FACULTY OF ARTS AND PHILOSOPHY

DEPARTMENT OF TRANSLATION, INTERPRETING AND COMMUNICATION
LT3 LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

TAILORING MT FOR SCIENTIFIC LITERATURE THROUGH TOPIC FILTERING AND FM AUGMENTATION

Thomas Moerman, Tom Vanallemeersch, Sara Szoc and Arda Tezcan

GHENT UNIVERSITY

language and translation technology team

CROSSLANG TRANSLATION AUTOMATION

fwo

THE WHY – INTRO & MOTIVATION

- **Problem**
  - Science is global, but language barriers limit access
  - English dominance creates inequalities
  - Scientific information reaches a limited audience

THE WHY – INTRO & MOTIVATION

- **Challenge for MT**
  - Lack of domain-specific parallel data
    - Specialized terminology
    - Complex syntax
    - Domain-specific discourse

THE WHY – INTRO & MOTIVATION

- **Our Goal**
  - Improve MT for specialized scientific domains
    - Neuroscience
    - Climatology
    - Mobility
  - By making efficient use of existing data
    - Topic Filtering + Fuzzy Match Augmentation

GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

BACKGROUND

BACKGROUND: NMT & LLMS

- **Transformer-based NMT** has significantly improved translation quality
- **LLMs** show impressive results in general translation tasks (Kocmi et al., 2024)
- For **specialized domains**, well-trained NMT can still outperform LLMs (Wassie et al., 2025; Higashiyama, 2024)
- However, **NMT vs LLMs** has not been done extensively

GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

BACKGROUND: DATA SELECTION & FMs

- **Data Selection:** Intelligently picking relevant sentences from general corpora (Chu & Wang, 2018)
  - **through Topic Filtering:** Selecting (potentially) domain-relevant data based on topic signals
- **Fuzzy Match Augmentation:** Leveraging translations of similar sentences (Tezcan et al., 2021; Moslem et al., 2023)
  - Known to work well in combination with other data augmentation methods (e.g. backtranslation, Tezcan et al., 2024)

OUR CONTRIBUTION

- Investigate the **combination** of **Topic Filtering** and **Fuzzy Match Augmentation** in the **scientific** domain
- Apply it to both NMT and LLMs and compare, focusing on:
  - Comparing performance across model architectures
  - Trade-offs between specialized NMT and (off-the-shelf) LLMs

GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

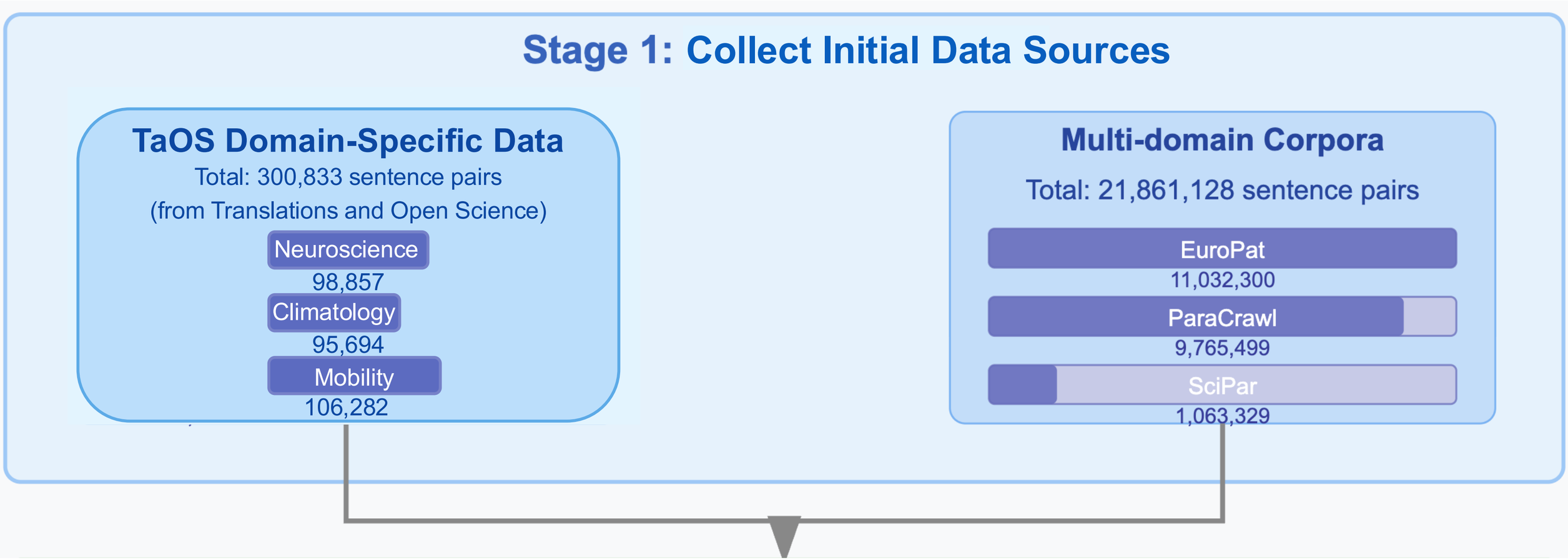
METHOD 1 – TOPIC FILTERING

- **Observation**
  - Large general corpora (ParaCrawl, EuroPat, SciPar) contain useful sentences
  - BUT most are irrelevant to specific scientific domains
- **Solution: data selection through topic filtering**

GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

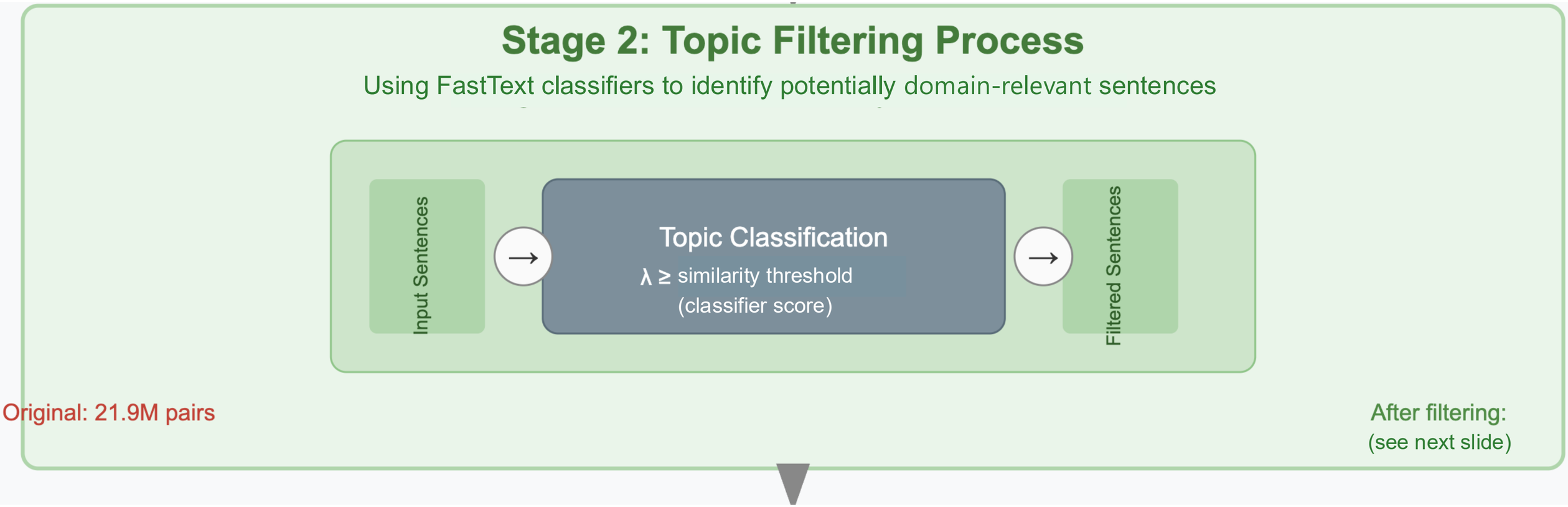
METHOD 1 – TOPIC FILTERING



GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

METHOD 1 – TOPIC FILTERING



METHOD 1 – TOPIC FILTERING

Neuroscience

5,057,229 sentence pairs

(23.1% of original data)

Original: 21.9M pairs

Filtered: 5.1M pairs (23.1%)

Data Sources

EuroPat: 2,156,482 (19.5%)

ParaCrawl: 2,508,710 (25.7%)

SciPar: 392,037 (36.9%)

Percentages show proportion kept after filtering

Climatology

10,185,899 sentence pairs

(46.6% of original data)

Original: 21.9M pairs

Filtered: 10.2M pairs (46.6%)

Data Sources

EuroPat: 6,998,414 (63.4%)

ParaCrawl: 2,713,013 (27.8%)

SciPar: 474,472 (44.6%)

Percentages show proportion kept after filtering

Mobility

8,824,756 sentence pairs

(40.4% of original data)

Original: 21.9M pairs

Filtered: 8.8M pairs (40.4%)

Data Sources

EuroPat: 2,610,923 (23.7%)

ParaCrawl: 5,879,689 (60.2%)

SciPar: 334,144 (31.4%)

Percentages show proportion kept after filtering

METHOD 1 – TOPIC FILTERING

Neuroscience | Climatology | Mobility | External Data (Original) | Topic-Filtered External Data

1d

Single Discipline Data

~100K pairs per domain

Neuroscience
98,857 sentence pairs
TaOS domain-specific data

Climatology
95,694 sentence pairs
TaOS domain-specific data

Mobility
106,282 sentence pairs
TaOS domain-specific data

One model trained per domain

3d

All Disciplines Combined

~300K pairs total

98,857 pairs

95,694 pairs

106,282 pairs

Total: 300,833 pairs
All TaOS domain data

One model trained on all domains

3d+Ext

All TaOS + All External

~22.2M pairs total

TaOS domains

300,833 pairs

(shown to scale)

External Data
21.9M pairs (unfiltered)

Large but potentially noisy dataset

3d+ExtTF

TaOS + Topic-Filtered Data

~5-10M pairs per domain

TaOS domains

300,833 pairs

(shown to scale)

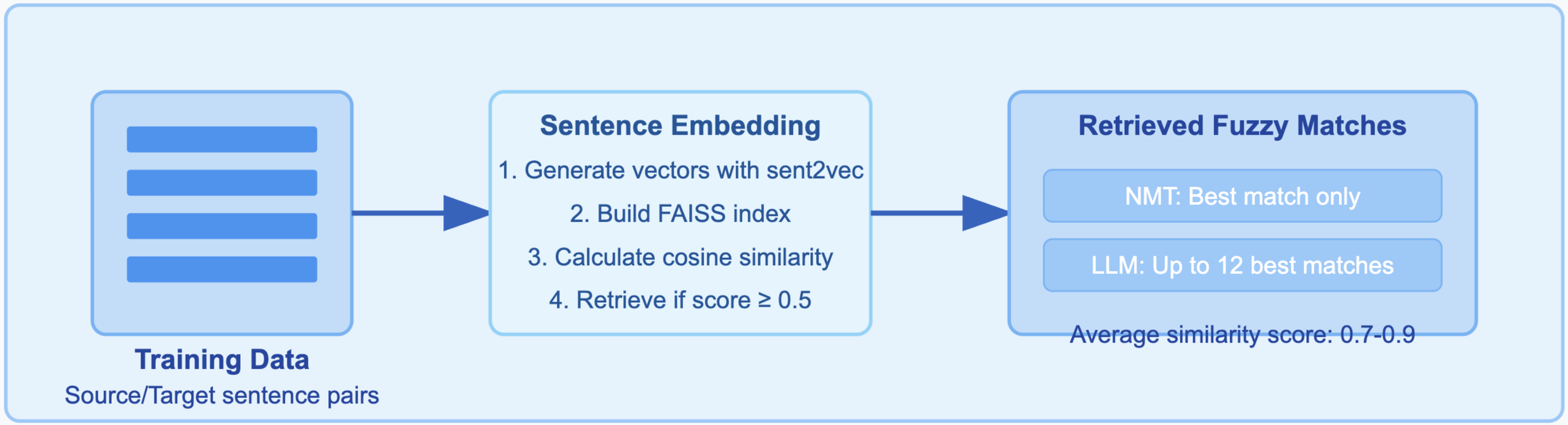
Neuroscience-filtered
5.1M pairs (23.1% of original)

Climatology-filtered
10.2M pairs (46.6% of original)

Mobility-filtered
8.8M pairs (40.4% of original)

Each configuration was further augmented with Fuzzy Match techniques for both NMT and LLM experiments

METHOD 2 – FUZZY MATCH AUGMENTATION



GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

METHOD 2 – FUZZY MATCH AUGMENTATION

During Training

Step 1: Original Sentence Pair

Source (S): We found three studies for inclusion in the review.

Target (T): Nous avons trouvé trois études pour l'inclusion dans la revue.

Step 2: Retrieved Fuzzy Match (Similarity Score: 0.9309)

Fuzzy Match Source (FMS): We identified **nine** eligible studies for inclusion in the review.

Fuzzy Match Target (FMT): Nous avons identifié **neuf** études éligibles pour l'inclusion dans la revue.

Differences highlighted in bold

Step 3: Creating Augmented Input (S*)

Formula: S* = Source **<sep>** FMT

Actual S*: We found three studies for inclusion in the review.
<sep> Nous avons identifié neuf études éligibles pour l'inclusion dans la revue.

Same process applied at inference time - source is augmented with FMT before translation

GHENT
UNIVER

METHOD 2 – FUZZY MATCH AUGMENTATION

Using fuzzy matches as few-shot examples in the prompt

Example 1 (Fuzzy Match)

Translate the source text from English to French.

Source: **We identified nine eligible studies for inclusion in the review.**

Target: **Nous avons identifié neuf études éligibles pour l'inclusion dans la revue.**

Up to 12 fuzzy match examples...

Actual Translation Request

Translate the source text from English to French.

Source: **We found three studies for inclusion in the review.**

Target:

Benefits: Improves terminology consistency, adapts to domain-specific patterns

BRINGING IT ALL TOGETHER

TF Method 1: Topic Filtering

Large Multi-domain Corpora

- ParaCrawl
- EuroPat
- SciPar

FastText Classifier

Domain-Specific Filtered Data

- Neuroscience (23.1%)
- Climatology (46.6%)
- Mobility (40.4%)

Key Points:

- Reduces data size while increasing relevance
- Uses classifiers trained on domain data
- Keeps only sentences related to domain

Data Configurations

1d | 3d | 3d+Ext | 3d+ExtTF

FM Method 2: Fuzzy Match Augmentation

Input Sentence

We found three studies for inclusion in the review.

Need: Similar sentence

Fuzzy Match Retrieval

Retrieved Fuzzy Match

We identified nine studies for inclusion in the review.

Similarity score: 0.93

Application by System Type:

NMT: Augment source with FM translation

LLM: Use FM as few-shot examples (up to 12)

RESULTS – NMT

NMT Model	Neuroscience			Climatology			Mobility		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
1d	39.11	65.15	79.28	29.57	57.99	76.05	30.14	59.45	76.98
3d	43.11	68.40	83.06	35.24	62.62	81.01	33.96	62.32	81.73

* $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$

Key Findings for NMT Models

- Adding more data helps: 3d > 1d, 3d+Ext > 3d

RESULTS – NMT

NMT Model	Neuroscience			Climatology			Mobility		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
3d	43.11	68.40	83.06	35.24	62.62	81.01	33.96	62.32	81.73
3d+Ext	44.40	69.42	84.70	35.70	63.35	82.48	36.11	64.01	84.88

* $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$

Key Findings for NMT Models

- Adding more data helps: 3d > 1d, 3d+Ext > 3d

RESULTS – NMT

NMT Model	Neuroscience			Climatology			Mobility		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
3d+Ext	44.40	69.42	84.70	35.70	63.35	82.48	36.11	64.01	84.88
3d+ExtTF	44.99	69.75	84.73	36.28	63.76	82.54	36.89	64.49	84.96

* $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$

Key Findings for NMT Models

- Adding more data helps: 3d > 1d, 3d+Ext > 3d
- Topic Filtering is effective (3d+ExtTF improves over 3d+Ext in all cases)

RESULTS – NMT

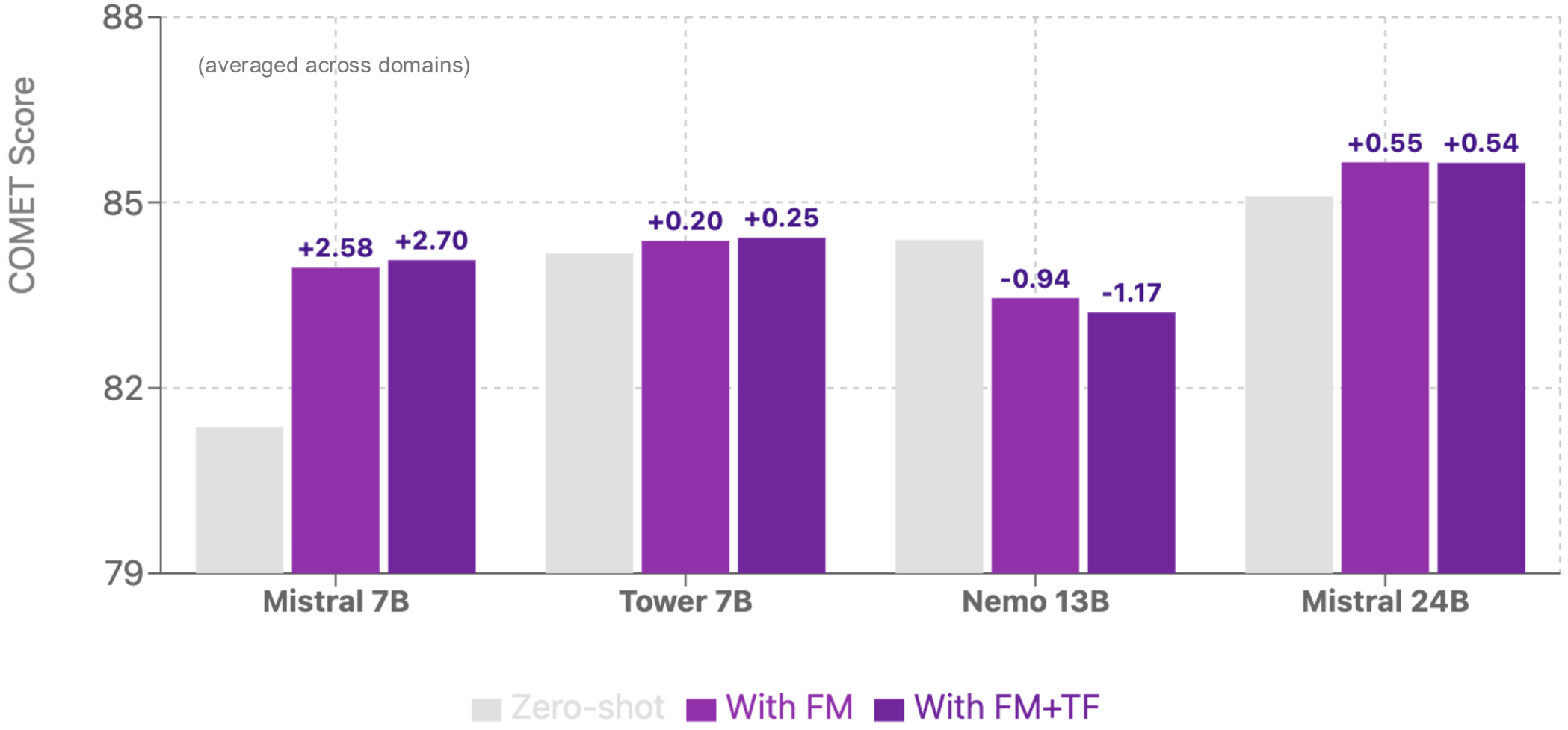
NMT Model	Neuroscience			Climatology			Mobility		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
1d	39.11	65.15	79.28	29.57	57.99	76.05	30.14	59.45	76.98
3d	43.11	68.40	83.06	35.24	62.62	81.01	33.96	62.32	81.73
3d+Ext	44.40	69.42	84.70	35.70	63.35	82.48	36.11	64.01	84.88
3d+ExtTF	44.99	69.75	84.73	36.28	63.76	82.54	36.89	64.49	84.96
3d+ExtTF_FM	**46.33‡**	**70.58‡**	**85.30†**	**36.97†**	**64.00†**	**82.82**	**37.68‡**	**64.81***	**85.27‡**

* p < 0.05, † p < 0.01, ‡ p < 0.001

Key Findings for NMT Models

- Adding more data helps: 3d > 1d, 3d+Ext > 3d
- Topic Filtering is effective (3d+ExtTF improves over 3d+Ext in most cases
- FM Augmentation consistently improves results
- **Best NMT: 3d+ExtTF_FM significantly outperforms all other systems**

RESULTS – LLMs



GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

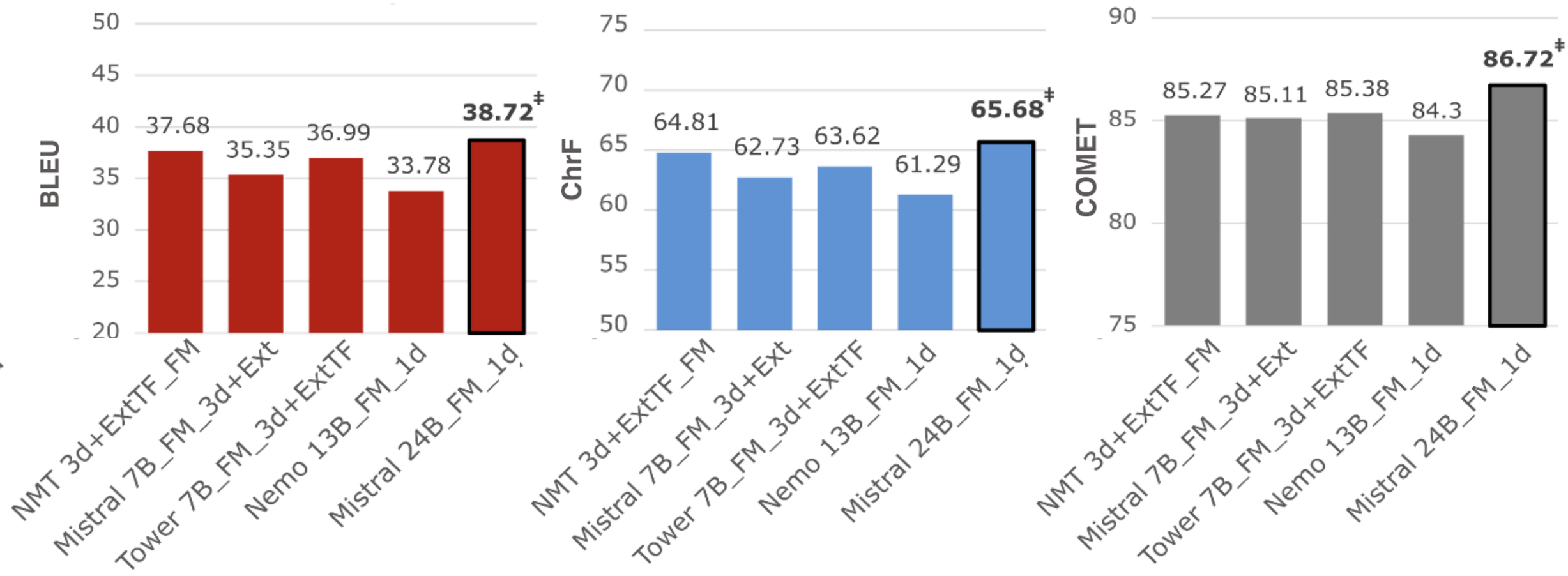
RESULTS – NMT vs. LLM COMPARISON

GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

RESULTS – NMT vs. LLM COMPARISON

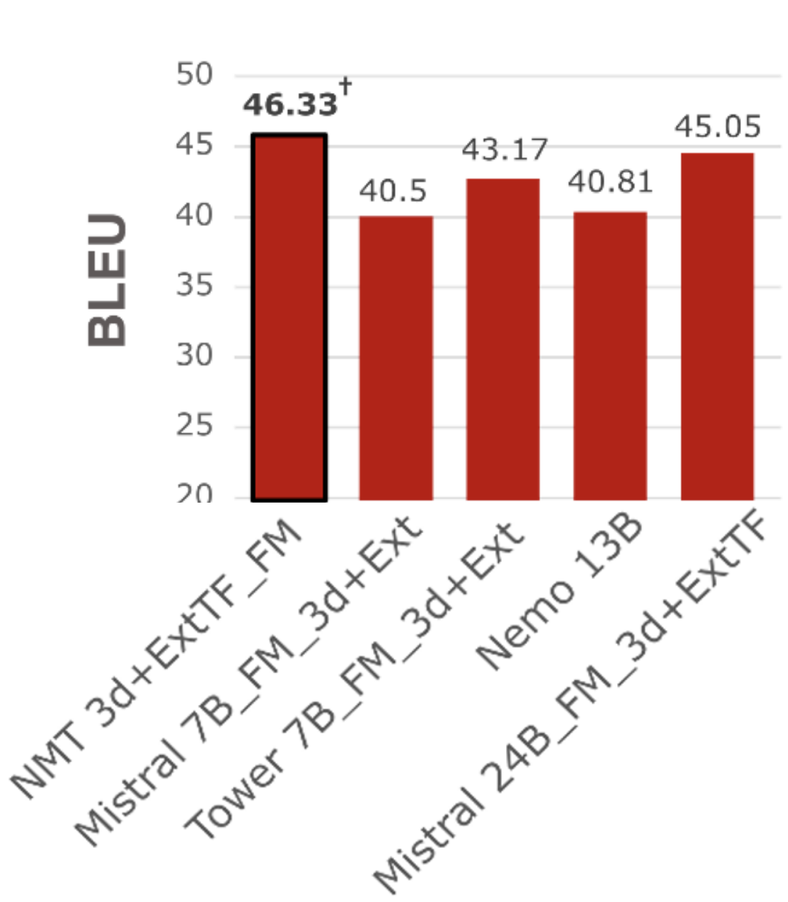
– **Mobility**

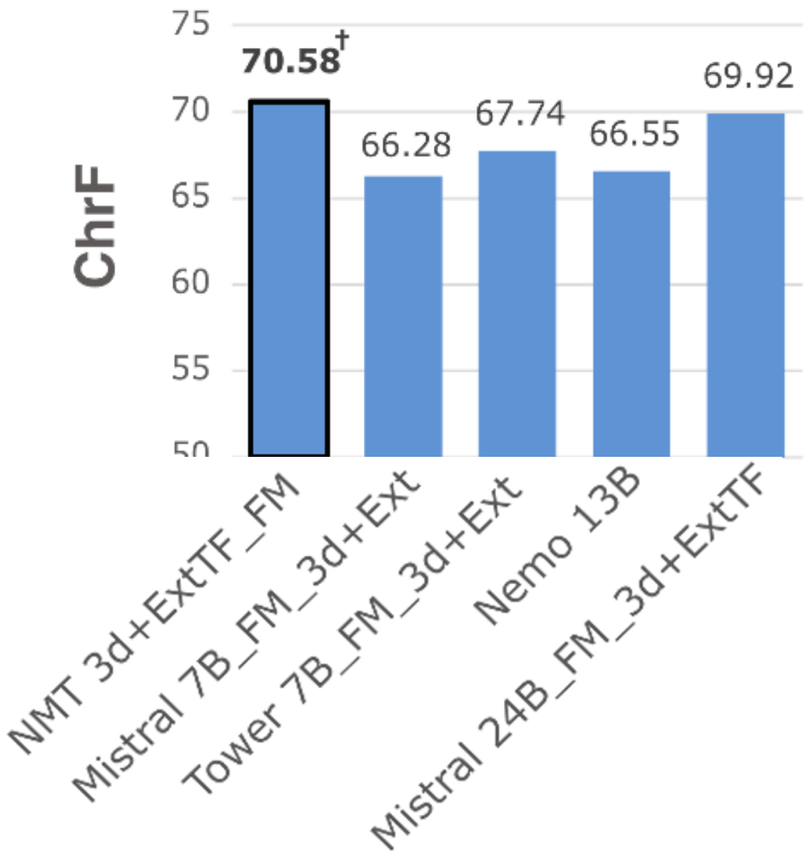


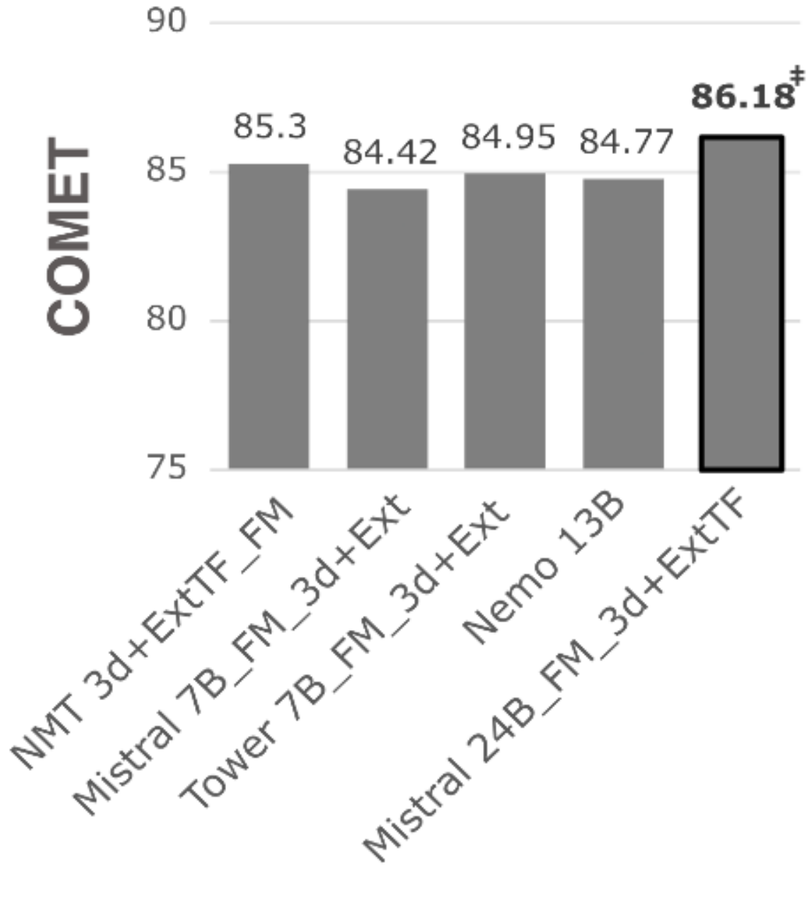
GHENT UNIVERSITY

RESULTS – NMT vs. LLM COMPARISON

– **Neuroscience**







CONCLUSION & DISCUSSION

- **Main Findings**
  - Combining Topic Filtering and FM Augmentation is effective for scientific domain adaptation
  - Specialized NMT models remain competitive with many off-the-shelf LLMs

CONCLUSION & DISCUSSION

- **LLM Findings**
  - FM augmentation via in-context learning is highly effective for LLMs
  - Small, domain-specific data for FM retrieval can be as effective as larger datasets
  - Larger LLMs generally perform better

LIMITATIONS

- Only English>French language pair tested
- Limited to three scientific domains
- Only automatic metrics used (no human evaluation)
- Did not explore fine-tuning LLMs with these methods
- Did not compare any other data selection methods

FUTURE WORK

- Fine-tuning LLMs in combination with in-context learning + how does it compare to NMT
- Combining back-translation with topic filtering+FM-augmentation

GHENT UNIVERSITY

CROSSLANG
TRANSLATION AUTOMATION

FACULTY OF ARTS AND PHILOSOPHY

Thomas Moerman

PhD Researcher

LANGUAGE AND TRANSLATION TECHNOLOGY TEAM (LT3)

✉ Thomas.Moerman@UGent.be

www.ugent.be

https://lt3.ugent.be/

GHENT UNIVERSITY

language and translation technology team

CROSSLANG TRANSLATION AUTOMATION

fwo