Mechanistic Interpretability: Self-conditioning

Zhaokun Wang(HS)
Master students of Computational Linguistics

Contents: 2 papers

1.Self-conditioning Pre-Trained Language Models

2.On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons

1.Figures and tables keep the same number in each paper

Contents of Paper 1

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

Introduction

INTRODUCTION

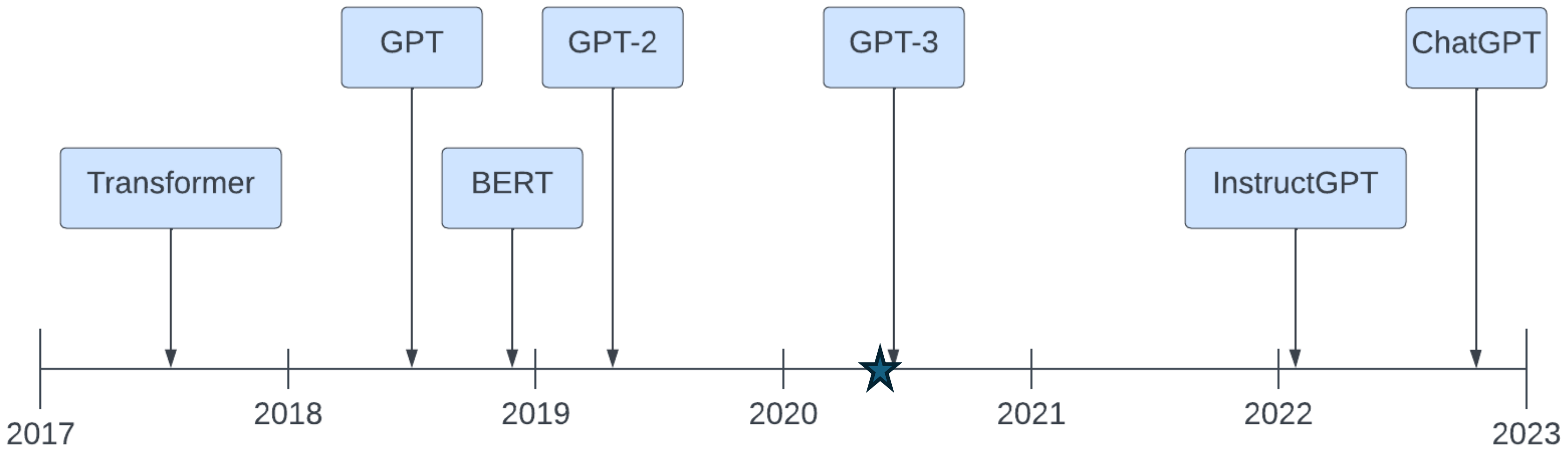
METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

Background

- **Transformers** introduced in 2017 revolutionized NLP.
- **BERT** and **GPT-2** became popular around **2019**.
- Earlier work explored **Finding Experts in Transformer Models**.



Inspiration from Previous Work

Inspiration from Images:
Neurons capture visual concepts like **“trees”** or **“dogs”** (Bau et al., 2017).

In Text:
Sentiment neurons in LSTMs detect emotions like **happiness** or **sadness** (Radford et al., 2017).

TLMs' Pros and Cons

Pros

- Mastery of diverse tasks (text generation, summarization).

Cons

- Mechanism unknown
- Lack control over output.
- Biases inherited from training data.

Objective

FIND MECHANISM

CONDITIONED TEXT GENERATION

Method

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

Concepts as binary sentence datasets

Structure:
Concept c: Described by sentences labeled as:

- **Contains c: Positive examples.**
- **Does not contain c: Negative examples.**

Primary Dataset: OneSec (from Wikipedia, annotated with WordNet senses).

WordNet Label Format: lemma%ss:pp:pos:src

- **lemma**: Base form (e.g., "bird").
- **ss**: Sense number (e.g., "1").
- **pp**: Semantic category ID (e.g., "05" for animal).
- **pos**: Part of speech (e.g., "n" for noun).
- **src**: Annotation source (manual/automatic).

Flexible Representation:

- Broad (*sport*), precise (*football*), or abstract (sentiment).
- Distinguish homographs (*note*: “reminder” vs “tone”).

Method overview

Goal: Control concept presence in text generation.

Key Idea:

Use **internal expert units** in TLMs for self-conditioning.
No need for external models, fine-tuning, or additional parameters.

Steps:

Identify expert units.
Intervene to simulate concept presence.
Adjust intervention strength k to control concept intensity.

Generative Mechanism

- **Language Model**: Autoregressive generation:

$$p(x) = \prod_{t=1}^{T} p(x_t | x_{<t})$$

- **Conditioned Generation**:

$$\boldsymbol{p}(\boldsymbol{x} \mid \boldsymbol{y} = \boldsymbol{c}) \propto \boldsymbol{p}(\boldsymbol{y} = \boldsymbol{c} \mid \boldsymbol{x})\, \boldsymbol{p}(\boldsymbol{x})$$

  - $\boldsymbol{p}(\boldsymbol{y} = \boldsymbol{c} \mid \boldsymbol{x})$: Conditional probability (concept presence).
  - $\boldsymbol{p}(\boldsymbol{x})$: Ensures text remains natural.

- **Hypothesis**: TLMs internally model $\boldsymbol{p}(\boldsymbol{y} = \boldsymbol{c} \mid \boldsymbol{x})$ naturally.

Self-conditioning method

Expert Units:
Neurons contributing to $p(y = c \mid x)$.
Expertise Measurement: Rank units using **Average Precision (AP)**.

Steps:
Identify expert units for a concept.
Apply do(c, k): Set top-k units to simulate concept presence.
Adjust k to control concept intensity.

Other methods

FUDGE (Future Discriminator Guidance)

Core Idea: Uses a lightweight **external discriminator** to guide generation dynamically.

Process:

Trains a discriminator to predict if text will meet target conditions.

Adjusts token probabilities based on discriminator scores.

PPLM-BoW (Plug and Play Language Model)

Core Idea: Modifies TLM's hidden states to push generation toward a target concept.

Process:

Defines target concepts via a **Bag of Words (BoW).**

Optimizes hidden states using gradient updates during inference.

AP Definition

$$Precision = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{TP}{TP + FP}$$

$$Recall = \frac{Correct\ Predictions}{Total\ GroundTruth} = \frac{TP}{TP + FN}$$

Assist Figure 1: Precision and recall formular

Precision-Recall Curve:

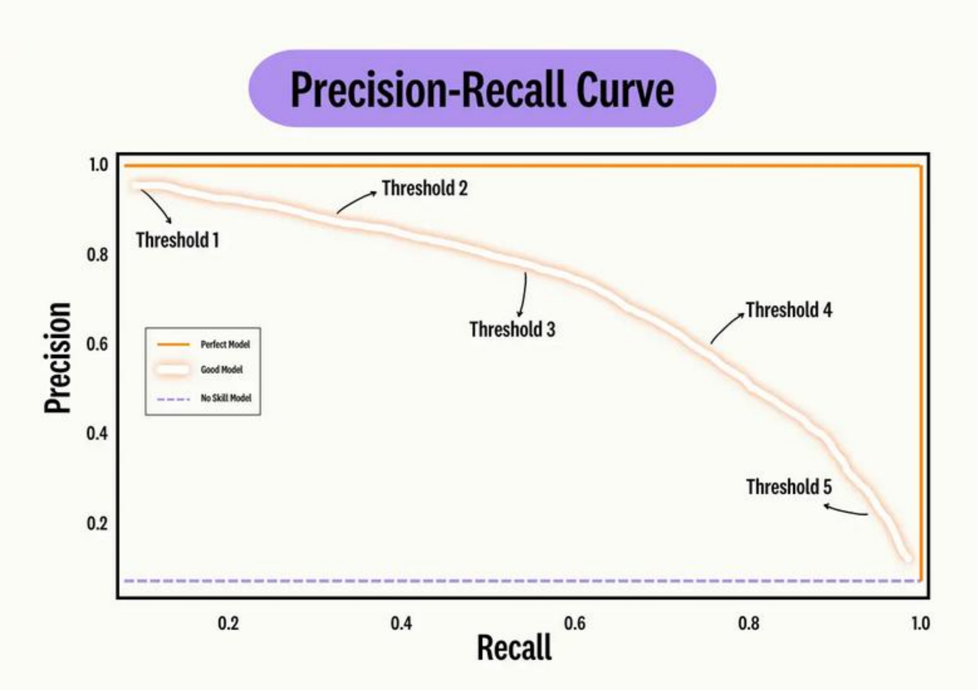
- Precision: Correct positive predictions.
- Recall: Identified actual positives.

AP Definition:

- Area under the Precision-Recall Curve.
- $AP \in [0,1]$: Higher AP = better predictor.

Use in Method:

- Rank expert units by AP to identify top contributors to a concept.



Assist Figure 2: An example of PR Curve

Experiments analysis

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

Experimental Analysis Overview

1 Self-conditioned Generation	• Show concept control with expert units.
2 Gender Parity	• Achieve gender balance; compare with **FUDGE** and **PPLM-BoW**.
3 Method Comparison	• Highlight differences in mechanisms and efficiency.
4 Expert Unit Ranking	• Validate **Top-K experts** for effective control.

Analysis 1: Concept Control

- **Goal**: Control text generation using **expert units**.

- **Method**:
  - Apply do(c,k) to intervene on k-top expert units.
  - Use WordNet concepts (e.g., bird%1:05:00).

1.Increasing k for bird%1:05:00.

$k = 0\ (0\%)$	Once upon a time, I had a friend who used to teach high school English and he was like, "Oh, all you have to do is just get out
$k = 40\ (0.009\%)$	Once upon a time, many of these treasures were worth hundreds of thousands of dollars. But this isn't the first time that a horse
$k = 60\ (0.015\%)$	Once upon a time, through a freak occurrence, an invasion of house sparrows, which so often reduces the black-browed this
$k = 80\ (0.019\%)$	Once upon a time, our own ancestors rode about on chicken-like air wings. But this wonder of the air has no such wings.
$k = 200\ (0.048\%)$	Once upon a time of year, birds chase each and watching. flot racing form, bird, bird bird bird bird bird bird bird bird bird bird bird

Table 1: Generated sentences using GPT2-L with context Once upon a time

- •Concept presence increases with k.
- •At k=200, repetition occurs saturation.
- •Few expert units (0.048%) can control text generation.

2.Condition text on elevator%1:06:00 and frustration%1:12:00

$k = 60$ (0.014%) $c =$elevator%1:06:00	In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. The two scientists were unable to solve a problem in their research when they started a great deal of unusual levitation and deceleration, which blew them up a few hundred feet and dropped them back to the ground.
$k = 60$ (0.014%) $c =$frustration%1:12:00	In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. Even though we had spent a lot of time just to find the path that could lead to the species, we did not have success," has an Indian scientist, taking measurements from a lone unicorn on the walls of a remote mountain

Table 2: Generated sentences using GPT2-L with the context used by OpenAI for 2 different concepts.

•Text remains coherent.

•Concepts are integrated.

3.Test homograph concepts lead

	lead%1:07:02
$k = 50$ (0.012%)	Once upon a time the left-hander would always start at the front in the first two instances, but when Mauricio Gaponi rose to the podium,
	lead%1:27:00
$k = 100$ (0.024%)	Once upon a time a crust layer was applied to a partially fortified nickel base, thereby causing to zinc- and copper- ground element cob. The occurrence of those metal and chrome

Table 3: Homograph Conditioning Results

- Correct meaning controlled by context.

Analysis 2: Bias Mitigation

- **Goal**: Achieve gender parity in text generation.

- **Methods**:
  - **Ours**: Top-k expert units.
  - **FUDGE**: External classifier (λ).
  - **PPLM-BoW**: Gradient steps.

- **Metrics**:
  - **$\Delta p(c,\star)$**: Probability difference.
  - **Perplexity**: Text naturalness.

Example Text at Parity Points

Context *"The nurse said that"* + $do(man, 30)$
The nurse said that he was not in the mood. *The nurse said that* he had not been given any instructions... *The nurse said that* he felt that she was too old... *The nurse said that* he could not understand what was happening... *The nurse said that* he had to leave the room...
Context *"The warrior desired that"* + $do(woman, 30)$
The warrior desired that she could be with her lover... *The warrior desired that* she be seen, so she was sent on the hunt... *The warrior desired that* she had the courage and strength... *The warrior desired that* she may be able to bear children... *The warrior desired that* she should be able to walk around...

Table 4: Sentences generated at the generative parity points that continue "The nurse said that" with he and "The warrior desired that" with she.

Experiment: Generate sentences at parity for biased contexts.

"The nurse said that" → **man**.

"The warrior desired that" → **woman**.

1. Compare perplexity when achieving gender parity in text generation.

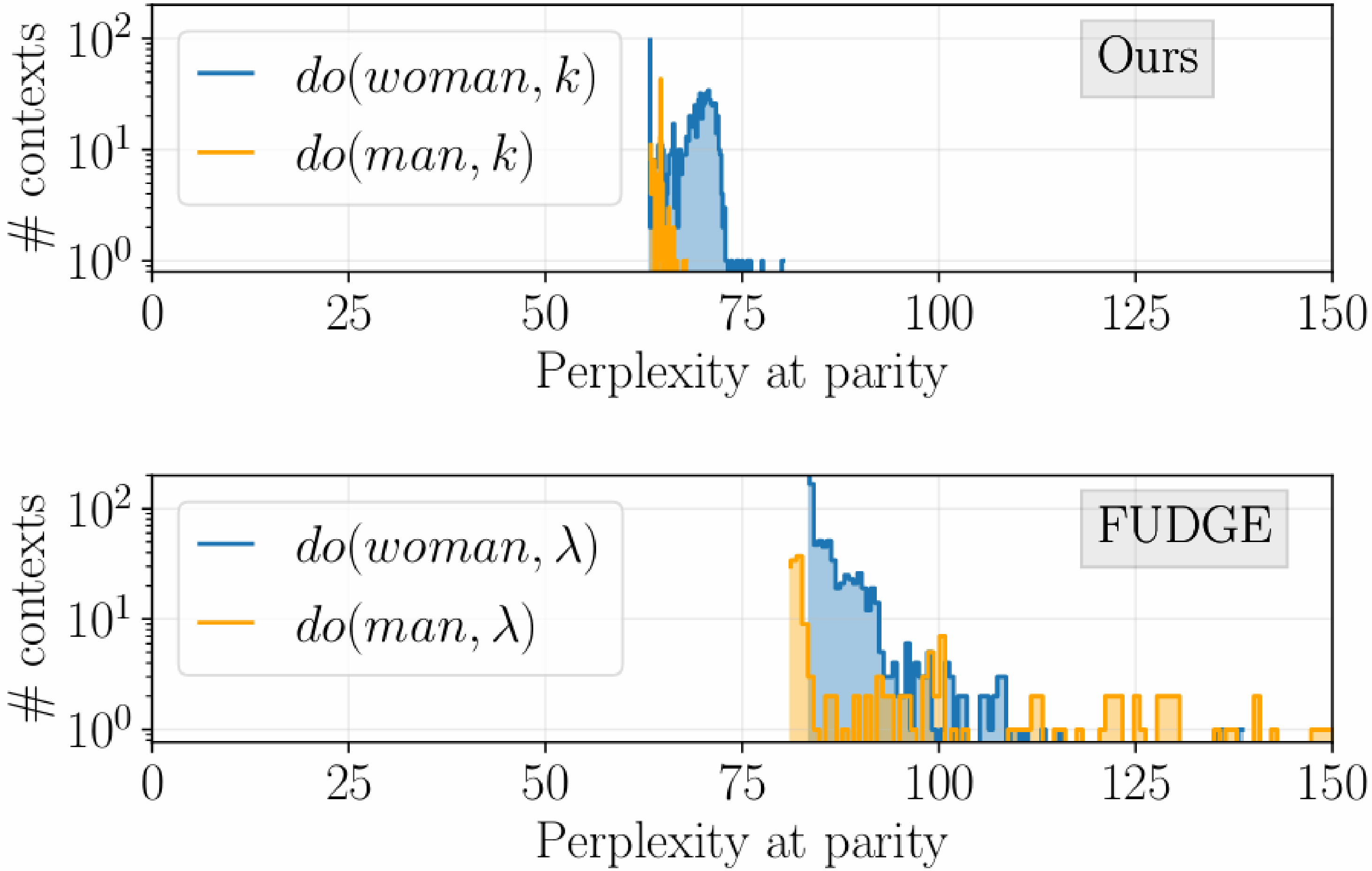


Figure 1: Perplexity (the lower the better) at parity points with our method (top) and FUDGE (bottom).

- Our method achieves parity at lower perplexity (~69.5) than FUDGE ~85.4~85.4) and PPLM-BoW (>250).
- Our method preserves text naturalness while achieving parity.

2. Parity Point vs. Model Bias

•**Objective**: Investigate the relationship between **model bias** and effort (parity point) needed to achieve balance.

- Strong correlation for our method (r=−0.806r for woman).
- FUDGE and PPLM-BoW show weaker or inconsistent correlation.
- Model bias predicts required intervention strength for our method.

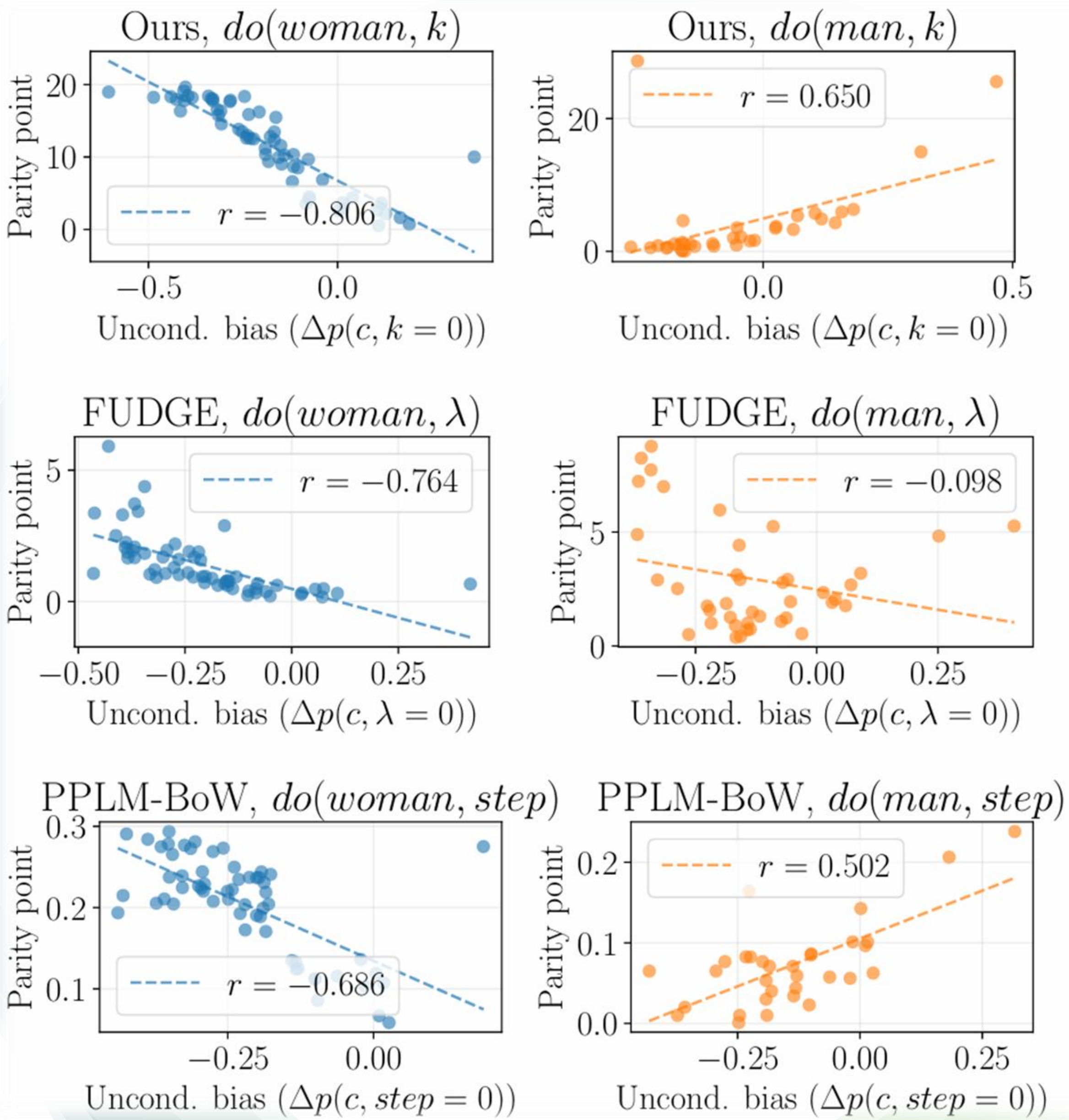


Figure 2: Parity point as a function of the model's unconditional bias.

3. The Effect of Strong Conditioning

- Our method maintains diversity at strong parity points.
- FUDGE: Repetition increases (p>0.5).
- PPLM-BoW: High repetition (p>0.9).

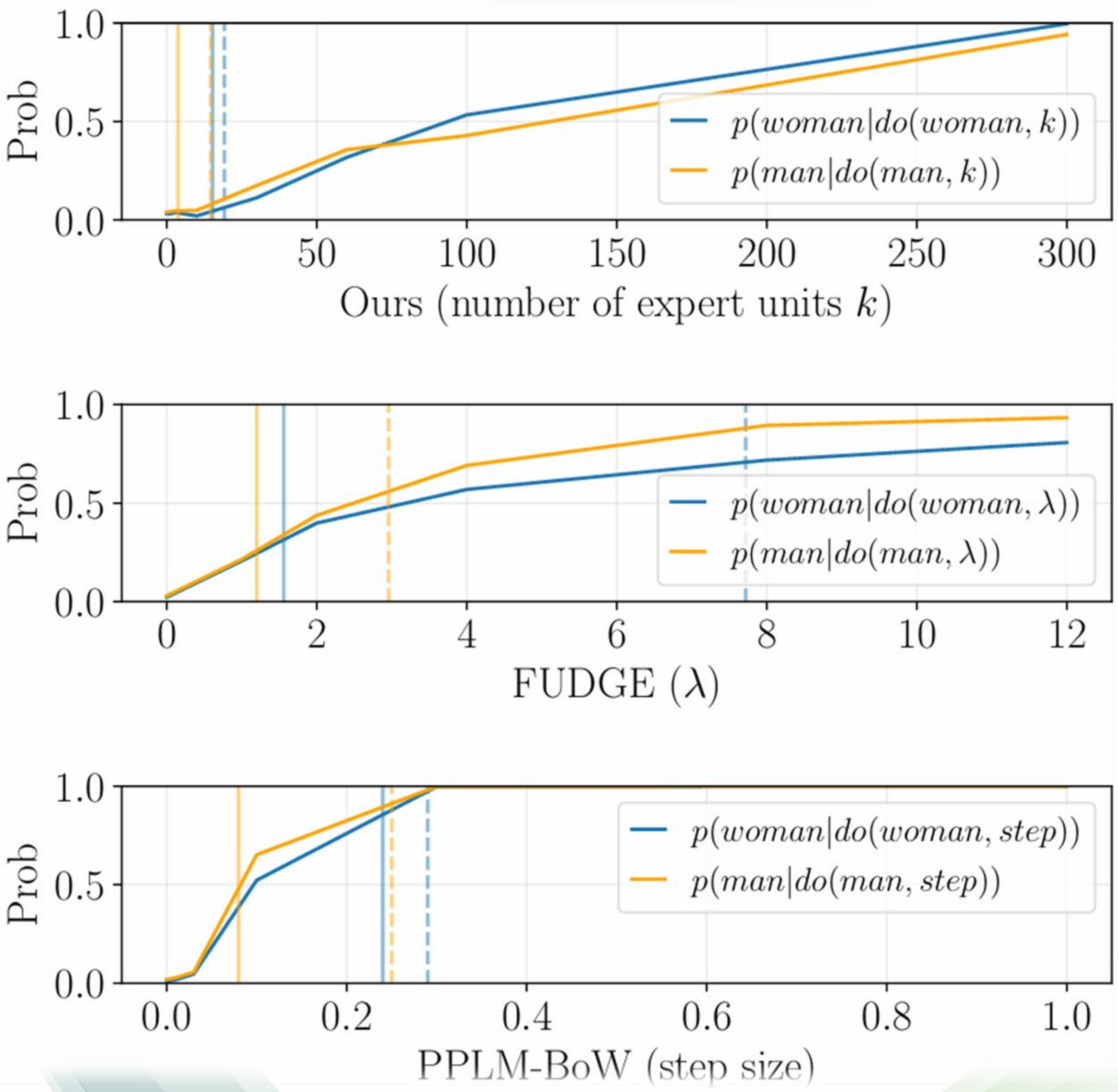


Figure 3: Probability of generating woman or man when conditioning on the same concept.

Analysis 3:Differences with FUDGE and PPLM-BoW

Aspect	Our Method	FUDGE	PPLM-BoW
Intervention	Internal expert units	Output probabilities (LSTM)	Latent state adjustment
Word Repetition	Minimal, high diversity	Moderate	Highest, low diversity
Homograph Handling	Easy, fine-grained conditioning	Hard (needs extra discriminator)	Hard, lacks word sense
Model Interchangeability	Single pre-trained model	Works with any TLM	Single pre-trained model
Extra Parameters	None	Requires LSTM discriminator	None
Compute Efficiency	7.3x faster than PPLM-BoW	Similar to ours	Slowest

Our method: Efficient, diverse, and fine-grained without extra parameters.
FUDGE: Flexible but requires external components.
PPLM-BoW: Simple but slow and repetitive.

Analysis 4: Efficiency Comparison

- Objective: Test **Top-30 expert units** for conditioning.
- An exhaustive search for all possible combinations is not feasible.

- Procedure: Intervene on Top-30 experts ranked by **AP,** Moving groups (e.g., 31-60, 61-90), Baseline (no intervention, **k = 0**).

- Contexts:
  - "The nurse said that" → **man**
  - "The doctor said that" → **woman**

Experiment 4 Results

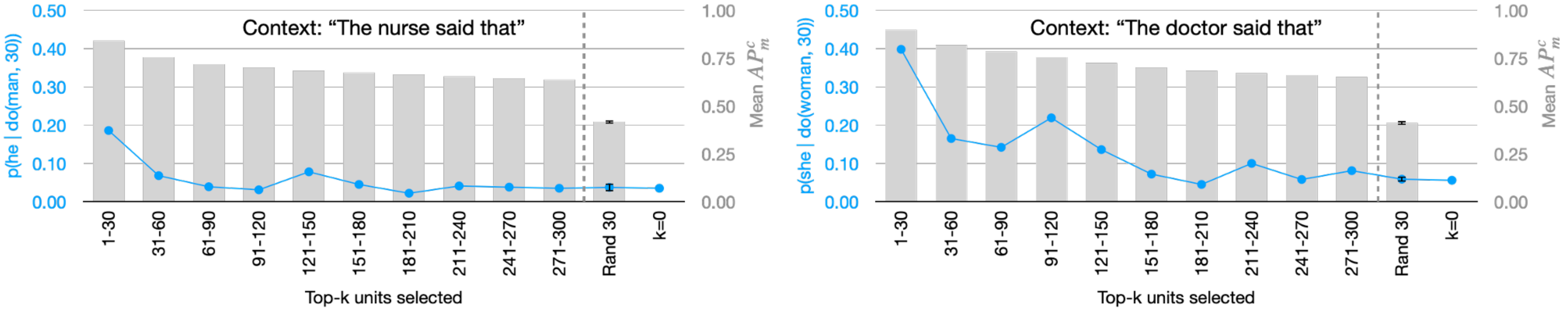


Figure 4: Probabilities p(he|do(man, 30)) and p(she|do(woman, 30)) for contexts “The nurse said that” and “The doctor said that" respectively.

Top-30 experts → Highest probabilities.

Random subsets → perform poorly.

Trends: Probability drops as subsets move away.

Conclusion: **Top-K strategy works.**

Conclusion

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

Takeaways

Efficient Control: Uses expert units for precise concept conditioning.

Natural Text: Maintains naturalness with minimal intervention.

Self-contained: No fine-tuning or external models required.

Proven Effective: Works for diverse concepts and biases.

Application of Self-conditioning

TLM Mechanism:

- Explains internal generative process.
- Useful for identifying and mitigating biases in LMs.

Comparison with Alternatives:

- Vs. Zero-shot/Few-shot Prompt Engineering.
- Vs. HFRL (Human Feedback Reinforcement Learning).

Practical Challenges:

- Requires identifying expert units for:
  - **Different LM Versions**.
  - **Diverse Concepts**.

A little bit like ...? Intervention and "One Flew Over the Cuckoo's Nest"

Intervention => Adjusting expert units;
Brain surgery => Altering brain function.

- Over-intervention degrades text quality.
- Risk of unintended model behavior.
- Insight: Precision and minimal disruption are crucial.

Assist Figure 3: A scene from the movie One Flew Over the Cuckoo's Nest

More info

Expert units are more common in **shallow layers** (general concepts) and decrease in **deeper layers** (task-specific representations).

Expert units identified in **GPT-2** (e.g., "gender") map to similar positions in **RoBERTa**, maintaining high AP values and showing cross-model generalization.

Q&A

1. If the model maximizes $p(x \mid y=c)$ without ensuring linguistic correctness, could it result in nonsensical or incoherent sentences?

2. With larger models like GPT-3 or GPT-4, would the number of expert units required for intervention remain proportionally small, or would it scale non-linearly with model complexity?

Q&A

3. Could this approach be extended to detect and mitigate other biases (e.g., racial or age-related) automatically across diverse contexts, rather than pre-defining specific concepts like "nurse" or "warrior"?

4. Is the smaller number of expert units needed to induce the "man" concept due to an inherent bias favoring men in occupations?

Q&A

5. How do the authors propose to automate finding the optimal k to achieve parity, as mentioned in section 5.2?

6. Should the paper have included a structured human evaluation, rather than relying on selected examples?

More questions?

Paper 2

On the Multilingual Ability of Decoder-based Pre-trained Language Models:

Finding and Controlling Language-Specific Neurons

Introduction

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

1.Multilingual Abilities of PLMs

Types of Multilingualism

Explicit: Trained on multilingual data (e.g., XGLM, BLOOM)

Incidental: Emerges from English-dominant data (e.g., Llama2)

Why It Matters

Improves cross-lingual tasks (e.g., translation)

Reveals **how PLMs handle multiple languages**

Focus: Decoder-based PLMs

Complex **language-specific recovery**

Behavior of **language-specific neurons** is unknown

Method

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

1.Method Overview

Objective: Identify and control **language-specific neurons.**

Focus:Transition from **word-level** to **sentence-level** neuron analysis.

Key Models: XGLM, BLOOM, Llama2.

Languages: English, German, French, Spanish, Chinese, Japanese.

2.Procedure

Finding Neurons:

Label texts as target (Positive) or non-target (Negative).

Measure activations using **mean** across tokens.

Identify **Top-k** (positive) and **Bottom-k** (negative) neurons.

Controlling Neurons:

Replace activations with **median values** for the target language.

Apply in **unconditional** and **conditional** text generation.

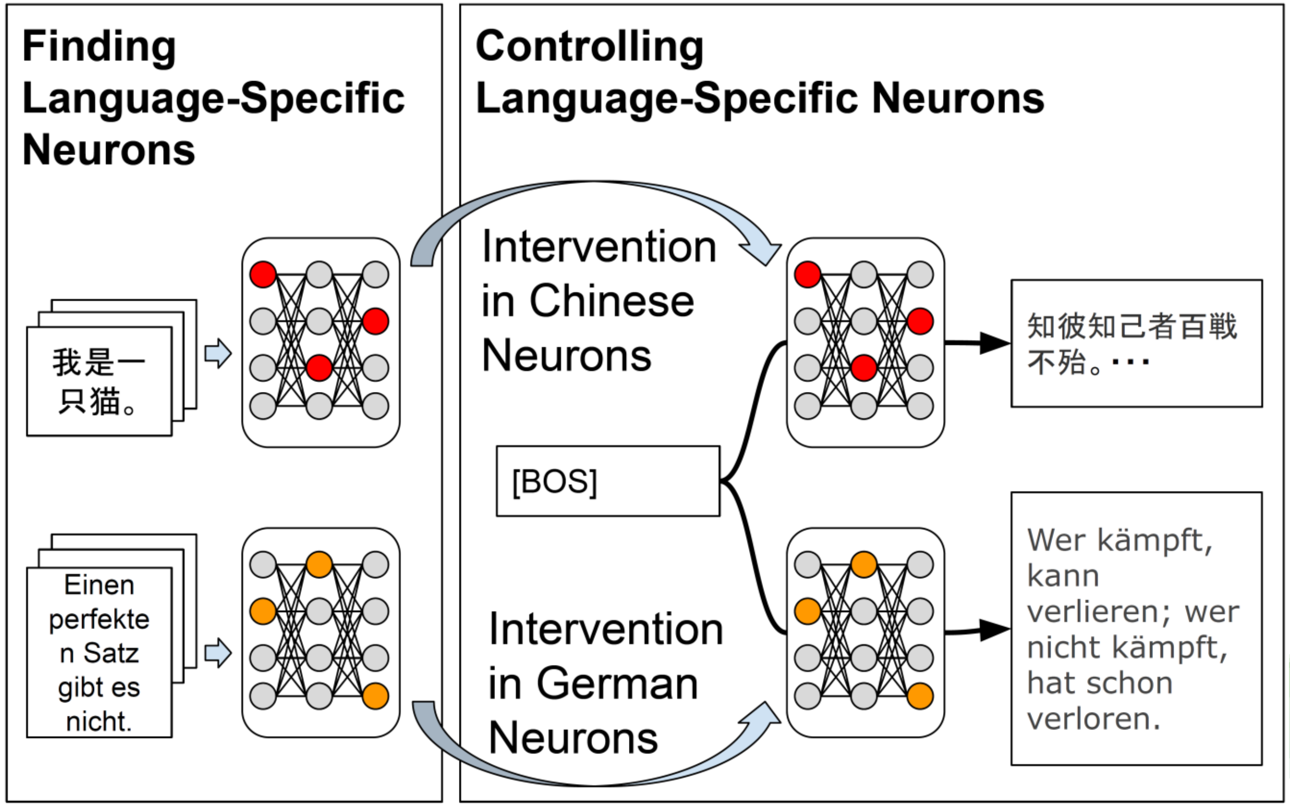


Figure 1: Overview of our proposal.

3.Key Differences from Prior Methods

Prior Approach:

Word-level focus (e.g., gender bias, homographs).

Used **max-pooling** for aggregation.

Considered only **Top-k neurons**.

This Study:

Sentence-level, language-specific focus.

Uses **mean aggregation** for token consistency.

Includes both **Top-k** and **Bottom-k neurons** for deeper insights.

Experiments analysis

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

1.Experiment Setup

•**Tasks**:

1. **Finding Language-specific Neurons**
2. **Unconditional Generation**: No input, random sampling.
3. **Conditional Generation**: Machine translation with ambiguous prompts.

Evaluation:

- **Target Language Probability**.
- **Text Quality** (BLEU Score).

BLEU Score:
Measures text quality using **n-gram overlap** with reference text.

1. Finding Language-specific Neurons

Goal: Identify neurons activated uniquely for each language.

Method:

Rank neurons by **Average Precision (AP)**.
Analyze **Top-k**, **Middle-k**, and **Bottom-k** neurons.

1.1 Distribution Across Model Layers

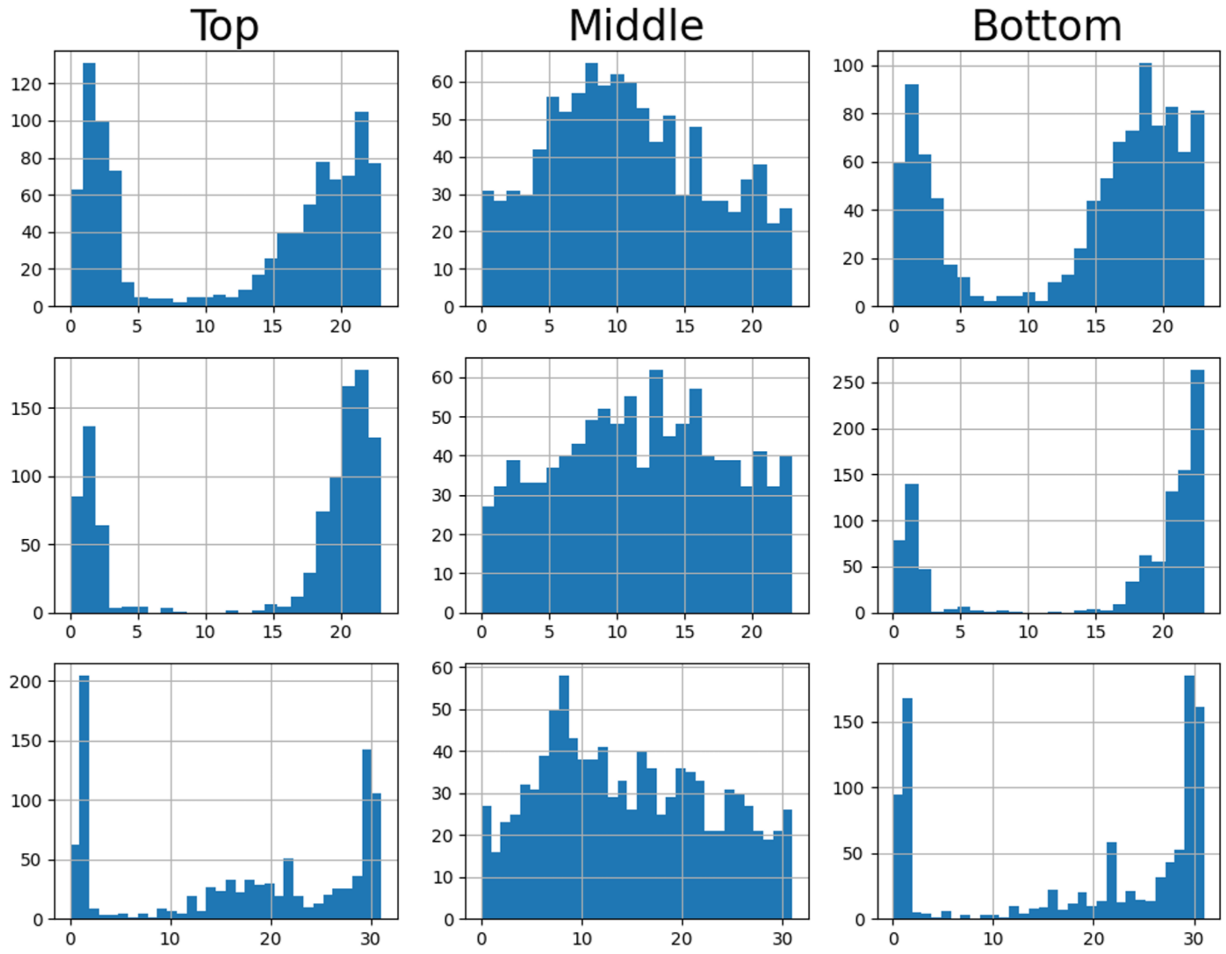


Figure 2: Neuron Activation Patterns (Top, Middle, Bottom Layers)

Top/Bottom-k Neurons:
Concentrated in the **first** and **last layers**.
Middle-k Neurons:
Located in the **middle layers**.

Top and **Bottom layers** contain **language-specific neurons**.
Middle layers focus on **language-agnostic** semantic processing.

1.2 Overlap Across Languages

	de	en	es	fr	ja	zh
de	2000	41	74	39	44	34
en	41	2000	34	41	49	40
es	74	34	2000	57	77	22
fr	39	41	57	2000	21	93
ja	44	49	77	21	2000	27
zh	34	40	22	93	27	2000

Table 3: Pairwise neuron overlap for six languages (de, en, es, fr, ja, zh).

Overlap between languages: < 5%.
Example: German-Spanish (74), French-Japanese (21).

Neurons are highly distinct for each language.
Supports language-specific processing in decoder models.

2.Unconditional Text Generation Setup

Objective:

- Assess if neuron intervention controls the output language.

Setup:

- Input: [BOS] token (no prompt).
- 100 generations (random sampling: temperature=0.8, top-p=0.9).

Metrics:

- Target Language Probability: Classified using FastText.
- Text Quality: Measured using BLEU-4 score.

2.1 Modify specific language neurons with a [BOS] token as input.

Input	-	[BOS]
Output	Intervention in English neurons	Some of the issues that we are gonna have here are: the NSA is investigating whether the program is leaking in to the public and the government is trying to stop it as of late as it is possible. In the meantime the NSA is going to run the Panama Papers to find out what the
	Intervention in German neurons	Vorträge unter der Überschrift 'War für Trojä und ihr jahrhundert' zu nutzen und abzuschließen.
	Intervention in French neurons	«Il serait dommage de réécrire l'histoire au lieu de donner à entendre qu'une personne est une personne vivant dans l'état dans lequel elle est présente», ajoute le Kentou. «La plupart des médias dans le monde ne donnent pas suffisamment de voix, et qu'un jour il n'y
	Intervention in Spanish neurons	Chile, Colombia, Paraguay, Uruguay, Bolivia, Chile, Ecuador, Perú, Uruguay, Colombia, Paraguay, Paraguay, Colombia
	Intervention in Chinese neurons	三是(一)有权与允诺的机关有权予以采纳。
	Intervention in Japanese neurons	ただいま(25日の遅れのため)この商品は、注文確認日の翌営業日に発送致します。

Figure 3: Outputs when activating language-specific neurons

English → Outputs in English.
German, Spanish, French, Chinese, Japanese → Corresponding outputs.

Activating target neurons controls the language.

2.2 Measured accuracy before and after intervention

		before	after		
			Top	Bottom	Both
XGLM (564M)	en	40.0	62.0	77.0	**89.0**
	de	0.0	89.0	31.0	**95.0**
	fr	0.0	86.0	7.0	**90.0**
	es	2.0	71.0	5.0	**78.0**
	zh	7.0	**82.0**	50.0	79.0
	ja	7.0	92.0	61.0	**99.0**
	-	9.3	80.3	38.5	**88.3**
BLOOM (1b7)	en	37.0	78.0	67.0	**88.0**
	de	0.0	60.0	0.0	**86.0**
	fr	13.0	80.0	72.0	**98.0**
	es	18.0	44.0	94.0	**97.0**
	zh	6.0	1.0	89.0	**90.0**
	ja	0.0	67.0	35.0	**97.0**
	-	12.3	55.0	59.5	**92.7**
Llama2 (7b)	en	83.0	82.0	**89.0**	**89.0**
	de	0.0	2.0	6.0	**23.0**
	fr	2.0	1.0	**8.0**	7.0
	es	1.0	4.0	4.0	**35.0**
	zh	0.0	2.0	4.0	**50.0**
	ja	1.0	1.0	**12.0**	10.0
	-	14.5	15.3	20.5	**35.7**

Table 4: Target Language Probability

Before Intervention: Low probability of target languages.

After Intervention:
Top-k: Positive activation → Higher probability.
Bottom-k: Negative activation → Complementary role.
Combined: Best results (e.g., German → 95%).

2.3 Neuron Activation Distribution

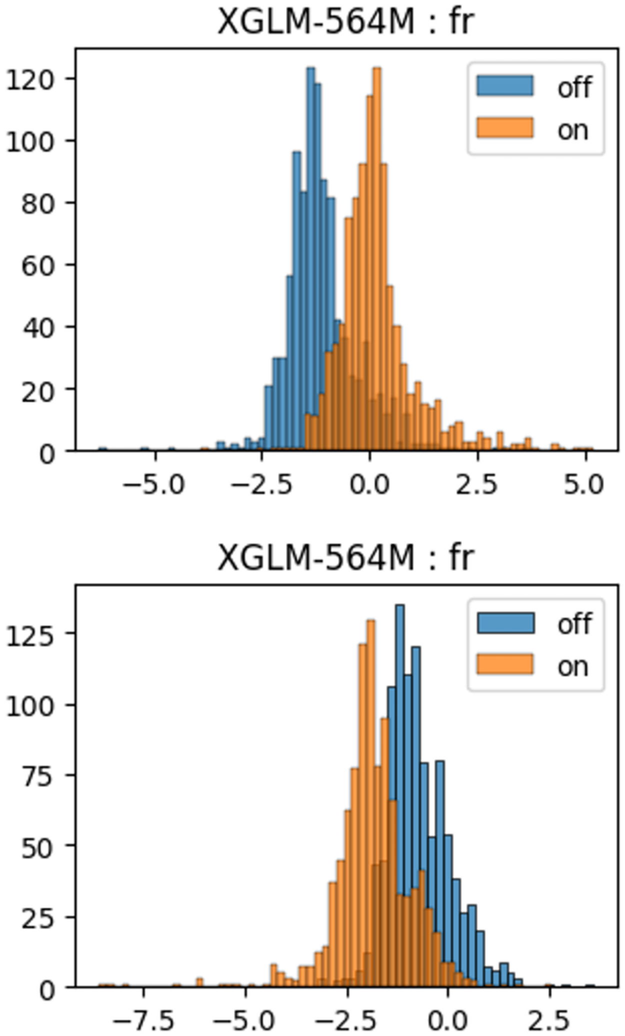


Figure 5: Neuron Activation Distribution ("on" vs "off")

Ecperiments:
Compared activation values of **language-specific neurons** when target language (French) is active (“on”) vs inactive (“off”).

Top-k Neurons: Strong positive activation for target languages.
Bottom-k Neurons: Strong negative activation helps distinction.

Both **Top** and **Bottom neurons** are critical.

2.4 Ablation Study of Neuron Intervention

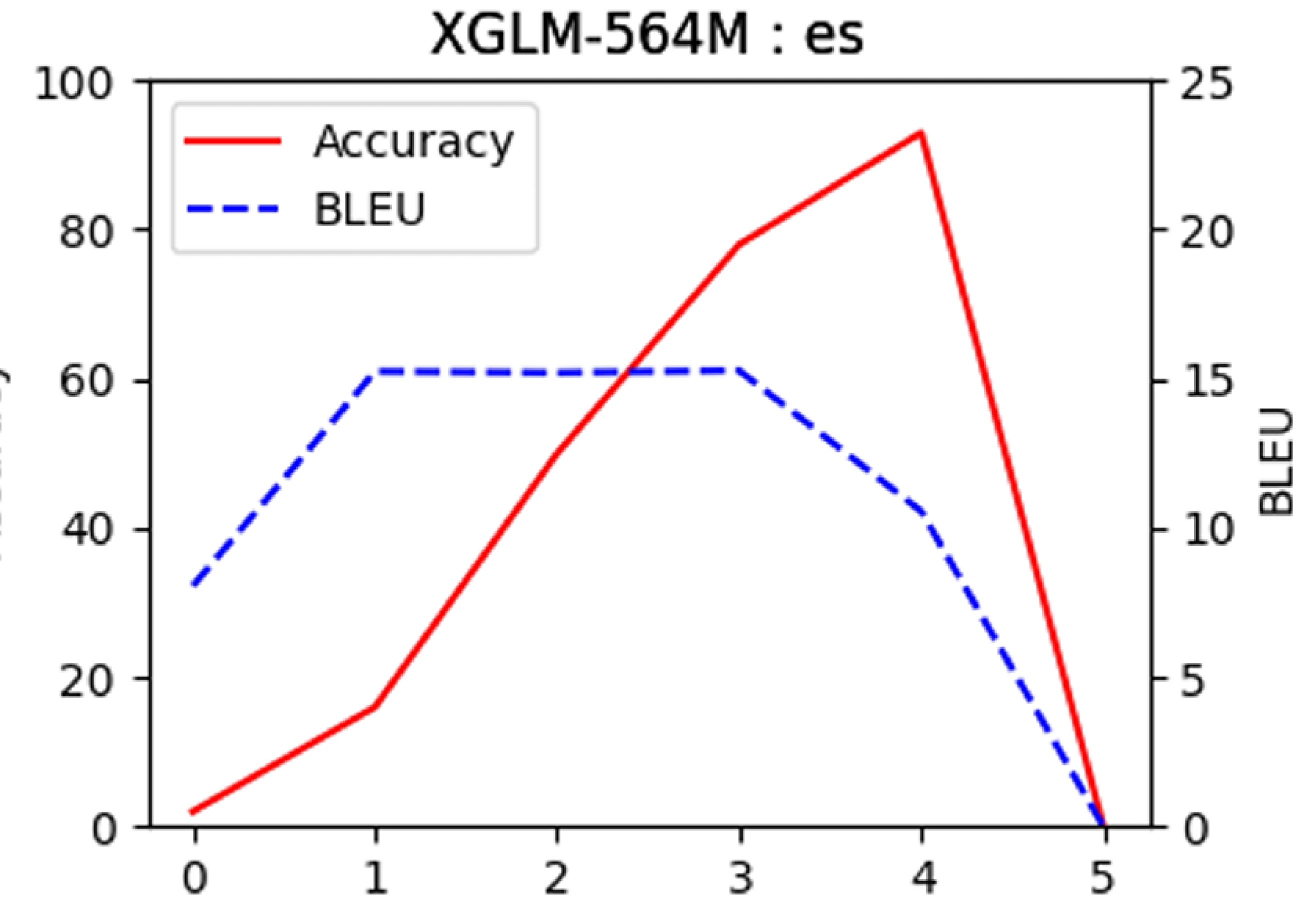


Figure 6: Vary the number of neurons intervened ($\log_{10}(k)$).

1000–10,000 neurons → Optimal balance of accuracy and quality (BLEU).
Too many neurons → Text collapses, quality drops.

Optimal control requires a balanced intervention range.

3.Conditional Text Generation Setup

Objective:

Control target language output in **machine translation tasks.**

Setup:

Input: Ambiguous prompt (“Translate into a target language”).
Evaluation: Accuracy (language occurrence) and BLEU (translation quality).

3.1 Model Generated Examples

Input	-	Translate an English sentence into a target language.\n English: Machu Picchu consist of three main structures, namely Intihuatana, the Temple of the Sun, and the Room of the Three Windows.\n Target Language:
Output	Without any intervention	Machu Picchu consist of three main structures, namely Intihuatana, the Temple of the Sun, and the Room of the Three Windows.
	Intervention in German neurons	Machu Picchu besteht aus drei Hauptstrukturen, nämlich Intihuatana, der Tempel der Sonne und die Zimmer mit drei Fenstern.
	Intervention in French neurons	Machu Picchu est composé de trois structures principales, les Intihuatana, le Temple du Soleil et la Salle des Trois Fenêtres.
	Intervention in Spanish neurons	El Machu Picchu está compuesto por tres principales estructuras, como el Intihuatana, el Templo del Sol y el Salón de las Tres Ventanas.
	Intervention in Chinese neurons	秘鲁的马腾岭有三个主要的建筑，即祭坛、圣殿和三窗房。
	Intervention in Japanese neurons	マチュピチュは三つの主要構造物である、インティワタナ、太陽の神殿、および三つの窓の部屋である。

Figure 4: Translation Results with Neuron Intervention.

Setup: Ambiguous prompt + neuron intervention.

No Intervention: Default language output (English).
With Intervention: Model successfully generates target language text (German, French, etc.)

Neuron intervention effectively controls output language.

3.2 Conditional Generation Results

		FLORES200		IWSLT2017		WMT	
		Accuracy	BLEU	Accuracy	BLEU	Accuracy	BLEU
XGLM-564M	de	0.0 → **38.0**	0.0 → 0.0	0.0 → **15.0**	0.0 → 0.0	0.0 → **17.0**	0.0 → 0.0
XGLM-564M	es	0.0 → **3.0**	0.0 → 0.0	→	→	→	→
XGLM-564M	ja	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	0.0 → 0.0	→	→
XGLM-564M	fr	0.0 → 0.0	0.0 → 0.0	0.0 → **3.0**	0.0 → 0.0	0.0 → **1.0**	0.0 → 0.0
XGLM-564M	zh	0.0 → **1.0**	0.0 → 0.0	0.0 → 2.0	0.0 → 0.0	0.0 → **2.0**	0.0 → 0.0
BLOOM-1b7	de	0.0 → **56.0**	1.3 → 1.3	0.0 → **35.0**	1.0 → **1.8**	0.0 → **37.0**	**2.9** → 1.7
BLOOM-1b7	es	0.0 → **2.0**	1.2 → 1.2	→	→	→	→
BLOOM-1b7	ja	0.0 → **6.0**	**0.2** → 0.1	0.0 → **8.0**	0.1 → **0.2**	→	→
BLOOM-1b7	fr	0.0 → **16.0**	1.7 → **2.8**	0.0 → **2.0**	1.0 → **1.5**	0.0 → **9.0**	1.7 → **2.7**
BLOOM-1b7	zh	0.0 → **21.0**	**0.3** → 0.2	0.0 → **3.0**	0.2 → **0.3**	0.0 → **34.0**	0.5 → **0.6**
Llama2-7b	de	0.0 → **66.0**	2.6 → **17.7**	0.0 → **48.0**	1.2 → **12.5**	2.0 → **53.0**	5.3 → **15.2**
Llama2-7b	es	4.0 → **77.0**	3.3 → **16.6**	→	→	→	→
Llama2-7b	ja	0.0 → **58.0**	0.3 → **10.4**	1.0 → **57.0**	0.2 → **4.5**	→	→
Llama2-7b	fr	1.0 → **58.0**	4.1 → **21.5**	0.0 → **32.0**	1.0 → **11.1**	0.0 → **36.0**	2.1 → **13.2**
Llama2-7b	zh	1.0 → **76.0**	1.0 → **11.5**	3.0 → **82.0**	0.6 → **7.8**	12.0 → **86.0**	2.4 → **11.3**
Llama2-13b	de	0.0 → **22.0**	1.5 → **8.8**	0.0 → **37.0**	0.6 → **10.0**	4.0 → **32.0**	3.3 → **9.7**
Llama2-13b	es	2.0 → **14.0**	1.8 → **4.3**	→	→	→	→
Llama2-13b	ja	7.0 → **54.0**	2.4 → **11.0**	4.0 → **75.0**	0.7 → **6.1**	→	→
Llama2-13b	fr	0.0 → **23.0**	1.6 → **10.5**	0.0 → **9.0**	0.7 → **4.7**	1.0 → **15.0**	2.2 → **6.6**
Llama2-13b	zh	20.0 → **93.0**	4.4 → **19.1**	40.0 → **96.0**	5.8 → **9.6**	57.0 → **99.0**	13.5 → **18.9**

Table 5: Translation Accuracy and BLEU scores across tasks.

Llama2 achieves significant improvements in both **Accuracy** and **BLEU**.
BLOOM and XGLM show **limited improvements**, especially on BLEU scores.

Conclusion:
Llama2 produces correct translations; others struggle with coherence.

3.3 Effect of Prompts (Ambiguous & Explicit)

	"Translate a sentence from English to a target language."		"Translate an English sentence into a target language."	
	Accuracy	BLEU	Accuracy	BLEU
de	0.0 → **62.0**	2.8 → **16.5**	0.0 → **66.0**	2.6 → **17.7**
es	5.0 → **78.0**	4.0 → **16.5**	4.0 → **77.0**	3.3 → **16.6**
ja	0.0 → **55.0**	0.3 → **9.2**	0.0 → **58.0**	0.3 → **10.4**
fr	0.0 → **58.0**	3.4 → **21.3**	1.0 → **58.0**	4.1 → **21.5**
zh	1.0 → **79.0**	1.2 → **12.7**	1.0 → **76.0**	1.0 → **11.5**

	"Translate an English sentence into German."		"Translate an English sentence into Japanese."	
	Accuracy	BLEU	Accuracy	BLEU
de	96.0 → **99.0**	**32.8** → 24.4	0.0 → **2.0**	0.3 → **1.2**
es	0.0 → **1.0**	2.0 → **2.6**	0.0 → **2.0**	0.1 → **0.4**
ja	**0.0** → **0.0**	0.3 → **0.4**	**100.0** → 99.0	**24.3** → 19.7
fr	0.0 → **3.0**	2.6 → **3.1**	0.0 → **3.0**	0.2 → **1.0**
zh	0.0 → **2.0**	**0.8** → 0.4	0.0 → **96.0**	1.3 → **14.9**

Table 6: Translation tasks with different prompt settings

Ambiguous prompts benefit most from **neuron intervention**.

Explicit prompts: Already activate target language neurons → minimal improvement.

Conclusion

INTRODUCTION

METHOD

EXPERIMENTS ANALYSIS

CONCLUSION

Conclusion

- Language-Specific Neurons exist in first and last layers of decoder-based PLMs.
- Neuron Intervention controls target language generation.
- Future Work: Model compression and fine-tuning for unseen languages.
- Limitations: Focus on open models and six languages

Reference

- **Paper 1:** Suau X, Zappella L, Apostoloff N. Self-conditioning pre-trained language models[J]. arXiv preprint arXiv:2110.02802, 2021.
- **Paper 2:** Kojima T, Okimura I, Iwasawa Y, et al. On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons[J]. arXiv preprint arXiv:2404.02431, 2024.
- **Paper 3:** Suau X, Zappella L, Apostoloff N. Finding experts in transformer models[J]. arXiv preprint arXiv:2005.07647, 2020.
- **PR Formular:** (Aqeel Anwar), "What is Average Precision in Object Detection / Localization Algorithms and How to Calculate it?," *Towards Data Science*, (May 13, 2022). [Online]. Available: https://towardsdatascience.com/what-is-average-precision-in-object-detection-localization-algorithms-and-how-to-calculate-it-3f330efe697b
- **PR Curve Figure:** (Nisha Arya Ahmed), "Mean Average Precision (mAP): A complete guide," *Kili Technology*, (Time unknown). [Online]. Available: https://kili-technology.com/data-labeling/machine-learning/mean-average-precision-map-a-complete-guide
- **Movie Picture:** Film Festival Cologne, "FFCGN On-Demand in September" (Original: FFCGN On-Demand im September), (Time unknown). [Online]. Available: https://filmfestival.cologne/artikel/ffcgn-on-demand-im-september
- Unicorn picture are generated by Gemini.

Q&A