	20.64	3.63	14.96
stack-adapter	29.64	8.14	17.19	35.31	5.83	22.14	19.17	2.34	13.85
meta-learning	32.21	7.02	16.73	37.13	5.50	18.91	22.68	1.70	15.23
m*^{4}*Adapter	**33.53**	**9.87**	**18.43**	**39.05**	**8.56**	**23.21**	**25.22**	**4.33**	**17.48**
Δ	+1.25	+1.15	+1.03	+0.11	+2.12	+0.64	+2.26	+0.69	+2.43

Table 5: Domain robustness during meta-testing.

Findings

- *m*^{4}*Adapter* successfully adapts to new domains and language pairs simultaneously.
- All baselines fail to outperform the original m2m model, showing insufficient knowledge transfer.

Results (Domain Transfer via Languages)

- We define domain transfer via languages, i.e., the ability to transfer domains while keeping the languages unchanged.

	meta-adaptation domain							specific DLP (hr-sr)						
	EUbookshop	KDE	OpenSubtitles	QED	TED	Ubuntu	Bible	EUbookshop	KDE	OpenSubtitles	QED	TED	Ubuntu	Bible
m2m	17.77	22.05	14.13	18.34	16.20	20.62	9.80	11.43	25.37	19.01	12.25	8.14	22.33	2.01
m2m + FT	12.73	24.56	16.22	20.46	18.74	31.32	11.30	9.79	21.05	53.34	23.87	20.81	**34.08**	12.57
m2m + tag	13.03	25.34	16.12	17.75	17.04	26.29	11.49	10.13	29.64	49.54	19.78	20.43	34.15	13.25
agnostic-adapter	16.24	**25.85**	17.90	21.71	**20.08**	**31.53**	11.75	9.05	**30.64**	54.04	22.79	21.19	28.83	10.59
stack-adapter	13.25	24.19	17.21	19.56	18.37	28.27	10.38	10.55	24.50	42.94	22.02	20.95	25.41	10.14
meta-learning	13.61	24.91	16.22	17.70	16.40	24.93	11.84	7.90	27.85	52.50	20.41	19.00	31.24	10.42
m*^{4}*Adapter	**18.99**	25.22	**17.94**	**21.71**	19.86	31.37	**12.12**	**12.05**	30.49	**54.30**	**23.92**	**21.32**	33.71	**13.69**
Δ	+2.75	-0.63	+0.04	+0.00	-0.22	-0.16	+0.37	+3.00	-0.15	+0.26	+1.13	+0.13	+4.88	+3.1

Table 6: Domain transfer via languages.

Findings

- Most systems transfer language knowledge to new domains (all outperform m2m).
- *m*^{4}*Adapter* comparable to best baseline (*agnostic-adapter*).
- Particularly effective on distant domains.

Results (Language Transfer via Domains)

- We define language transfer via domains, i.e., the ability to transfer languages while keeping the domains unchanged.

	meta-adaptation language pair				**specific DLP (de-en)**					
	de-en	en-fr	fi-uk	is-lt	EUbookshop	KDE	OpenSubtitles	QED	TED	Ubuntu
m2m	24.52	29.20	12.34	12.55	19.59	26.48	15.89	26.34	28.14	30.65
m2m + FT	23.29	24.44	11.29	9.59	16.04	23.17	13.34	21.39	26.20	39.59
m2m + tag	22.52	24.97	11.71	11.22	15.86	23.67	11.72	20.64	25.97	37.25
agnostic-adapter	28.33	**30.93**	**15.42**	**14.38**	20.16	**28.72**	17.97	**27.66**	**33.63**	41.89
stack-adapter	23.37	24.96	11.51	11.09	16.14	22.51	13.84	22.29	27.67	36.73
meta-learning	25.08	28.26	13.40	12.83	17.88	21.20	16.32	24.96	30.32	39.81
m*^{4}*Adapter	**28.37**	30.80	15.24	14.05	**20.20**	28.19	**18.06**	27.18	33.32	**43.24**
Δ	+0.04	-0.13	-0.18	-0.33	+0.04	-0.53	+0.09	-0.48	-0.31	+1.35

Table 7: Language transfer via domains.

Findings

- Traditional fine-tuning approaches fail to transfer domain knowledge to new languages.
- *m*^{4}*Adapter* matches or exceeds the most competitive baseline (*agnostic-adapter*).

Summary

- *m*^{4}*Adapter* addresses multiple dimensions of data sparsity:
  - **Robustness:** Successfully handles domain and language shift simultaneously.
  - **Efficiency:** Uses only a few parameters compared to full fine-tuning.
  - **Generalization:** Enables knowledge transfer across both language and domain dimensions.
- **Practical impact:** Enables rapid adaptation in low-resource scenarios with minimal data (500 samples).

Paper Code Blog

EMNLP 2022

Adaptation to Distributional Diversity (Part II: Languages)

- **Research Papers:**
  - **Wen Lai**, Alexandra Chronopoulou, Alexander Fraser. Mitigating Data Imbalance and Representation Degeneration in Multilingual Machine Translation. (EMNLP 2023)
  - **Wen Lai**, Mohsen Mesgar, Alexander Fraser. LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback. (ACL 2024)
- **Research Focuses:**
  - Adapting multilingual machine translation to low-resource languages (EMNLP 2023).
  - Adapting large language models to low-resource languages (ACL 2024). ⇐ **Present Today!**

LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback

(Presented at ACL 2024; Bangkok, Thailand)

Motivation & Research Question

- Current LLMs show severe performance disparities across languages
  - Most models primarily trained on high-resource language (30-95% of pretraining data)
  - Low-resource languages severely underrepresented

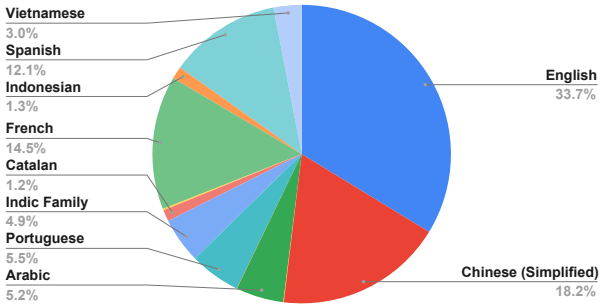


Figure 17: Language distribution in BLOOM (Workshop et al., 2022).

Motivation & Research Question

- Current LLMs show severe performance disparities across languages
  - Most models primarily trained on high-resource language (30-95% of pretraining data)
  - Low-resource languages severely underrepresented
- Two critical multilingual capabilities:
  - **Understanding Capability:** means understand instructions in diverse languages.
  - **Generation Capability:** means produce correct outputs in target languages.

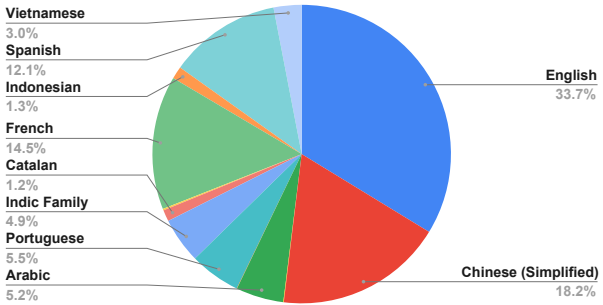


Figure 17: Language distribution in BLOOM (Workshop et al., 2022).

Motivation & Research Question

- Current LLMs show severe performance disparities across languages
  - Most models primarily trained on high-resource language (30-95% of pretraining data)
  - Low-resource languages severely underrepresented
- Two critical multilingual capabilities:
  - **Understanding Capability:** means understand instructions in diverse languages.
  - **Generation Capability:** means produce correct outputs in target languages.

Research Question

- How can we adapt LLMs to understand and generate content in 100+ languages without retraining from scratch?

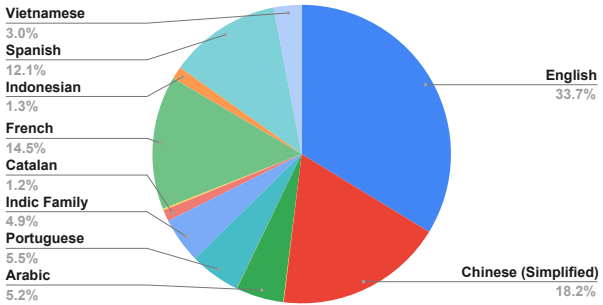


Figure 17: Language distribution in BLOOM (Workshop et al., 2022).

Our Solution: xLLMs-100

Dataset 1: Multilingual Instruction Dataset

- **Goal:** 100-language instruction dataset (FLORES-101)
- **Structure:** <instruction, input, response> triplets adapted from Self-Instruct
- **Translation Method:**
  - Google Translate API for supported languages
  - NLLB model for remaining languages
- **Response Generation:**
  - High-quality languages: Direct generation via ChatGPT
  - Low-quality languages: Translation-based approach

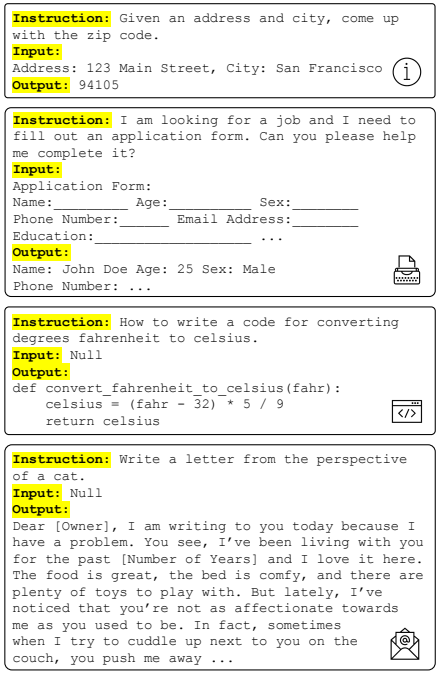


Figure 18: Example from Self-Instruct (Wang et al., 2023).

Our Solution: xLLMs-100

Dataset 2: Cross-Lingual Human Feedback Dataset

- **Goal:** Create human feedback dataset across 30 languages
- **Structure:** <instruction, input, positive response, negative response>
- **Innovation:** Cross-lingual format
  - Traditional: Same language for instruction and output
  - **Our approach:** Different languages for instruction and output
  - Enables $30 \times 29 = 870$ generation scenarios
  - Quality assured by ChatGPT

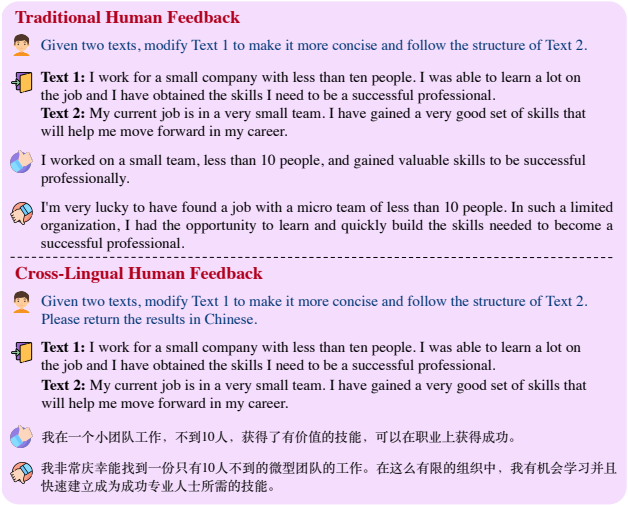


Figure 19: Example from cross-lingual human feedback.

Our Solution: xLLMs-100

- **Supervised Fine-Tuning (SFT)**
  - Aligns model with instruction-following data
- **Direct Preference Optimization (DPO)**
  - Simpler alternative to RLHF (Ouyang et al., 2022).
  - Optimizes preferences without reward models (Rafailov et al., 2023).

Training Process

- Fine-tune LLaMA-2/BLOOM on multilingual instruction data (SFT)
- Apply DPO with cross-lingual human feedback

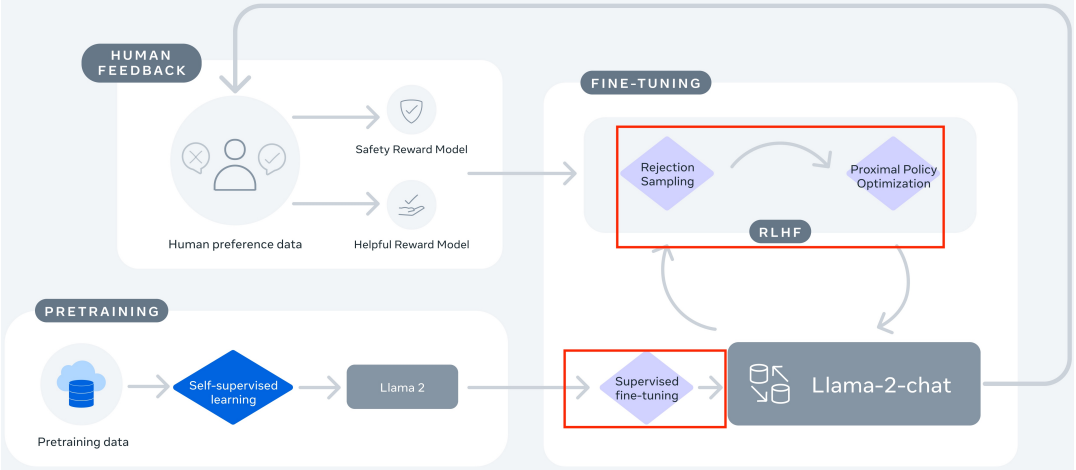


Figure 20: Source from Touvron et al., 2023 (Llama 2)

Experimental Setups

- **Datasets:**
  - **Understanding**: PAWS-X (Yang et al., 2019).
  - **Generation**: FLORES-101 (Goyal et al., 2022), XL-Sum (Hasan et al., 2021).
  - **Reasoning**: XCOPA (Ponti et al., 2020).
  - **Expert-written**: Self-Instruct* (Wang et al., 2023): Aligning LMs with Self-Generated Instructions. (translated to 5 high-resource: Ar, Cs, De, Zh, Hi; 5 low-resource: Hy, Ky, Yo, Ta, Mn)
- **Baselines:**
  - **Off-the-shelf LLMs**: LLaMA-2 (Touvron et al., 2023), BLOOM (Workshop et al., 2022).
  - **Multilingual instruction-tuned**: Bactrian-X (Li et al., 2023). (52 languages, variants: BX_{LLaMA}, BX_{BLOOM}).
  - **Our models (xLLMs-100)**: $\text{SFT}_{\text{LLaMA}}$, $\text{SFT}_{\text{BLOOM}}$. (fine-tuned on our multilingual instruction dataset)
- **Evaluation Metrics:**
  - **FLORES-101**: Case-sensitive detokenized BLEU.
  - **XCOPA, PAWS-X**: Accuracy.
  - **XL-Sum, Self-Instruct***: Multilingual ROUGE-1 (Lin, 2004).

Results (Understanding & Generation Capabilities)

Understanding Capabilities												
	PAWS-X	XCOPA		Self-Instruct*		XL-Sum			FLORES(f)		FLORES(t)	
		low	high	low	high	low	mid	high	low	high	low	high
LLaMA	38.10	47.44	47.22	7.09	12.57	4.07	5.44	2.84	3.07	4.95	2.96	6.61
BX_{LLaMA}	37.28	49.53	49.00	6.31	11.88	2.17	5.52	7.89	2.69	2.38	3.15	5.31
SFT_{LLaMA}	42.32	50.19	49.86	7.32	12.72	4.70	7.34	7.55	3.13	3.93	3.16	6.92
xLLMs-100	**46.95**	**51.53**	**51.96**	**12.94**	**15.35**	**8.83**	**13.90**	**17.29**	**3.27**	**8.09**	**4.04**	**14.18**
BLOOM	36.47	44.27	49.14	7.56	8.67	9.03	14.06	16.80	2.54	2.04	2.05	2.56
BX_{BLOOM}	36.42	46.28	50.35	4.81	8.11	4.89	8.47	11.71	2.14	1.74	2.41	1.57
SFT_{BLOOM}	36.67	49.42	52.31	6.31	11.88	5.62	10.12	14.33	**3.12**	3.79	2.62	2.52
xLLMs-100	**39.83**	**52.50**	**55.59**	**7.94**	**13.35**	**12.87**	**15.23**	**18.38**	3.02	**4.71**	**3.94**	**6.54**
Generating Capabilities												
	PAWS-X	XCOPA		Self-Instruct*		XL-Sum			FLORES(f)		FLORES(t)	
		low	high	low	high	low	mid	high	low	high	low	high
LLaMA	50.22	49.33	51.52	5.38	8.81	6.26	5.80	8.08	1.35	3.90	2.11	4.95
BX_{LLaMA}	48.41	48.00	49.85	7.01	9.80	1.11	2.74	1.70	1.56	5.33	1.37	1.61
SFT_{LLaMA}	50.36	48.93	50.05	7.10	12.15	4.51	6.06	9.21	2.42	4.56	2.71	7.29
xLLMs-100	**61.94**	**49.71**	**54.68**	**9.16**	**14.71**	**9.99**	**13.57**	**16.61**	**2.89**	**9.07**	**5.64**	**16.98**
BLOOM	47.39	49.85	49.47	4.07	7.01	6.08	7.77	8.91	0.78	1.20	0.99	1.49
BX_{BLOOM}	47.26	47.72	49.98	5.88	8.21	1.98	3.59	4.58	0.47	0.82	1.95	2.33
SFT_{BLOOM}	48.50	49.13	49.28	7.78	11.51	3.89	8.87	10.89	2.59	3.12	2.05	2.56
xLLMs-100	**50.53**	**52.36**	**52.26**	**10.17**	**13.62**	**8.77**	**11.74**	**12.36**	**3.97**	**5.79**	**4.22**	**7.68**

Table 10: Understanding and generating capability of LLMs.

Findings

- Substantial gap between high and low-resource language performance.
- English instructions yield better results than non-English prompts.
- xLLMs-100 consistently outperforms all baselines across tasks.

Results (Off-Target Analysis)

- **Off-target issue:** Model generates content in wrong language (Zhang et al., 2020).
- **Metric:** Off-Target Ratio (OTR) – percentage of outputs in unintended language.

	FLORES(f)		FLORES(t)	
	Low	High	Low	High
LLaMA	23.26	16.76	14.15	10.16
BX_{LLaMA}	14.13	8.32	12.17	8.24
SFT_{LLaMA}	10.26	6.34	8.72	6.23
xLLMs-100	**8.82**	**3.47**	**6.95**	**1.46**

Table 11: OTR scores (lower is better) of examined multilingual LLMs on the FLORES benchmark.

Findings

- Off-target errors significantly higher in low-resource languages.
- xLLMs-100 reduces error rates compared to base models.
- Remaining gap indicates opportunity for further improvement.

Summary

- xLLMs-100 demonstrates how to address data sparsity in multilingual contexts:
  - **Robustness:** Improves performance across 100+ languages without retraining.
  - **Efficiency:** Achieves multilingual capability with minimal additional parameters.
  - **Generalization:** Cross-lingual feedback enables zero-shot performance on unseen language.
- **Key Findings:**
  - xLLMs-100 shows that **targeted adaptation strategies** can effectively address linguistic data sparsity, achieving more equitable performance across languages without the prohibitive costs of retraining.

Paper Code

ACL 2024 Bangkok, Thailand

Adaptation to Distribution Diversity (Part III: Styles)

- **Research Paper:**
  - **Wen Lai**, Viktor Hangya and Alexander Fraser. Style-Specific Neurons for Steering LLMs in Text Style Transfer. (EMNLP 2024)
- **Research Focuses:**
  - Text style transfer in low-resource style.

Style-Specific Neurons for Steering LLMs in Text Style Transfer

(Presented at EMNLP 2024; Miami, Florida)

Motivation & Research Question

- **Text Style Transfer (TST)**
  - Transform text from **source** to **target** style, preserving **content** and **fluency**.

Style Features	Example	
Formality	Formal: Informal:	Would you like to get a drink? Wanna get a drink?
Politeness	Polite: Impolite:	I see your point, but I have a different perspective. You're wrong.
Sentiment	Positive: Negative:	I'd enjoy something more exciting. This is boring.

Table 13: Examples of classic text style transfer task.

Motivation & Research Question

- **Text Style Transfer (TST)**
  - Transform text from **source** to **target** style, preserving **content** and **fluency**.

Style Features	Example	
Formality	Formal: Informal:	Would you like to get a drink? Wanna get a drink?
Politeness	Polite: Impolite:	I see your point, but I have a different perspective. You're wrong.
Sentiment	Positive: Negative:	I'd enjoy something more exciting. This is boring.

Table 13: Examples of classic text style transfer task.

- **Challenges in LLMs**: tend to **copy** input → weak style control

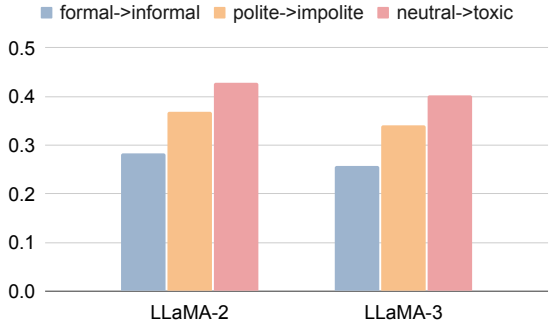


Figure 21: Copy Ratio on TST tasks in original LLMs.

Motivation & Research Question

- **Text Style Transfer (TST)**
  - Transform text from **source** to **target** style, preserving **content** and **fluency**.

Style Features	Example	
Formality	Formal: Informal:	Would you like to get a drink? Wanna get a drink?
Politeness	Polite: Impolite:	I see your point, but I have a different perspective. You're wrong.
Sentiment	Positive: Negative:	I'd enjoy something more exciting. This is boring.

Table 13: Examples of classic text style transfer task.

- **Neuron Analysis**
  - Identify and understand the roles of individual neurons within a neural network.
  - Language-specific neurons can markedly enhance the multilingual capabilities of LLMs during the decoding stage (Tang et al., 2024).

- **Challenges in LLMs**: tend to **copy** input → weak style control

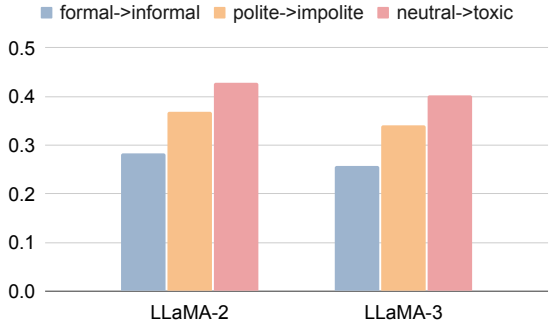


Figure 21: Copy Ratio on TST tasks in original LLMs.

Motivation & Research Question

- **Text Style Transfer (TST)**
  - Transform text from **source** to **target** style, preserving **content** and **fluency**.

Style Features	Example	
Formality	Formal: Informal:	Would you like to get a drink? Wanna get a drink?
Politeness	Polite: Impolite:	I see your point, but I have a different perspective. You're wrong.
Sentiment	Positive: Negative:	I'd enjoy something more exciting. This is boring.

Table 13: Examples of classic text style transfer task.

- **Neuron Analysis**
  - Identify and understand the roles of individual neurons within a neural network.
  - Language-specific neurons can markedly enhance the multilingual capabilities of LLMs during the decoding stage (Tang et al., 2024).

- **Challenges in LLMs**: tend to **copy** input → weak style control

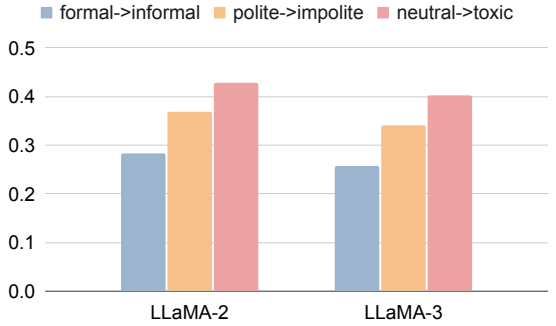


Figure 21: Copy Ratio on TST tasks in original LLMs.

Research Question

- **Q1:** Do LLMs contain neurons specialized for style control?
- **Q2:** If so, how can we leverage them during decoding to better enforce target style?

Our Solution: sNeuron-TST

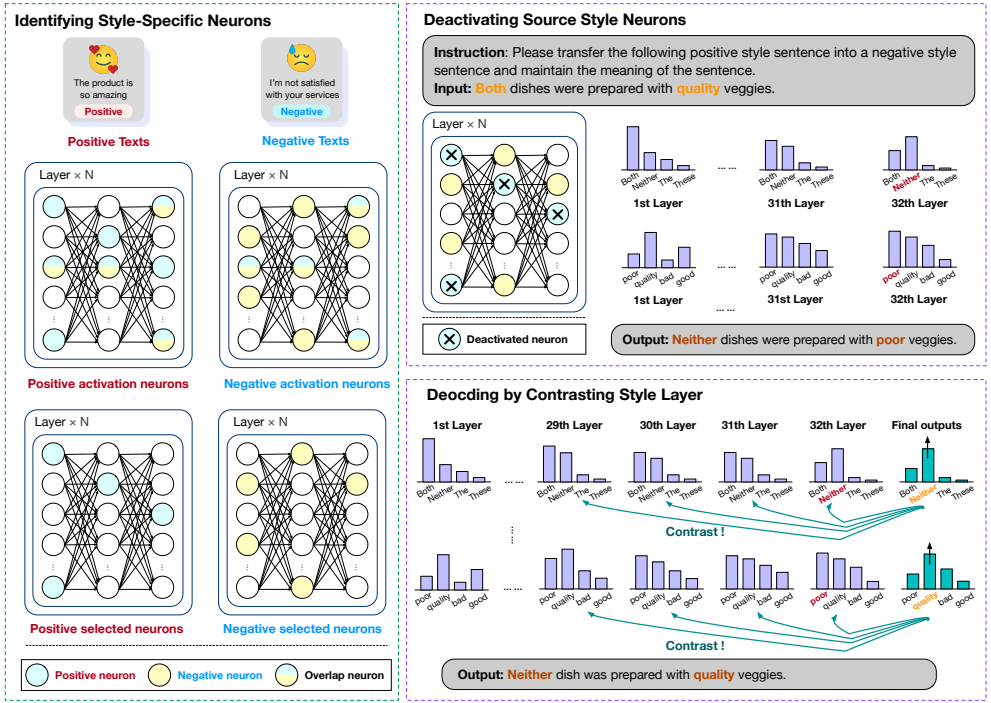


Figure 22: Proposed Framework (sNeuron-TST)

- **Step 1:** Identify Style Neurons
  - Find neurons that correlate with specific styles, remove overlaps between source/target.
- **Step 2:** Deactivate Source Style Neurons
  - Zero out activations of source-style neurons to shift output distribution.
- **Step 3:** Contrastive Decoding
  - Adjust decoding to recover fluency while preserving target style.

Identifying Style-Specific Neurons

- **Neurons in LLMs**
  - Feed-Forward Networks (FFNs) contain about **2/3 of model parameters** and encode key task-specific information (Yang et al., 2023).

$$a^{(j)} = \text{act\_fn}(W^{(j)} a^{(j-1)} + b^{(j)})$$

  - Here, $W^{(j)}$ and $b^{(j)}$ denote weights and biases of layer j; $a^{(j-1)}$ is the previous layer's activation; act_fn$(\cdot)$ is the activation function (e.g., GLU in LLaMA).
  - A neuron i in layer j is considered **active** if $a_i^{(j)} > 0$.

Findings

- Style neuron overlap is very high (95%), harming transfer.
- Remove neurons active for both styles to isolate style-specific features.

- **Neuron Selection**

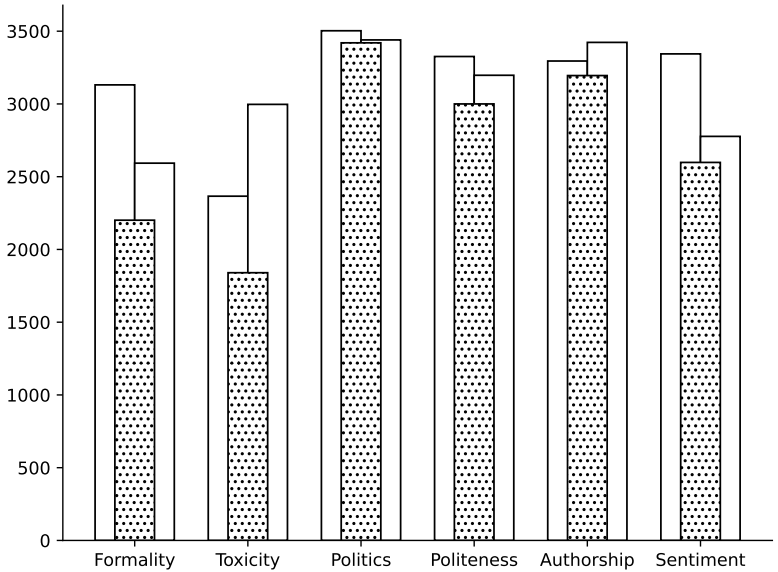


Figure 23: Overlap statistics of style-specific neurons.

Deactivating Source Style Neurons

- **Common Practice:** set selected neuron activations to zero.
- However, neurons are highly sensitive; removing important neurons can degrade quality.
- **Question:** Which neurons should be deactivated — **source** or **target**?

Findings

- Removing source-style neurons: improves target-style accuracy.
- Removing target-style neurons: harms accuracy.
- Fluency drops in all deactivation settings → requires compensation.

Style Accuracy					
Source	Target	Formality		Politeness	
		informal	formal	impolite	polite
55	55	80.00	11.20	79.50	14.80
51	55	**80.53**	**13.63**	**80.06**	**19.37**
55	51	76.25	8.51	65.50	9.27
51	51	78.42	9.27	73.48	10.36
Fluency					
Source	Target	Formality		Politeness	
		informal	formal	impolite	polite
55	55	**92.53**	**87.69**	**105.35**	**92.34**
51	55	104.17	96.83	127.26	105.12
55	51	113.14	106.23	136.10	112.51
51	51	108.22	100.79	131.22	108.64

Table 14: Which side should we deactivate? source or target?

Contrastive Decoding for TST

- **Goal:** improve fluency & factuality by contrasting early vs final layers. Ideas original comes from DoLa (Chuang et al., 2023).
- **Per-layer prediction:**

$$p^j(x_t \mid x_{<t}) = \text{softmax}\big(\phi(h_t^{(j)})\big)$$

- **Contrastive (next-token) prediction:**

$$\hat{p}(x_t \mid x_{<t}) = \text{softmax}\big(\mathscr{F}\big(p^N(x_t), p^M(x_t)\big)\big)$$

Contrastive Decoding for TST

- **Goal:** improve fluency & factuality by contrasting early vs final layers. Ideas original comes from DoLa (Chuang et al., 2023).
- **Per-layer prediction:**

$$p^j(x_t \mid x_{<t}) = \text{softmax}\left(\phi(h_t^{(j)})\right)$$

- **Contrastive (next-token) prediction:**

$$\hat{p}(x_t \mid x_{<t}) = \text{softmax}\left(\mathscr{F}\left(p^N(x_t), p^M(x_t)\right)\right)$$

- **Our adaptation to TST task:**
  - *Candidate layers:* last 4 layers (style concentrated there).
  - *Dynamic selection:* choose $M = \arg\max_{j \in \mathscr{J}} \text{JSD}\left(p^N, p^j\right)$ (largest deviation from final layer).
  - *Interpretation:*
    - small inter-layer changes $\rightarrow$ style-neutral token,
    - large final-layer change $\rightarrow$ style-specific token.

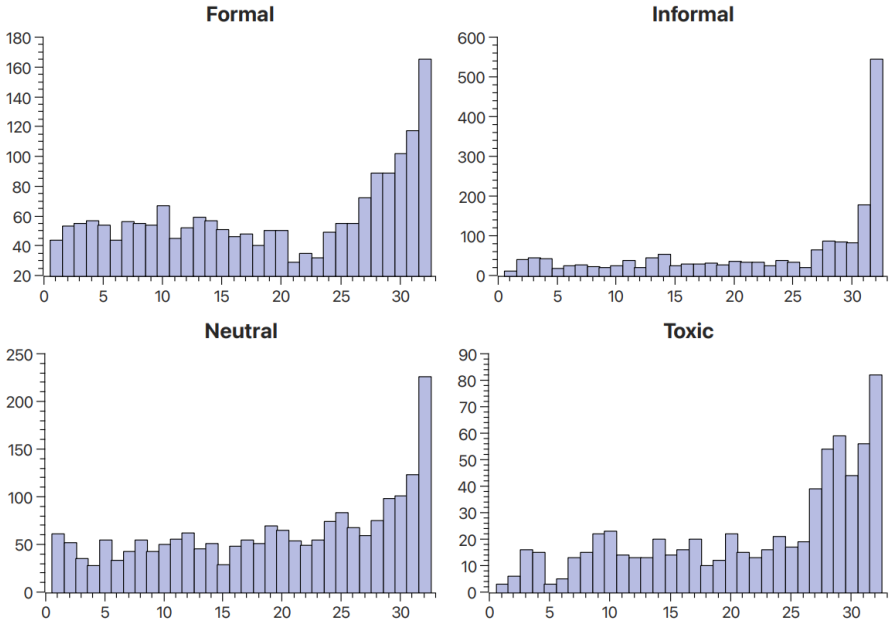


Figure 25: Neuron distribution by style in LLaMA-3 layers.

Experimental Setup

- **Datasets**
  - Six typical TST tasks: Formality (GYAFC; Rao and Tetreault, 2018), Toxicity (ParaDetox; Logacheva et al., 2022), Politics (Political; Voigt et al., 2018), Politeness (Politness; Madaan et al., 2020), Authorship (Shakespeare; Xu et al., 2012), and Sentiment (Yelp; Shen et al., 2017).

Benchmark	Dataset	Tasks	Size		
			train	vald	test
Politeness	Politness (Madaan et al., 2020)	impolite ↔ polite	100k	2000	2000
Toxicity	ParaDetox (Logacheva et al., 2022)	toxic ↔ neutral	18k	2000	2000
Formality	GYAFC (Rao and Tetreault, 2018)	informal ↔ formal	52k	500	500
Authorship	Shakespeare (Xu et al., 2012)	shakespeare ↔ modern	27k	500	500
Politics	Political (Voigt et al., 2018)	democratic ↔ republican	100k	1000	1000
Sentiment	Yelp (Shen et al., 2017)	positive ↔ negative	100k	1000	1000

Table 15: Data statistics on six benchmarks.

Experimental Setup

- **Baselines**
  - **LLaMA-3** (Grattafiori et al., 2024): Vanilla baseline without fine-tuning.
  - **APE** (Tang et al., 2024): Identifies style neurons via activation probability entropy.
  - **AVF** (Tan et al., 2024): Detects style neurons using activation value frequency thresholding.
  - **PNMA** (Kojima et al., 2024): Finds neurons active for source but not target styles.
- **Evaluation Metrics**
  - **Style Accuracy:** Classification accuracy using a trained style classifier.
  - **Content Preservation:** Sentence embedding cosine similarity (LaBSE; Feng et al., 2022).
  - **Fluency:** GPT-2 perplexity (Radford et al., 2019).

Results (Main)

Style Transfer Accuracy

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	80.00	11.20	47.67	29.04	35.50	48.20	79.50	14.80	63.80	43.80	76.40	52.80
APE	74.00	12.20	47.57	28.44	**40.90**	44.80	77.10	18.20	55.80	44.60	78.90	48.00
AVF	76.00	12.40	47.57	28.44	38.80	44.20	77.90	18.70	55.60	44.40	**79.20**	47.90
PNMA	73.85	8.70	42.43	23.79	35.57	37.05	72.84	14.16	53.74	37.58	75.39	41.71
Our	**80.80**	**14.40**	**55.36**	**31.98**	37.81	**50.30**	**80.63**	**23.27**	**73.40**	**45.14**	77.93	**54.73**

Content Preservation

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	**85.95**	74.71	73.54	82.71	82.48	75.77	75.32	**89.14**	78.75	**62.28**	76.17	**74.47**
APE	76.72	85.06	**76.72**	83.00	**87.99**	**82.21**	76.80	87.89	80.07	57.61	**76.52**	73.53
AVF	75.21	84.53	76.63	**83.57**	86.92	80.68	**76.94**	87.32	**80.94**	58.98	76.15	73.95
PNMA	75.52	84.11	75.67	82.54	86.79	80.67	76.04	86.93	79.22	57.42	75.04	72.67
Our	85.84	**86.28**	75.85	80.10	82.32	74.96	75.65	82.47	77.19	60.92	75.25	74.21

Fluency

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	92.53	87.69	113.84	191.30	88.22	68.49	105.35	92.34	197.62	136.03	177.01	125.98
APE	94.27	89.93	133.12	188.34	88.51	69.06	108.24	95.17	250.65	133.92	**151.06**	126.73
AVF	96.63	89.36	131.10	191.29	87.93	75.94	112.67	97.50	220.30	**126.42**	151.33	130.17
PNMA	103.61	90.85	136.27	194.71	96.31	77.95	111.77	101.61	260.52	135.00	154.85	129.49
Our	**90.79**	**81.46**	**85.65**	**172.26**	**85.28**	**66.68**	**104.92**	**83.36**	**151.71**	134.86	174.46	**110.48**

Table 16: Style transfer accuracy (higher values are better; ↑), content preservation (↑) and fluency (↓) on 6 datasets across 12 transfer directions.

Overall Performance

- Baselines (APE, AVF, PNMA) show limited improvement over **LLaMA-3**.
- Style transfer requires fine-grained semantic reasoning beyond token-level matching.
- **Our method** achieves the best balance of style accuracy and fluency.

Results (Main)

Style Transfer Accuracy												
	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	80.00	11.20	47.67	29.04	35.50	48.20	79.50	14.80	63.80	43.80	76.40	52.80
APE	74.00	12.20	47.57	28.44	**40.90**	44.80	77.10	18.20	55.80	44.60	78.90	48.00
AVF	76.00	12.40	47.57	28.44	38.80	44.20	77.90	18.70	55.60	44.40	**79.20**	47.90
PNMA	73.85	8.70	42.43	23.79	35.57	37.05	72.84	14.16	53.74	37.58	75.39	41.71
Our	**80.80**	**14.40**	**55.36**	**31.98**	37.81	**50.30**	**80.63**	**23.27**	**73.40**	**45.14**	77.93	**54.73**
Content Preservation												
	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	**85.95**	74.71	73.54	82.71	82.48	75.77	75.32	**89.14**	78.75	**62.28**	76.17	**74.47**
APE	76.72	85.06	**76.72**	83.00	**87.99**	**82.21**	76.80	87.89	80.07	57.61	**76.52**	73.53
AVF	75.21	84.53	76.63	**83.57**	86.92	80.68	**76.94**	87.32	**80.94**	58.98	76.15	73.95
PNMA	75.52	84.11	75.67	82.54	86.79	80.67	76.04	86.93	79.22	57.42	75.04	72.67
Our	85.84	**86.28**	75.85	80.10	82.32	74.96	75.65	82.47	77.19	60.92	75.25	74.21
Fluency												
	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	92.53	87.69	113.84	191.30	88.22	68.49	105.35	92.34	197.62	136.03	177.01	125.98
APE	94.27	89.93	133.12	188.34	88.51	69.06	108.24	95.17	250.65	133.92	**151.06**	126.73
AVF	96.63	89.36	131.10	191.29	87.93	75.94	112.67	97.50	220.30	**126.42**	151.33	130.17
PNMA	103.61	90.85	136.27	194.71	96.31	77.95	111.77	101.61	260.52	135.00	154.85	129.49
Our	**90.79**	**81.46**	**85.65**	**172.26**	**85.28**	**66.68**	**104.92**	**83.36**	**151.71**	134.86	174.46	**110.48**

Table 16: Style transfer accuracy (higher values are better; ↑), content preservation (↑) and fluency (↓) on 6 datasets across 12 transfer directions.

Content Preservation

- High scores mainly reflect **copying behavior**, not real style control.
- Models often preserve meaning but lack stylistic transformation.
- Cosine similarity underestimates large semantic gaps (e.g., formal $\rightleftharpoons$ informal).

Results (Main)

Style Transfer Accuracy	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	80.00	11.20	47.67	29.04	35.50	48.20	79.50	14.80	63.80	43.80	76.40	52.80
APE	74.00	12.20	47.57	28.44	**40.90**	44.80	77.10	18.20	55.80	44.60	78.90	48.00
AVF	76.00	12.40	47.57	28.44	38.80	44.20	77.90	18.70	55.60	44.40	**79.20**	47.90
PNMA	73.85	8.70	42.43	23.79	35.57	37.05	72.84	14.16	53.74	37.58	75.39	41.71
Our	**80.80**	**14.40**	**55.36**	**31.98**	37.81	**50.30**	**80.63**	**23.27**	**73.40**	**45.14**	77.93	**54.73**

Content Preservation	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	**85.95**	74.71	73.54	82.71	82.48	75.77	75.32	**89.14**	78.75	**62.28**	76.17	**74.47**
APE	76.72	85.06	**76.72**	83.00	**87.99**	**82.21**	76.80	87.89	80.07	57.61	**76.52**	73.53
AVF	75.21	84.53	76.63	**83.57**	86.92	80.68	**76.94**	87.32	**80.94**	58.98	76.15	73.95
PNMA	75.52	84.11	75.67	82.54	86.79	80.67	76.04	86.93	79.22	57.42	75.04	72.67
Our	85.84	**86.28**	75.85	80.10	82.32	74.96	75.65	82.47	77.19	60.92	75.25	74.21

Fluency	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	92.53	87.69	113.84	191.30	88.22	68.49	105.35	92.34	197.62	136.03	177.01	125.98
APE	94.27	89.93	133.12	188.34	88.51	69.06	108.24	95.17	250.65	133.92	**151.06**	126.73
AVF	96.63	89.36	131.10	191.29	87.93	75.94	112.67	97.50	220.30	**126.42**	151.33	130.17
PNMA	103.61	90.85	136.27	194.71	96.31	77.95	111.77	101.61	260.52	135.00	154.85	129.49
Our	**90.79**	**81.46**	**85.65**	**172.26**	**85.28**	**66.68**	**104.92**	**83.36**	**151.71**	134.86	174.46	**110.48**

Table 16: Style transfer accuracy (higher values are better; ↑), content preservation (↑) and fluency (↓) on 6 datasets across 12 transfer directions.

Different Directions

- Strong asymmetry: impolite →polite 80%, polite → impolite 23%.
- Bias from polite/positive data and safety-driven generation reduces stylistic diversity.

Results (Copy Problem)

- The **copy problem** occurs when models reproduce the input verbatim — a common issue in multilingual MT (Lai et al., 2023).
- In TST, maintaining semantic consistency often leads LLMs to **over-copy** input sentences.
- We analyze this behavior in **formality**, **politeness**, and **toxicity** transformation tasks.

Findings

- Neuron-based methods (APE, AVF, PNMA) partially reduce copying via neuron control.
- Our method further lowers copy rate by deactivating source-side neurons and using a novel decoding strategy.

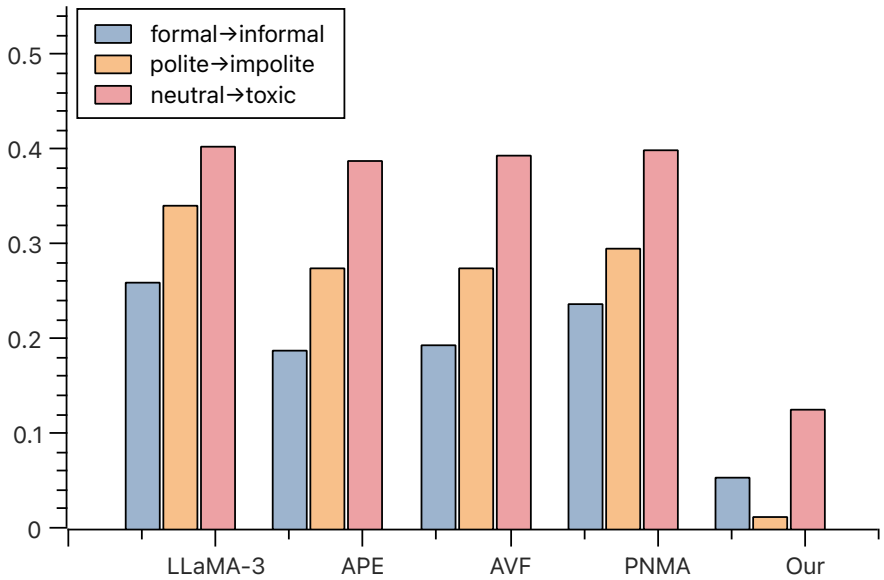


Figure 26: Copy Ratio on three selected TST tasks.

Summary

- **sNeuron-TST** tackles data sparsity in style adaptation:
  - **Robustness:** Neuron-level interventions reduce over-copying and improve target-style accuracy across six TST tasks.
  - **Efficiency:** No fine-tuning or extra data; adaptation is applied fully at inference time.
  - **Generalization:** Works across diverse styles and offers transferable neuron insights for broader ADaS problems.
- **Impact:** A lightweight and interpretable approach for reliable style control under scarce data.

Paper Code

EMNLP 2024

Adaptation to Data Scarcity (Part I: Task-Agnostic)

- **Research Papers:**
  - **Wen Lai**, Alexander Fraser and Ivan Titov. Joint Localization and Activation Editing for Low-Resource Fine-Tuning. (ICML 2025)
- **Research Focuses:**
  - Adaptation to low-resource settings, beyond the different distribution (e.g., domains, languages and styles).

Joint Localization and Activation Editing for Low-Resource Fine-Tuning

(Presented at ICML 2025; Vancouver, Canada)

Backgrounds & Motivations

(I) Parameter-Efficient Fine-Tuning (PEFT)

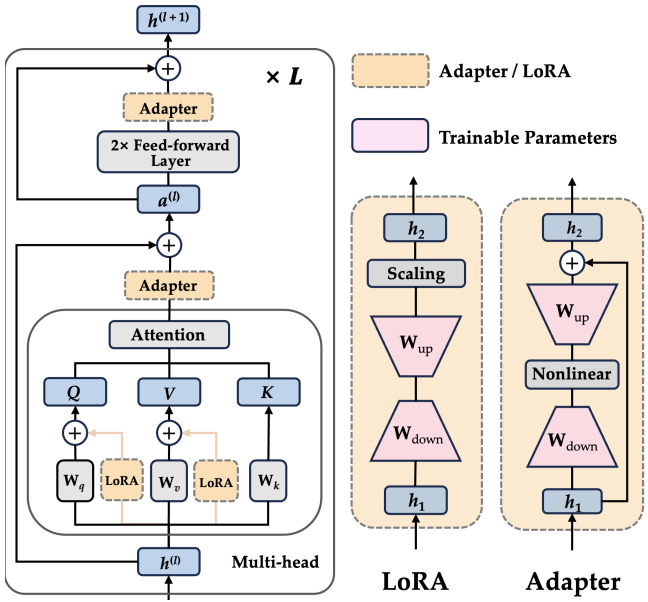


Figure 27: Architecture of classic PEFT methods.

- **Adapters** (Houlsby et al., 2019): Learnable modules inserted after sub-layers.
- **LoRA** (Hu et al., 2022): Adds low-rank matrices in parallel to W_q and W_v in attention layers.

Takeaway

- PEFT avoids modifying the original weights, which makes it deployment-friendly. However, methods like LoRA still introduce additional parameters (e.g., 0.826% for LLaMA-3-8B).
- The effectiveness of standard PEFT is limited in low-resource scenarios with only a few hundred examples.

Backgrounds & Motivations

(II) Activation Editing: A New Paradigm

- Activation editing enables parameter-efficient fine-tuning by learning small interventions on internal activations, achieving strong results with minimal updates (e.g., 0.0035% for LoFIT), even in low-data regimes.

- **Intervention component**
  - **BitFIT** (Ben Zaken et al., 2022): Bias term.
  - **RED** (Wu et al., 2024a): MLP layers output.
  - **ReFT** (Wu et al., 2024b): Hidden outputs (representation) within MLP layers.
  - **LoFIT** (Yin et al., 2024): Attention head outputs.

- **Intervention Strategy**
  - Given an activation output $z_t^{(l,i)} \in \mathbb{R}^{d_l}$ for i-th component at layer l, we apply the transformation: $z_t^{(l,i)'} = f(z_t^{(l,i)})$

Three Strategies

- **Additive:** $z_t^{(l,i)'} = z_t^{(l,i)} + a^{(l,i)}$
- **Multiplicative:** $z_t^{(l,i)'} = m^{(l,i)} \odot z_t^{(l,i)}$
- **Hybrid:** $z_t^{(l,i)'} = m^{(l,i)} \odot z_t^{(l,i)} + a^{(l,i)}$

Backgrounds & Motivations

(III) What Remains Unclear for Activation Editing?

- Which internal component yields the best intervention outcome?
  - Attention head outputs are the most effective intervention targets.
- Should we use additive, multiplicative, or hybrid operations for optimal results?
  - Additive bias offsets consistently lead to greater performance improvements than multiplicative scaling.
- Can activation editing perform well in **low-resource settings** (e.g., 200 samples)?
  - Performance is highly sensitive to hyperparameter choices, requiring careful manual tuning for each task.

Our Solution: JoLA

(I) JoLA Framework: An overview

- **Our goal:** Design a simple and general approach to **dynamically learn where and how** to edit activations in low-resource settings.

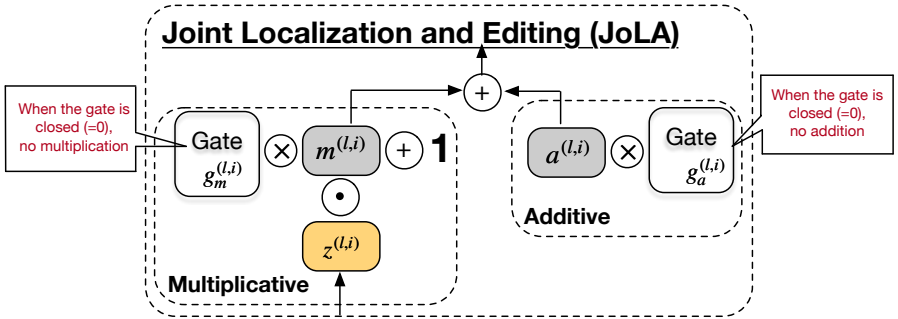


Figure 28: Architecture of gate design.

- $z^{(l,i)}$ – original (head / MLP) activation
- $a^{(l,i)}$ – additive modification
- $m^{(l,i)}$ – component-wise multiplicative modification

Takeaway

- Each gate is just a **random variable during training** (no input!) and becomes a scalar expectation at inference time.

Our Solution: JoLA

(II) Gating Mechanism: Learn Sparse Edits

- Gates follow a **mixed discrete-continuous distribution**, implemented via the **Hard Concrete** distribution (Louizos et al., 2018).
- The probability that a gate is non-zero acts as a L_0 **regularizer**, encouraging sparsity by controlling the expected number of active (open) gates.

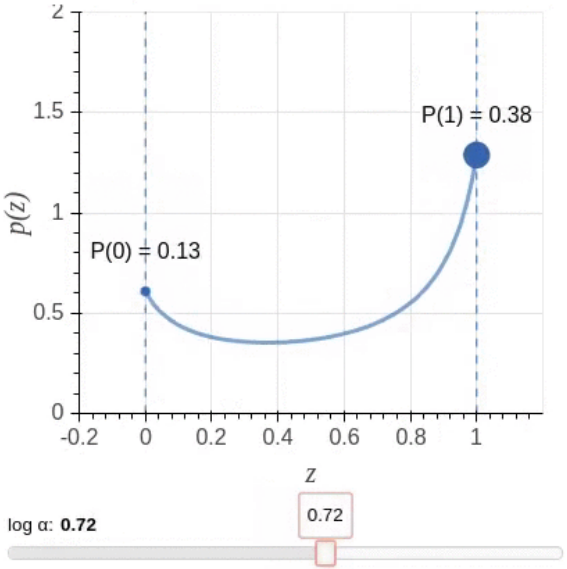


Figure 29: Hard Concrete Distribution.

Our Solution: JoLA

(III) Activation Status

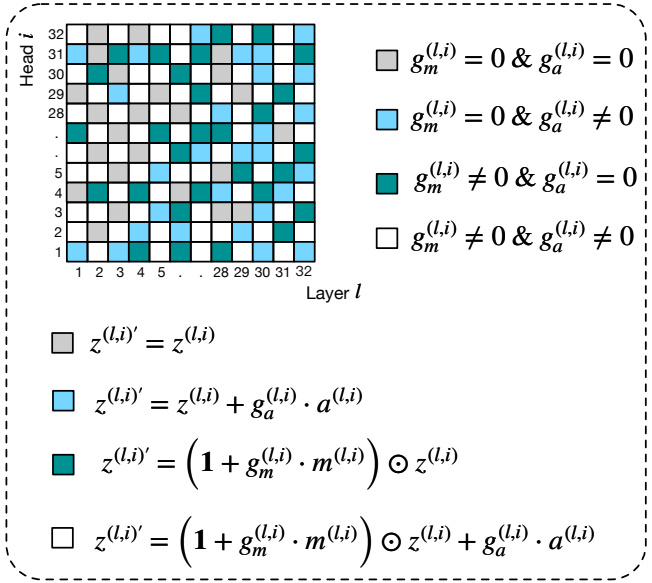


Figure 30: Four activation status during training.

- Given an activation output $z_t^{(l,i)} \in \mathbb{R}^{d_l}$ for i-th head at layer l, the activation can be optimized to four status during training.

Four Status

1. original activation (no modification)
2. add a bias vector (additive modification)
3. add a scale vector (multiplicative modification)
4. both scale and bias vector (hybrid modification)

Our Solution: JoLA

(IV) Training Objectives

$$L(\mathbf{m}, \mathbf{a}, \phi) = L_{xent}(\mathbf{m}, \mathbf{a}) + \lambda L_C(\phi)$$

where,

- $L_{xent}(\cdot)$: Standard cross-entropy loss
- $L_C(\phi)$: L_0 regularizer defined as:

$$L_C(\phi) = \sum_{l,i} \left(1 - P(g_a^{(l,i)} = 0 \mid \phi_a^{(l,i)}) + 1 - P(g_m^{(l,i)} = 0 \mid \phi_m^{(l,i)})\right)$$

Takeaway

- $L_C(\phi)$ regularizes the number of open gates, encouraging the model to close gates as training progresses.
- Most gates are closed at convergence, i.e., only a few interventions are applied.

Our Solution: JoLA

(V) Gate Status During Training

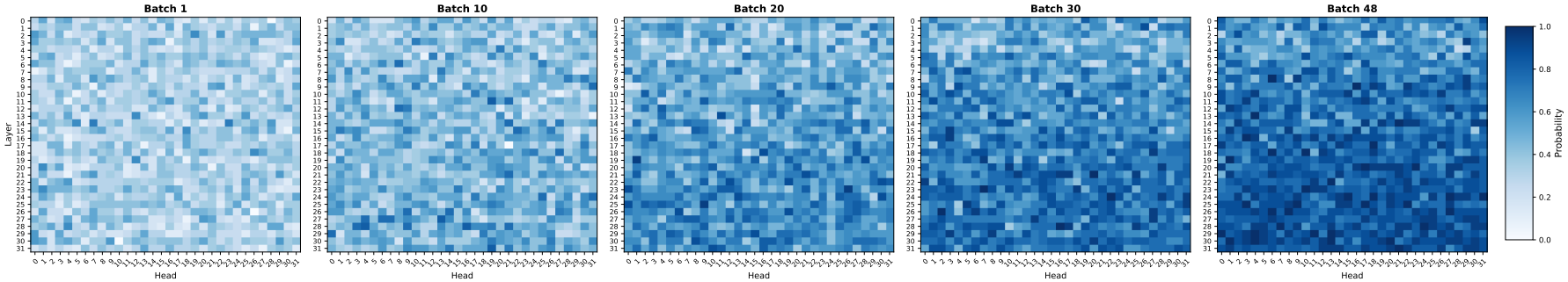


Figure 31: Gate status during training. Initially (open) and Finally (closed).

Takeaway

- All gates are initially open. Then, JoLA learns **which components** (e.g., attention heads) to modify and **how** (additively or multiplicatively)
- Interestingly, multiplicative gate (g_m) tends to **close more frequently**

Experimental Setup

- **Datasets (26 Benchmarks):**
  - **Commonsense Reasoning (8 datasets)**: ARC-c and ARC-e (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OBQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), and WinoGrande (Sakaguchi et al., 2021).
  - **Natural Language Understanding (14 datasets):** MMLU-Pro benchmark (Wang et al., 2024).
  - **Natural Language Generation (4 datasets in GEM benchmark Gehrmann et al., 2022):** CommonGen (Lin et al., 2020), E2E (Novikova et al., 2017), WebNLG (Gardent et al., 2017) and XSum (Narayan et al., 2018).

Experimental Setup

- **Datasets (26 Benchmarks):**
  - **Commonsense Reasoning (8 datasets)**: ARC-c and ARC-e (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OBQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), and WinoGrande (Sakaguchi et al., 2021).
  - **Natural Language Understanding (14 datasets):** MMLU-Pro benchmark (Wang et al., 2024).
  - **Natural Language Generation (4 datasets in GEM benchmark Gehrmann et al., 2022):** CommonGen (Lin et al., 2020), E2E (Novikova et al., 2017), WebNLG (Gardent et al., 2017) and XSum (Narayan et al., 2018).
- **Baselines:**
  - **Zero-shot**:LLaMA-3 (Grattafiori et al., 2024) and Qwen-2.5 (Yang et al., 2025).
  - **PEFT method**: LoRA (Hu et al., 2022).
  - **Activation editing during training**: BitFit (Ben Zaken et al., 2022), RED (Wu et al., 2024a), ReFT (Wu et al., 2024b), and LoFIT (Yin et al., 2024).
  - **Activation editing during inference**: RePE (Zou et al., 2023).
- **Evaluation Metrics:**
  - Accuracy for reasoning and understanding tasks, BLEU, ROUGE-L and BERTScore (Zhang et al., 2019) for generation tasks

Experimental Setup

- **Datasets (26 Benchmarks):**
  - **Commonsense Reasoning (8 datasets)**: ARC-c and ARC-e (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), OBQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), and WinoGrande (Sakaguchi et al., 2021).
  - **Natural Language Understanding (14 datasets):** MMLU-Pro benchmark (Wang et al., 2024).
  - **Natural Language Generation (4 datasets in GEM benchmark Gehrmann et al., 2022):** CommonGen (Lin et al., 2020), E2E (Novikova et al., 2017), WebNLG (Gardent et al., 2017) and XSum (Narayan et al., 2018).
- **Baselines:**
  - **Zero-shot**:LLaMA-3 (Grattafiori et al., 2024) and Qwen-2.5 (Yang et al., 2025).
  - **PEFT method**: LoRA (Hu et al., 2022).
  - **Activation editing during training**: BitFit (Ben Zaken et al., 2022), RED (Wu et al., 2024a), ReFT (Wu et al., 2024b), and LoFIT (Yin et al., 2024).
  - **Activation editing during inference**: RePE (Zou et al., 2023).
- **Evaluation Metrics:**
  - Accuracy for reasoning and understanding tasks, BLEU, ROUGE-L and BERTScore (Zhang et al., 2019) for generation tasks

Takeaway

- For all datasets, we sample **200 examples** to simulate low-resource scenarios.

Results (Main)

Llama-3.1-8B-Instruct					
	Reasoning	Understanding	Generation		
	ACC ↑	ACC ↑	BLEU ↑	ROUGE-L ↑	BERTScore ↑
zero_shot	53.70	40.00	12.56	36.70	77.23
LoRA	66.58	42.07	13.27	36.97	77.74
BitFit	63.05	35.02	9.25	28.81	74.83
RED	46.19	37.33	11.24	32.40	76.24
RePE	63.61	35.54	8.49	27.61	74.30
ReFT	65.95	40.89	12.60	36.89	77.21
LoFIT	56.19	27.76	11.88	32.09	76.71
JoLA	**70.55**	**47.00**	**17.07**	**40.65**	**80.54**

Table 17: Average performance comparison of JoLA and baselines.

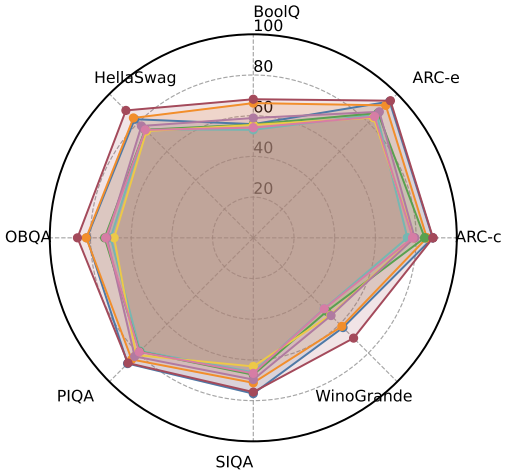


Figure 32: Performance comparison.

Takeaway

- Baseline methods: variable performance, sensitive to hyperparameters.
- **JoLA: Consistently superior across all tasks** with minimal tuning required.

Analysis: Ablation on Gating Mechanism

	Reasoning		Understanding		Generation	
	SIQA	WinoGrande	Law	Physics	E2E_NLG	WEB_NLG
MLP w/o gate	50.10	51.62	34.00	20.00	10.31	14.45
MLP with gate	**52.46**	**52.43**	**36.00**	**23.00**	**11.23**	**16.25**
Attention w/o gate	55.94	55.33	36.00	7.00	14.77	18.12
Attention with gate	**66.22**	**58.33**	**40.00**	**46.00**	**15.54**	**24.39**
Attention + MLP w/o gate	52.17	48.74	23.00	13.00	8.23	12.36
Attention + MLP with gate	**53.28**	**52.07**	**27.00**	**16.00**	**10.42**	**14.83**

Table 18: MLP and Attention interventions with/without gating.

Takeaway

- Gating mechanism significantly improves performance for both attention and MLP interventions.

Analysis: Ablation on Gating Mechanism

	Reasoning		Understanding		Generation	
	SIQA	WinoGrande	Law	Physics	E2E_NLG	WEB_NLG
MLP w/o gate	50.10	51.62	34.00	20.00	10.31	14.45
MLP with gate	**52.46**	**52.43**	**36.00**	**23.00**	**11.23**	**16.25**
Attention w/o gate	55.94	55.33	36.00	7.00	14.77	18.12
Attention with gate	**66.22**	**58.33**	**40.00**	**46.00**	**15.54**	**24.39**
Attention + MLP w/o gate	52.17	48.74	23.00	13.00	8.23	12.36
Attention + MLP with gate	**53.28**	**52.07**	**27.00**	**16.00**	**10.42**	**14.83**

Table 18: MLP and Attention interventions with/without gating.

Takeaway

- Combined attention+MLP with gating shows improvement, but single interventions still perform better, confirming previous findings.

Analysis: Different Data Size

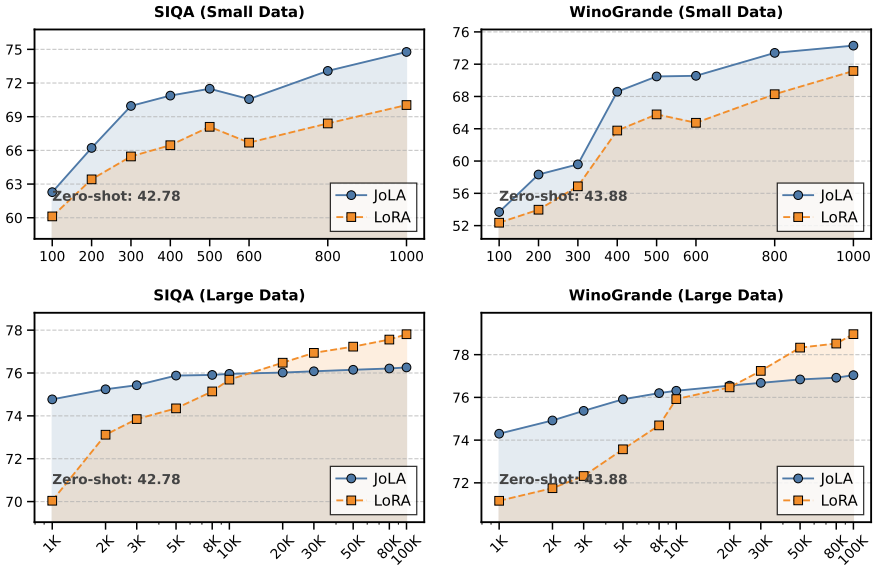


Figure 35: Performance across data sizes.

Data Size:

- **small data size:** 100 - 1,000 samples;
- **large data size:** 1,000 - 100,0000 samples;

Takeaway

- JoLA excels with small datasets (100-1,000 samples).
- Comparable performance at medium scale (5,000-10,000).
- LoRA slightly better at large scale (20k-100k).

Analysis: Different Model Size

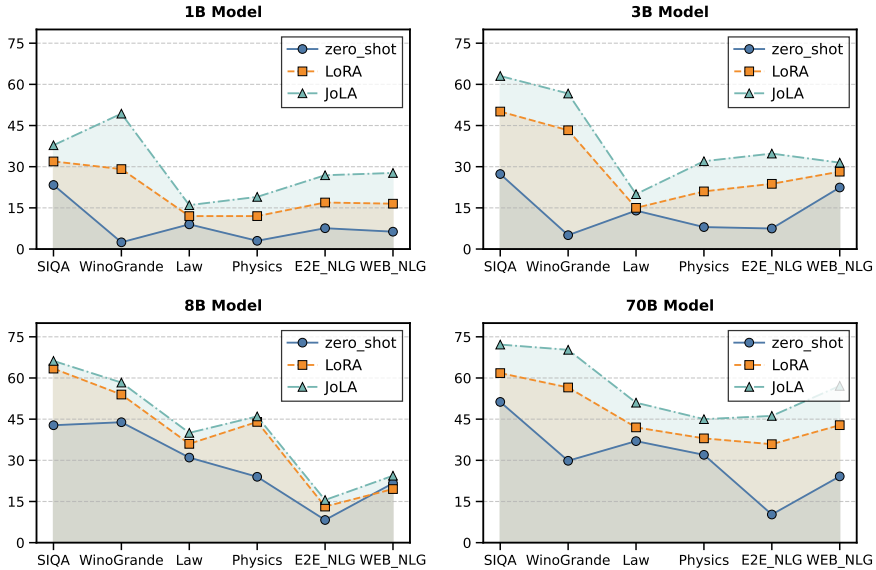


Figure 36: Performance across model scales.

Model Size:

- LLaMA (1B, 3B, 8B, 70B)

Takeaway

- JOLA improves performance across all model sizes.
- Larger models show more substantial benefits.

Summary

- **JoLA** addresses the three ADaS goals:
  - **Robustness:** Stable low-data adaptation (100–1k samples) via learnable activation interventions.
  - **Efficiency:** Sparse Hard-Concrete gating $\rightarrow$ few active edits, tiny parameter cost, low overhead.
  - **Generalization:** A unified additive/multiplicative update rule that transfers across tasks and domains.
- **Impact:** A simple and deployable recipe for reliable LLM adaptation under distributional sparsity.

Paper Code Blog

ICML International Conference On Machine Learning

Adaptation to Data Scarcity (Part II: Dataset Construction)

- **Research Papers:**
  - Yingli Shen*, **Wen Lai***, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser and Maosong Sun. DCAD-2000: A Multilingual Dataset across 2000+ Languages with Data Cleaning as Anomaly Detection. (NeurIPS 2025)
  - Yingli Shen*, **Wen Lai***, Shuo Wang, Ge Gao, Kangyang Luo, Alexander Fraser and Maosong Sun. From Unaligned to Aligned: Scaling Multilingual LLMs with Multi-Way Parallel Corpora. (EMNLP 2025)
- **Research Focuses:**
  - Multilingual dataset construction for LLMs continue pretraining (NeurIPS 2025).
  - Multi-Way Parallel dataset construction for scaling multilingual capability (EMNLP 2025).

Research Focus

DCAD-2000 (Multilingual Dataset)

- **Massive, language-agnostic dataset:** 2,282 languages, 8.63B docs (46.7TB) for continued pre-training.
- **Key idea:** Reframe data cleaning as *anomaly detection* to avoid fragile, language-specific rules.
- **Result / impact:** Cleaner multilingual corpora → better LLM performance across diverse downstream tasks.

Paper Code Data

Research Focus

DCAD-2000 (Multilingual Dataset)

- **Massive, language-agnostic dataset:** 2,282 languages, 8.63B docs (46.7TB) for continued pre-training.
- **Key idea:** Reframe data cleaning as *anomaly detection* to avoid fragile, language-specific rules.
- **Result / impact:** Cleaner multilingual corpora → better LLM performance across diverse downstream tasks.

Paper Code Data

TED2025 (Multi-Way Parallel Corpora)

- **Resource:** High-quality multi-way parallel corpus covering **113 languages** (up to 50-way).
- **Goal / motivation:** Improve multilingual alignment in LLMs where parallel signal is scarce.
- **Findings / best practices:** Multi-way data substantially aids multilingual capabilities; provides guidelines for effective instruction-tuning with parallel data.

Paper Code & Data

Contributions & Future Work

Contributions

- We investigate adaptation to data sparsity (ADaS), a challenging problem in Machine Translation and Large Language Models.

Contributions

- We investigate adaptation to data sparsity (ADaS), a challenging problem in Machine Translation and Large Language Models.
- We published several works addressing ADaS across multiple dimensions:
  - **Domains:** RMLNMT and *m*^{4}*Adapter* for robust domain adaptation.
  - **Languages:** Bi-ACL and xLLMs-100 for low-resource languages.
  - **Styles:** sNeuron-TST for style-specific control.
  - **Task-Agnostic:** JoLA for efficient adaptation with minimal data.
  - **Dataset Construction:** DCAD-2000 and TED2025 for scaling multilingual capabilities of LLMs.

Contributions

- We investigate adaptation to data sparsity (ADaS), a challenging problem in Machine Translation and Large Language Models.
- We published several works addressing ADaS across multiple dimensions:
  - **Domains:** RMLNMT and *m*^{4}*Adapter* for robust domain adaptation.
  - **Languages:** Bi-ACL and xLLMs-100 for low-resource languages.
  - **Styles:** sNeuron-TST for style-specific control.
  - **Task-Agnostic:** JoLA for efficient adaptation with minimal data.
  - **Dataset Construction:** DCAD-2000 and TED2025 for scaling multilingual capabilities of LLMs.

Key advancements

- Improved **robustness** to distributional shift.
- Enhanced **efficiency** through parameter-sparse methods.
- Better **generalization** across unseen distributions.
- Novel dataset construction for **extremely low-resource** settings.

List of Publications

- **Included in the dissertation:**
  1. **Wen Lai**, Alexander Fraser, and Ivan Titov. Joint Localization and Activation Editing for Low-Resource Fine-Tuning. (ICML 2025)
  2. **Wen Lai**, Viktor Hangya, and Alexander Fraser. 2024. Style-Specific Neurons for Steering LLMs in Text Style Transfer. (EMNLP 2024)
  3. **Wen Lai**, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback. (ACL 2024)
  4. **Wen Lai**, Alexandra Chronopoulou, and Alexander Fraser. 2023. Mitigating Data Imbalance and Representation Degeneration in Multilingual Machine Translation. (EMNLP 2023)
  5. **Wen Lai**, Alexandra Chronopoulou, and Alexander Fraser. 2022. m4 Adapter: Multilingual Multi-Domain Adaptation for Machine Translation with a Meta-Adapter. (EMNLP 2022)
  6. **Wen Lai**, Jindřich Libovický, and Alexander Fraser. 2022. Improving Both Domain Robustness and Domain Adaptability in Machine Translation. (COLING 2022)
- **Not included:**
  1. Yingli Shen*, **Wen Lai***, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. DCAD-2000: A Multilingual Dataset across 2000+ Languages with Data Cleaning as Anomaly Detection. (NeurIPS 2025) (* Equal Contribution)
  2. Yingli Shen*, **Wen Lai***, Shuo Wang, Kangyang Luo, Alexander Fraser, and Maosong Sun. From Unaligned to Aligned: Scaling Multilingual LLMs with Multi-Way Parallel Corpora. (EMNLP 2025). (* Equal Contribution) Nominated for Outstanding Paper Award and Resource Award.
  3. **Wen Lai**, Viktor Hangya, and Alexander Fraser. Extending multilingual machine translation through imitation learning. *arXiv preprint arXiv:2311.08538 (2023)*.
  4. **Wen Lai**, Jindřich Libovický, and Alexander Fraser. 2021. The LMU Munich System for the WMT 2021 Large-Scale Multilingual Machine Translation Shared Task.(WMT 2021)

Future Work

- **Unified Low-Resource Adaptation**
  - Developing frameworks that handle all dimensions of data sparsity simultaneously.

Future Work

- **Unified Low-Resource Adaptation**
  - Developing frameworks that handle all dimensions of data sparsity simultaneously.
- **Cross-Modal and Multimodal Adaptation**
  - Extending adaptation techniques to incorporate visual, audio, and other modalities.

Future Work

- **Unified Low-Resource Adaptation**
  - Developing frameworks that handle all dimensions of data sparsity simultaneously.
- **Cross-Modal and Multimodal Adaptation**
  - Extending adaptation techniques to incorporate visual, audio, and other modalities.
- **Continual and Lifelong Adaptation**
  - Building systems that can continuously adapt to new distributions without forgetting.

Future Work

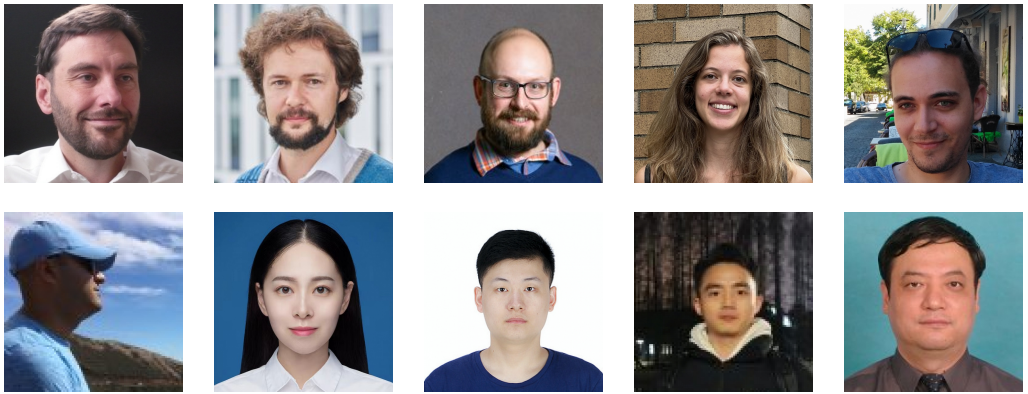
- **Unified Low-Resource Adaptation**
  - Developing frameworks that handle all dimensions of data sparsity simultaneously.
- **Cross-Modal and Multimodal Adaptation**
  - Extending adaptation techniques to incorporate visual, audio, and other modalities.
- **Continual and Lifelong Adaptation**
  - Building systems that can continuously adapt to new distributions without forgetting.
- **Human-in-the-Loop and Interactive Adaptation**
  - Incorporating human feedback more efficiently in adaptation processes.

Future Work

- **Unified Low-Resource Adaptation**
  - Developing frameworks that handle all dimensions of data sparsity simultaneously.
- **Cross-Modal and Multimodal Adaptation**
  - Extending adaptation techniques to incorporate visual, audio, and other modalities.
- **Continual and Lifelong Adaptation**
  - Building systems that can continuously adapt to new distributions without forgetting.
- **Human-in-the-Loop and Interactive Adaptation**
  - Incorporating human feedback more efficiently in adaptation processes.
- **Social and Ethical Implications**
  - Addressing the broader impacts of adaptation, particularly regarding equity across languages and cultures.

Thank You! & QA

Co-authors:



References I

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. (2022). BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Bisk, Y., Zellers, R., Gao, J., Choi, Y., et al. (2020). Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. (2023). Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (2019). BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cooper Stickland, A., Berard, A., and Nikoulina, V. (2021a). Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Cooper Stickland, A., Li, X., and Ghazvininejad, M. (2021b). Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., et al. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

References II

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). The WebNLG challenge: Generating text from RDF data. In Alonso, J. M., Bugarín, A., and Reiter, E., editors, *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Gehrmann, S., Bhattacharjee, A., Mahendiran, A., Wang, A., Papangelis, A., Madaan, A., Mcmillan-major, A., Shvets, A., Upadhyay, A., Bohnet, B., Yao, B., Wilie, B., Bhagavatula, C., You, C., Thomson, C., Garbacea, C., Wang, D., Deutsch, D., Xiong, D., Jin, D., Gkatzia, D., Radev, D., Clark, E., Durmus, E., Ladhak, F., Ginter, F., Winata, G. I., Strobelt, H., Hayashi, H., Novikova, J., Kanerva, J., Chim, J., Zhou, J., Clive, J., Maynez, J., Sedoc, J., Juraska, J., Dhole, K., Chandu, K. R., Beltrachini, L. P., Ribeiro, L. F. . R., Tunstall, L., Zhang, L., Pushkarna, M., Creutz, M., White, M., Kale, M. S., Eddine, M. K., Daheim, N., Subramani, N., Dusek, O., Liang, P. P., Ammanamanchi, P. S., Zhu, Q., Puduppully, R., Kriz, R., Shahriyar, R., Cardenas, R., Mahamood, S., Osei, S., Cahyawijaya, S., Štajner, S., Montella, S., Jolly, S., Mille, S., Hasan, T., Shen, T., Adewumi, T., Raunak, V., Raheja, V., Nikolaev, V., Tsai, V., Jernite, Y., Xu, Y., Sang, Y., Liu, Y., and Hou, Y. (2022). GEMv2: Multilingual NLG benchmarking in a single line of code. In Che, W. and Shutova, E., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (2021). XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

References III

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Kojima, T., Okimura, I., Iwasawa, Y., Yanaka, H., and Matsuo, Y. (2024). On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Lai, W., Hangya, V., and Fraser, A. (2023). Extending multilingual machine translation through imitation learning. *arXiv preprint arXiv:2311.08538*.

Lai, W., Libovický, J., and Fraser, A. (2022). Improving both domain robustness and domain adaptability in machine translation. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5191–5204, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Li, H., Koto, F., Wu, M., Aji, A. F., and Baldwin, T. (2023). Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y., and Ren, X. (2020). CommonGen: A constrained text generation challenge for generative commonsense reasoning. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., Semenov, N., and Panchenko, A. (2022). ParaDetox: Detoxification with parallel data. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*.

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., and Prabhumoye, S. (2020). Politeness transfer: A tag and generate approach. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

References IV

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Novikova, J., Dušek, O., and Rieser, V. (2017). The E2E dataset: New challenges for end-to-end generation. In Jokinen, K., Stede, M., DeVault, D., and Louis, A., editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., and Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2021). Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. (2019). Social IQa: Commonsense reasoning about social interactions. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

References V

Sharaf, A., Hassan, H., and Daumé III, H. (2020). Meta-learning for few-shot NMT adaptation. In Birch, A., Finch, A., Hayashi, H., Heafield, K., Junczys-Dowmunt, M., Konstas, I., Li, X., Neubig, G., and Oda, Y., editors, *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 43–53, Online. Association for Computational Linguistics.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Tan, S., Wu, D., and Monz, C. (2024). Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., and Wen, J.-R. (2024). Language-specific neurons: The key to multilingual capabilities in large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., and Tsvetkov, Y. (2018). RtGender: A corpus for studying differential responses to gender. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023). Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

References VI

Wu, M., Liu, W., Wang, X., Li, T., Lv, C., Ling, Z., JianHao, Z., Zhang, C., Zheng, X., and Huang, X. (2024a). Advancing parameter efficiency in fine-tuning via representation editing. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13445–13464, Bangkok, Thailand. Association for Computational Linguistics.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. (2024b). Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962.

Xu, W., Ritter, A., Dolan, B., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In Kay, M. and Boitet, C., editors, *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, S., Yu, X., Tian, Y., Yan, X., Ma, H., and Zhang, X. (2023). Evolutionary neural architecture search for transformer in knowledge tracing. *Advances in Neural Information Processing Systems*, 36:19520–19539.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yin, F., Ye, X., and Durrett, G. (2024). Lofit: Localized fine-tuning on llm representations. *Advances in Neural Information Processing Systems*, 37:9474–9506.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.