Fuzzy matches and synthetic data for machine translation

Thomas Moerman

An overview of our work at LT3 with Arda Tezcan and the plan for this research stay

Supervised by Arda Tezcan and Els Lefever, LT3, Ghent University

Transducens, Universitat d'Alacant, October 2026

GHENT UNIVERSITY

language and translation technology team

TransTech

PART 1

Our work at LT3

Problem

A small model gets a legal term wrong until it sees a similar example

REQUEST, TRANSLATE INTO FRENCH

The applicant may **lodge an appeal** against the decision within fourteen days.

Small model
no domain training

Problem

A small model gets a legal term wrong until it sees a similar example

REQUEST, TRANSLATE INTO FRENCH

The applicant may **lodge an appeal** against the decision within fourteen days.

→

Small model
no domain training

→

OUTPUT, NO EXAMPLE

Le **demandeur** peut **déposer un appel** contre la décision dans un délai de quatorze jours.

demandeur and déposer un appel are not the legal terms

Problem

A small model gets a legal term wrong until it sees a similar example

REQUEST, TRANSLATE INTO FRENCH

The applicant may **lodge an appeal** against the decision within fourteen days.

A SIMILAR EXAMPLE ADDED TO THE PROMPT

The applicant may lodge an appeal within one month.
Le requérant peut interjeter appel dans un délai d'un mois.

→

Small model
no domain training

→

OUTPUT, NO EXAMPLE

Le demandeur peut déposer un appel contre la décision dans un délai de quatorze jours.

demandeur and déposer un appel are not the legal terms

Problem

A small model gets a legal term wrong until it sees a similar example

REQUEST, TRANSLATE INTO FRENCH

The applicant may **lodge an appeal** against the decision within fourteen days.

A SIMILAR EXAMPLE ADDED TO THE PROMPT

The applicant may lodge an appeal within one month.
Le requérant peut interjeter appel dans un délai d'un mois.

→

Small model
no domain training

→

OUTPUT, NO EXAMPLE

Le demandeur peut déposer un appel contre la décision dans un délai de quatorze jours.

demandeur and déposer un appel are not the legal terms

OUTPUT WITH THE EXAMPLE

Le requérant peut interjeter appel contre la décision dans un délai de quatorze jours.

requérant and interjeter appel are copied from the example

Problem

A small model gets a legal term wrong until it sees a similar example

REQUEST, TRANSLATE INTO FRENCH

The applicant may **lodge an appeal** against the decision within fourteen days.

A SIMILAR EXAMPLE ADDED TO THE PROMPT

The applicant may lodge an appeal within one month.
Le requérant peut interjeter appel dans un délai d'un mois.

→ **Small model** no domain training →

OUTPUT, NO EXAMPLE

Le demandeur peut déposer un appel contre la décision dans un délai de quatorze jours.

demandeur and déposer un appel are not the legal terms

OUTPUT WITH THE EXAMPLE

Le requérant peut interjeter appel contre la décision dans un délai de quatorze jours.

requérant and interjeter appel are copied from the example

The problem shifts from translating to finding and adding relevant examples

METHOD 1 OF 2

Fuzzy-match augmentation

The model receives a similar translation to repair

DATA SCENARIO WITH PLENTY OF IN-DOMAIN BILINGUAL PAIRS THAT FORM THE MEMORY

SOURCE SENTENCE (EN)

The applicant may **lodge an appeal** against the decision within fourteen days.

→

MODEL INPUT

The applicant may lodge an appeal against the decision within fourteen days.

→ **LLM or NMT** small → **Translation (FR)**

[Bulte & Tezcan 2019]

METHOD 1 OF 2

Fuzzy-match augmentation

The model receives a similar translation to repair

DATA SCENARIO WITH PLENTY OF IN-DOMAIN BILINGUAL PAIRS THAT FORM THE MEMORY

SOURCE SENTENCE (EN)

The applicant may **lodge an appeal** against the decision within fourteen days.

TRANSLATION MEMORY, SOURCE AND TARGET PAIRS

- The applicant may lodge an appeal within one month.
- You may appeal this decision in writing.
- Please bring your documents.

→

MODEL INPUT

The applicant may lodge an appeal against the decision within fourteen days.

→ **LLM or NMT** small → **Translation (FR)**

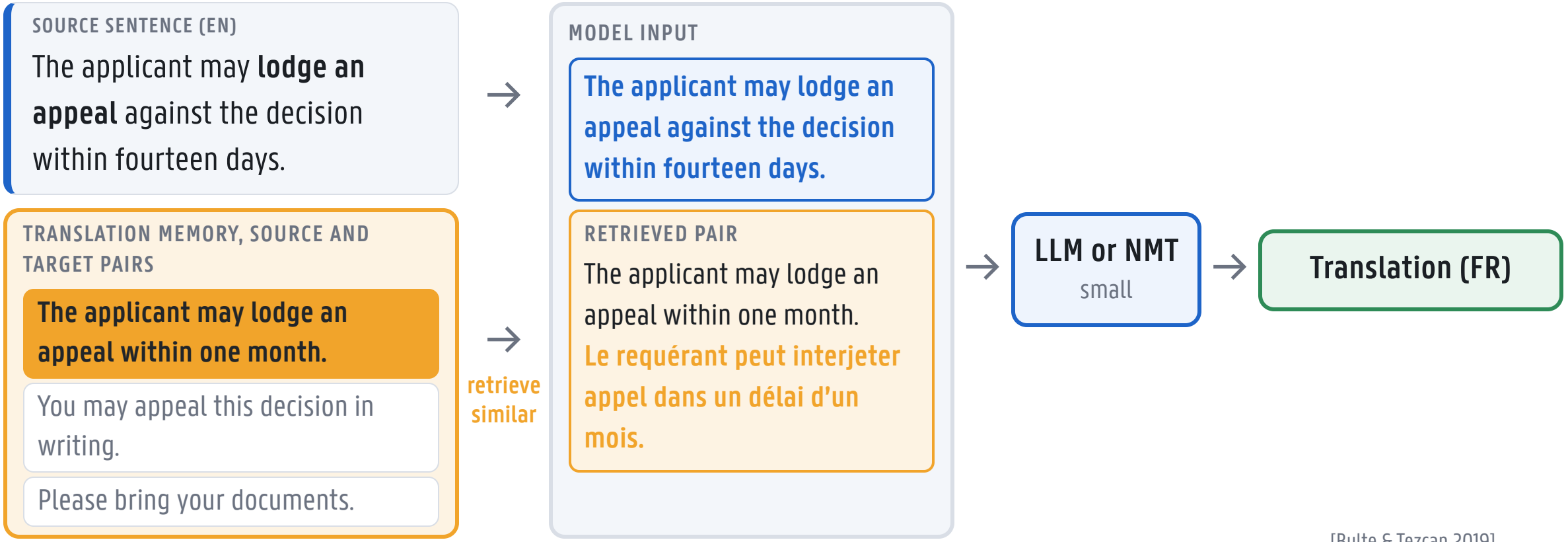
[Bulte & Tezcan 2019]

METHOD 1 OF 2

Fuzzy-match augmentation

The model receives a similar translation to repair

DATA SCENARIO WITH PLENTY OF IN-DOMAIN BILINGUAL PAIRS THAT FORM THE MEMORY



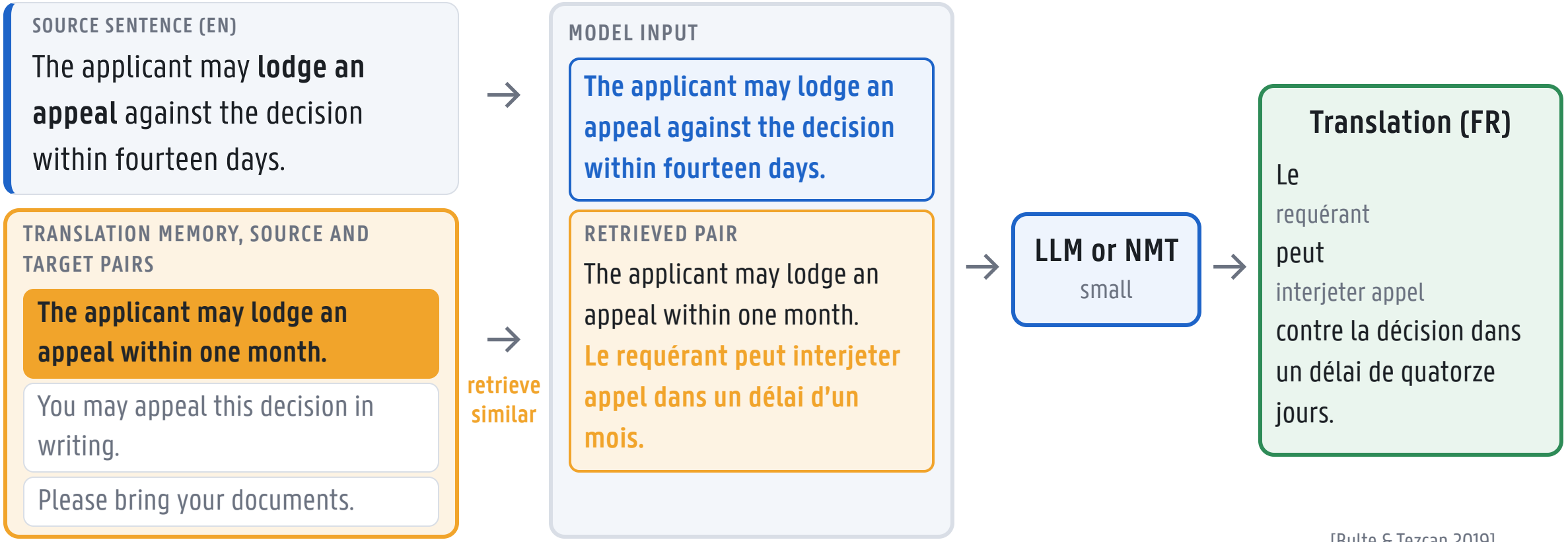
[Bulte & Tezcan 2019]

METHOD 1 OF 2

Fuzzy-match augmentation

The model receives a similar translation to repair

DATA SCENARIO WITH PLENTY OF IN-DOMAIN BILINGUAL PAIRS THAT FORM THE MEMORY



[Bulte & Tezcan 2019]

METHOD 1 OF 2

Fuzzy-match augmentation in NMT and in LLMs

NMT appends the match to the source and an LLM prompts with it

NMT, NEURAL FUZZY REPAIR, ONE AUGMENTED INPUT

The applicant may lodge an appeal against the decision within fourteen days. @@@ **Le requérant peut interjeter appel dans un délai d'un mois.**

The best match's translation after a separator. Trained on original plus augmented pairs. It also works without a match.

One match and target side only.

[Bulte & Tezcan 2019]

[Moerman, Lefever & Tezcan, under review]

METHOD 1 OF 2

Fuzzy-match augmentation in NMT and in LLMs

NMT appends the match to the source and an LLM prompts with it

NMT, NEURAL FUZZY REPAIR, ONE AUGMENTED INPUT

The applicant may lodge an appeal against the decision within fourteen days. @@@ **Le requérant peut interjeter appel dans un délai d'un mois.**

The best match's translation after a separator. Trained on original plus augmented pairs. It also works without a match. One match and target side only.

LLM, MATCHES AS EXAMPLES IN THE PROMPT

Translate from English to French.
Source: **The applicant may lodge an appeal within one month.**
Target: **Le requérant peut interjeter appel dans un délai d'un mois.**
… up to k retrieved pairs, most similar first
Source: **The applicant may lodge an appeal against the decision within fourteen days.**
Target:

[Bulte & Tezcan 2019]
[Moerman, Lefever & Tezcan, under review]

METHOD 1 OF 2

Fuzzy-match augmentation in NMT and in LLMs

NMT appends the match to the source and an LLM prompts with it

NMT, NEURAL FUZZY REPAIR, ONE AUGMENTED INPUT

The applicant may lodge an appeal against the decision within fourteen days. @@@ **Le requérant peut interjeter appel dans un délai d'un mois.**

The best match's translation after a separator. Trained on original plus augmented pairs. It also works without a match. One match and target side only.

LLM, MATCHES AS EXAMPLES IN THE PROMPT

Translate from English to French.
Source: **The applicant may lodge an appeal within one month.**
Target: **Le requérant peut interjeter appel dans un délai d'un mois.**
… up to k retrieved pairs, most similar first
Source: **The applicant may lodge an appeal against the decision within fourteen days.**
Target:

THE LLM SEES THE MATCHES AT INFERENCE OR IN FINE-TUNING OR BOTH

Without fine-tuning many examples keep helping

After fine-tuning 1 to 5 examples are usually enough

[Bulte & Tezcan 2019]
[Moerman, Lefever & Tezcan, under review]

METHOD 1 OF 2

Retrieval

We retrieve by embedding similarity and other measures fit other cases

MEASURE	WHAT IT COMPARES	WHEN IT FITS
Cosine over sentence embeddings	meaning. Paraphrases and reorderings match	our default in most settings with a FAISS index and a threshold of 0.5
Cross-lingual embeddings, LaBSE or SONAR	meaning across languages	retrieving from a monolingual target-language corpus
BM25	word overlap. Rare words count more	very large pools. Speed and no training
Edit distance, Levenshtein	surface form. The CAT-tool match percentage	near-duplicate memories. An interpretable score
Character-set Jaccard	characters shared	when no encoder covers the language

[Esplà-Gomis et al. 2022]
[Moerman & Tezcan 2026]

METHOD 1 OF 2

Retrieval illustrated

Retrieval scores the memory and hands the model the closest pair

QUERY, EN The applicant may lodge an appeal against the decision within fourteen days.

MEMORY, SOURCE (EN)	TARGET (FR)	COSINE SIMILARITY	
The applicant may lodge an appeal within one month.	Le requérant peut interjeter appel dans un délai d'un mois.		?
You may appeal this decision in writing.	Vous pouvez contester cette décision par écrit.		?
Please bring your documents to the appointment.	Veuillez apporter vos documents au rendez-vous.		?

[Tezcan, Skidanova & Moerman 2024]
[Moerman, Tezcan & Macken 2026]

METHOD 1 OF 2

Retrieval illustrated

Retrieval scores the memory and hands the model the closest pair

QUERY, EN The applicant may lodge an appeal against the decision within fourteen days.

MEMORY, SOURCE (EN)	TARGET (FR)	COSINE SIMILARITY
The applicant may lodge an appeal within one month.	Le requérant peut interjeter appel dans un délai d'un mois.	0.86
You may appeal this decision in writing.	Vous pouvez contester cette décision par écrit.	0.63
Please bring your documents to the appointment.	Veuillez apporter vos documents au rendez-vous.	0.19

[Tezcan, Skidanova & Moerman 2024]
[Moerman, Tezcan & Macken 2026]

METHOD 1 OF 2

Retrieval illustrated

Retrieval scores the memory and hands the model the closest pair

QUERY, EN The applicant may lodge an appeal against the decision within fourteen days.

MEMORY, SOURCE (EN)	TARGET (FR)	COSINE SIMILARITY
The applicant may lodge an appeal within one month.	Le requérant peut interjeter appel dans un délai d'un mois.	0.86
You may appeal this decision in writing.	Vous pouvez contester cette décision par écrit.	0.63
Please bring your documents to the appointment.	Veuillez apporter vos documents au rendez-vous.	0.19

The most similar pair becomes the fuzzy-match example

Cosine over MiniLM embeddings with matches above 0.5

[Tezcan, Skidanova & Moerman 2024]
[Moerman, Tezcan & Macken 2026]

METHOD 1 OF 2

Retrieval illustrated

Retrieval scores the memory and hands the model the closest pair

QUERY, EN The applicant may lodge an appeal against the decision within fourteen days.

MEMORY, SOURCE (EN)	TARGET (FR)	COSINE SIMILARITY
The applicant may lodge an appeal within one month.	Le requérant peut interjeter appel dans un délai d'un mois.	0.86
You may appeal this decision in writing.	Vous pouvez contester cette décision par écrit.	0.63
Please bring your documents to the appointment.	Veuillez apporter vos documents au rendez-vous.	0.19

SOURCE SENTENCE the sentence to translate → **SENTENCE ENCODER** MiniLM, 384 dimensions → **FAISS INDEX** fast lookup of the closest vectors → **TOP K MATCHES** score ≥ 0.5, no exact copies → **MODEL INPUT** 1 target for NMT, k pairs for an LLM

[Tezcan, Skidanova & Moerman 2024]
[Moerman, Tezcan & Macken 2026]

METHOD 2 OF 2

Back-translation

Back-translation turns monolingual text into synthetic pairs

DATA SCENARIO

Bilingual pairs, EN and UK
limited to about 50k pairs

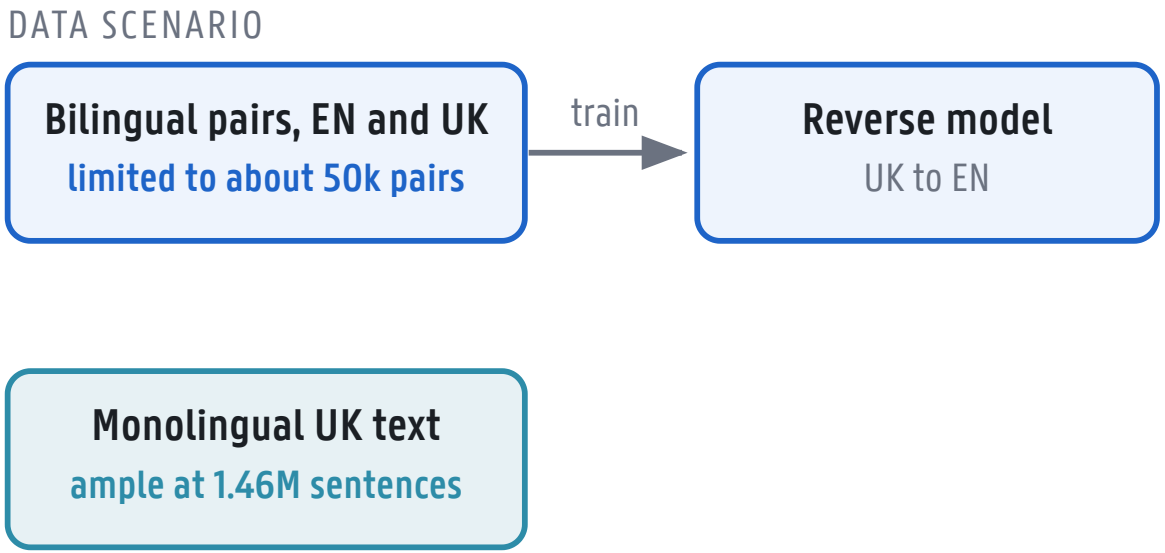
Monolingual UK text
ample at 1.46M sentences

[Sennrich et al. 2016]
[Tezcan, Skidanova & Moerman 2024]

METHOD 2 OF 2

Back-translation

Back-translation turns monolingual text into synthetic pairs



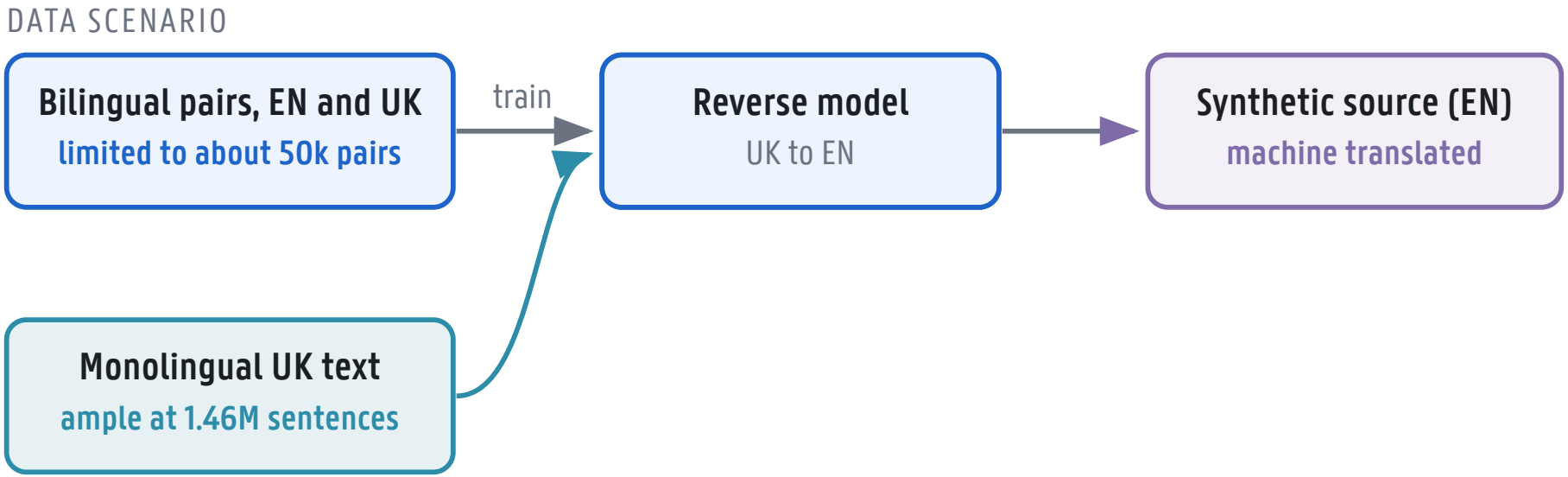
[Sennrich et al. 2016]

[Tezcan, Skidanova & Moerman 2024]

METHOD 2 OF 2

Back-translation

Back-translation turns monolingual text into synthetic pairs



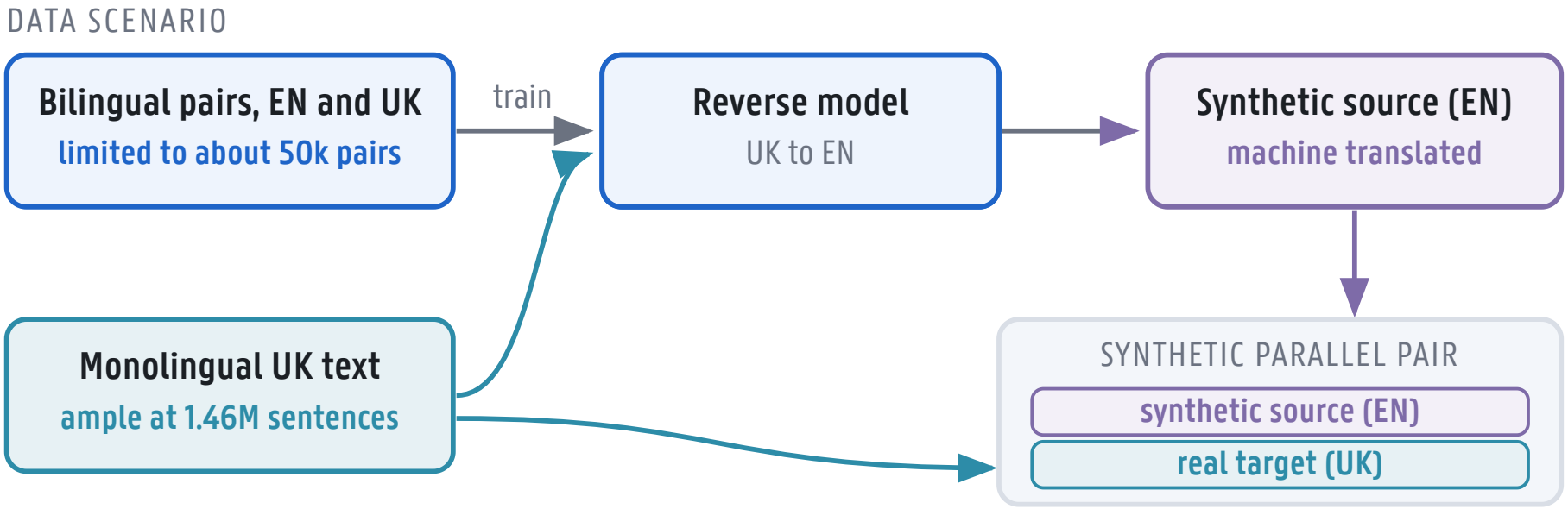
[Sennrich et al. 2016]

[Tezcan, Skidanova & Moerman 2024]

METHOD 2 OF 2

Back-translation

Back-translation turns monolingual text into synthetic pairs



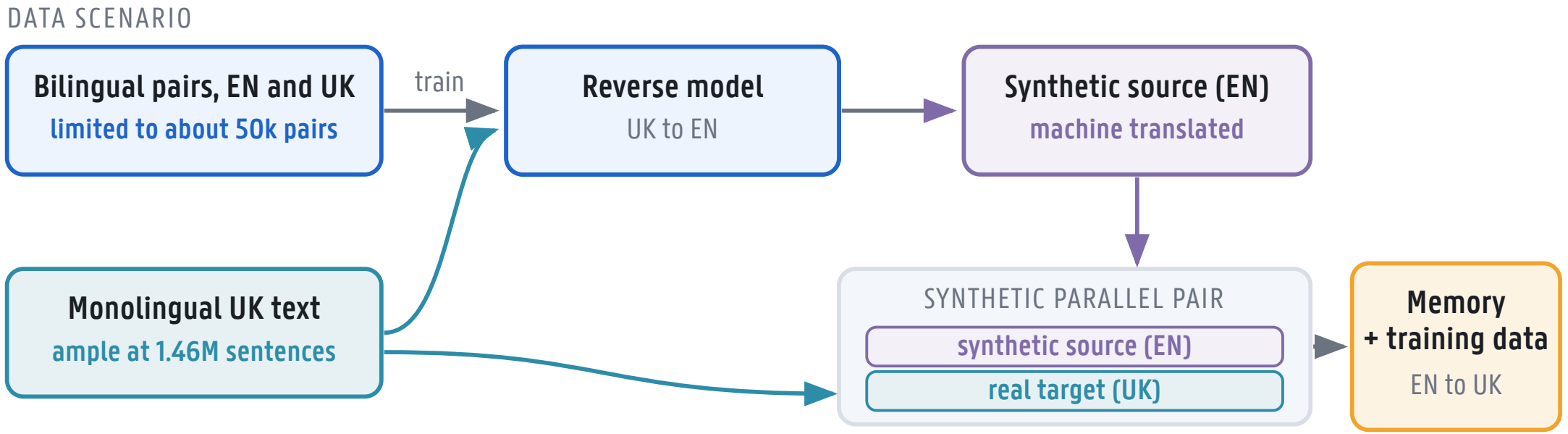
[Sennrich et al. 2016]

[Tezcan, Skidanova & Moerman 2024]

METHOD 2 OF 2

Back-translation

Back-translation turns monolingual text into synthetic pairs



The synthetic pairs feed both the retrieval memory and the training data

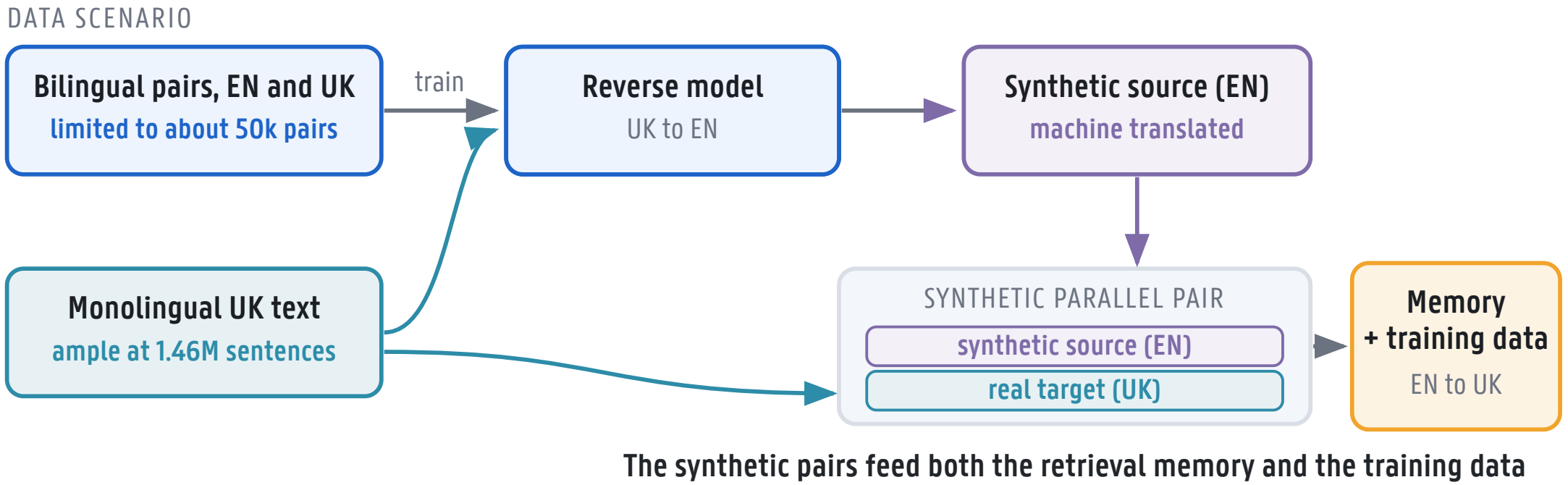
[Sennrich et al. 2016]

[Tezcan, Skidanova & Moerman 2024]

METHOD 2 OF 2

Back-translation

Back-translation turns monolingual text into synthetic pairs



The synthetic pairs feed both the retrieval memory and the training data

We found that 64 to 86 percent of the retrieved fuzzy matches were back-translated sentences

[Sennrich et al. 2016]

[Tezcan, Skidanova & Moerman 2024]

Setup

Three directions in two legal domains

LEGAL, EN TO UK

286,417 bilingual pairs and 1.46M monolingual Ukrainian sentences

EU LEGISLATION, DGT, EN TO FR AND FR TO EN

300,000 bilingual pairs and 1.5M monolingual sentences with their real translations held back

[Tezcan, Skidanova & Moerman 2024]

Setup

Three directions in two legal domains

LEGAL, EN TO UK

286,417 bilingual pairs and 1.46M monolingual Ukrainian sentences

EU LEGISLATION, DGT, EN TO FR AND FR TO EN

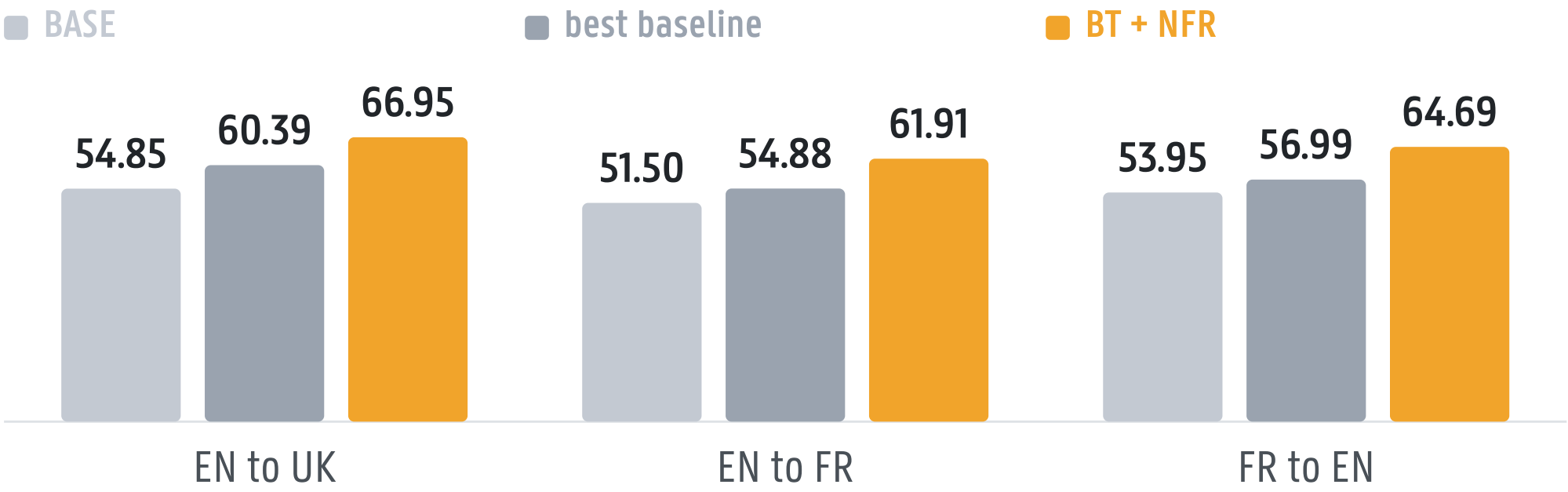
300,000 bilingual pairs and 1.5M monolingual sentences with their real translations held back

SYSTEM	BILINGUAL PAIRS	FUZZY MATCHES	BACK-TRANSLATED PAIRS	MATCHES FROM MONOLINGUAL TEXT	REAL TRANSLATIONS OF THE MONOLINGUAL TEXT AS A REFERENCE
BASE	●				
NFR	●	●			
NFR mono	●	●		●	
BT	●		●		
BT + NFR, ours	●	●	●		
BASE HQ	●				●
NFR HQ	●	●			●

[Tezcan, Skidanova & Moerman 2024]

Results

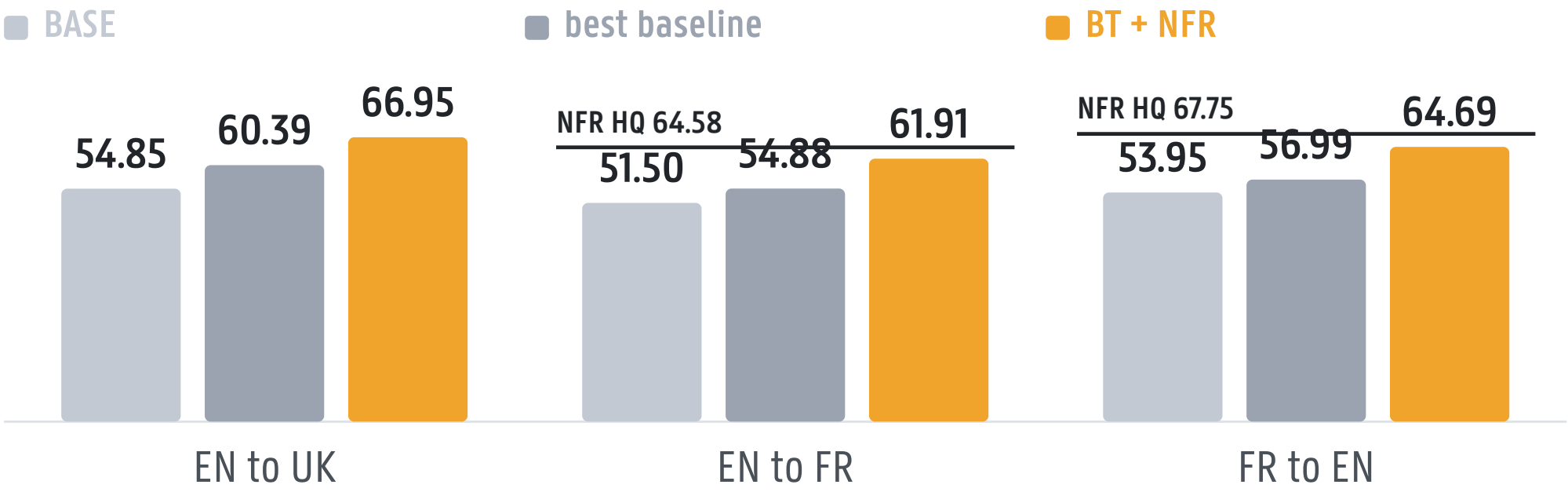
BT plus fuzzy matches add 6.6 to 7.7 BLEU over the best baseline



[Tezcan, Skidanova & Moerman 2024]

Results

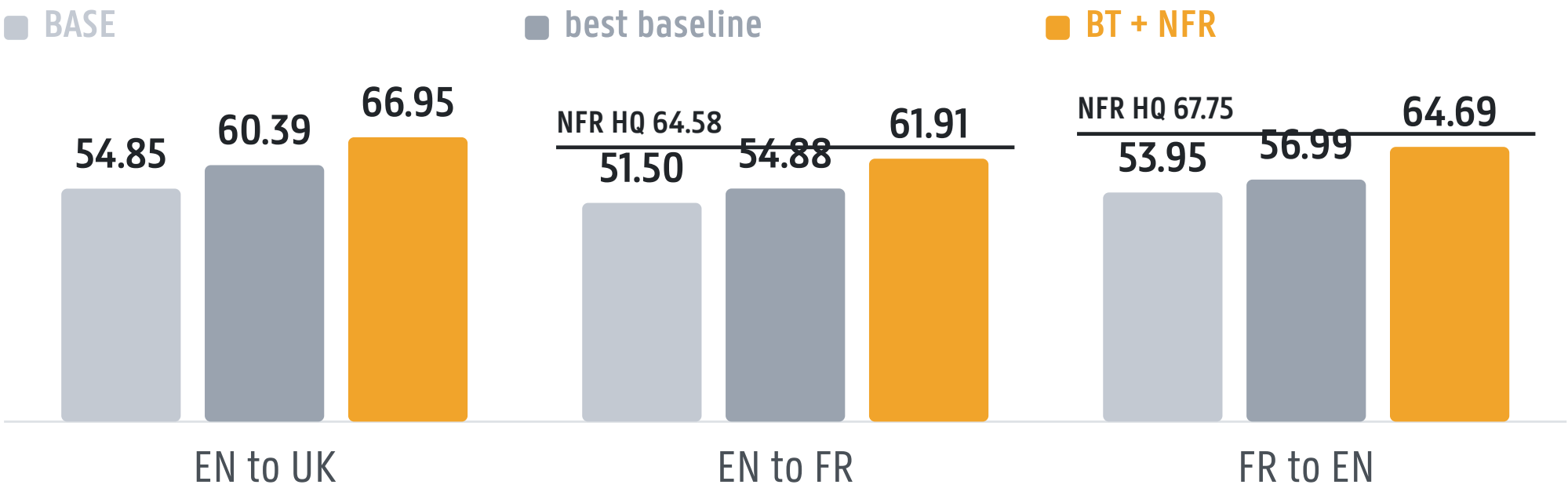
BT plus fuzzy matches add 6.6 to 7.7 BLEU over the best baseline



[Tezcan, Skidanova & Moerman 2024]

Results

BT plus fuzzy matches add 6.6 to 7.7 BLEU over the best baseline



Up to 12.10 BLEU over BASE at $p < 0.001$

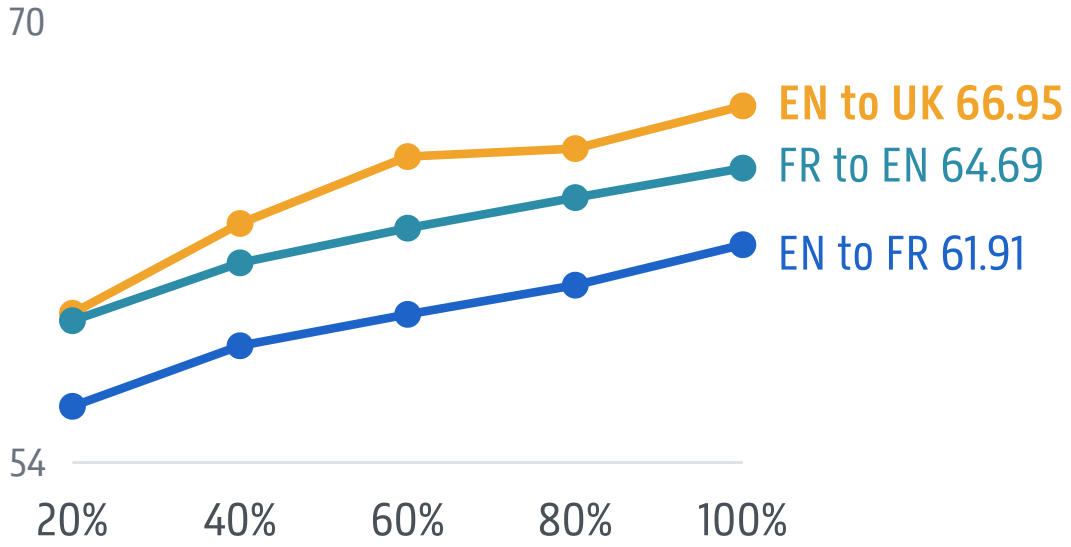
Ahead of GPT-4o and Llama 3.1 70B in every direction

[Tezcan, Skidanova & Moerman 2024]

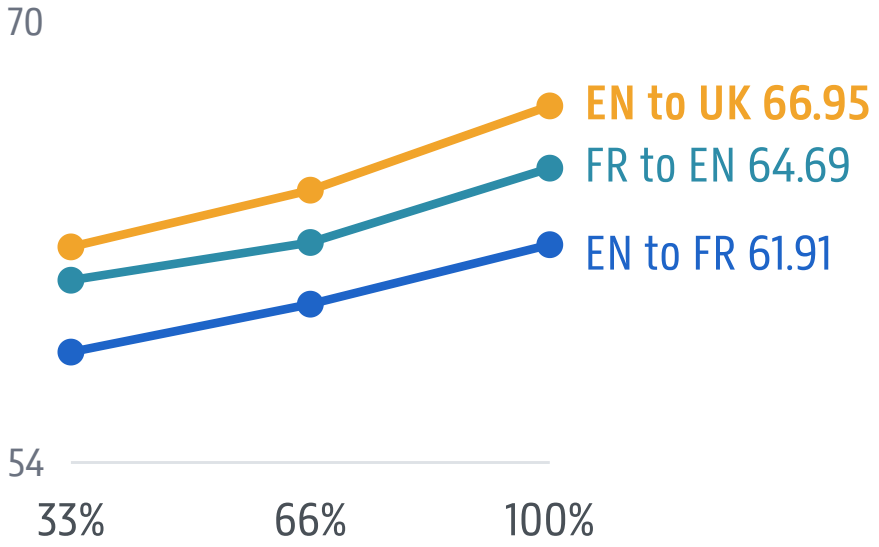
Reduced data

The gains hold with a third of the bilingual data

Less monolingual data, all bilingual pairs



Less bilingual data, all monolingual



BT + NFR BLEU

At a third of the bilingual data it still beats every baseline at $p < 0.001$

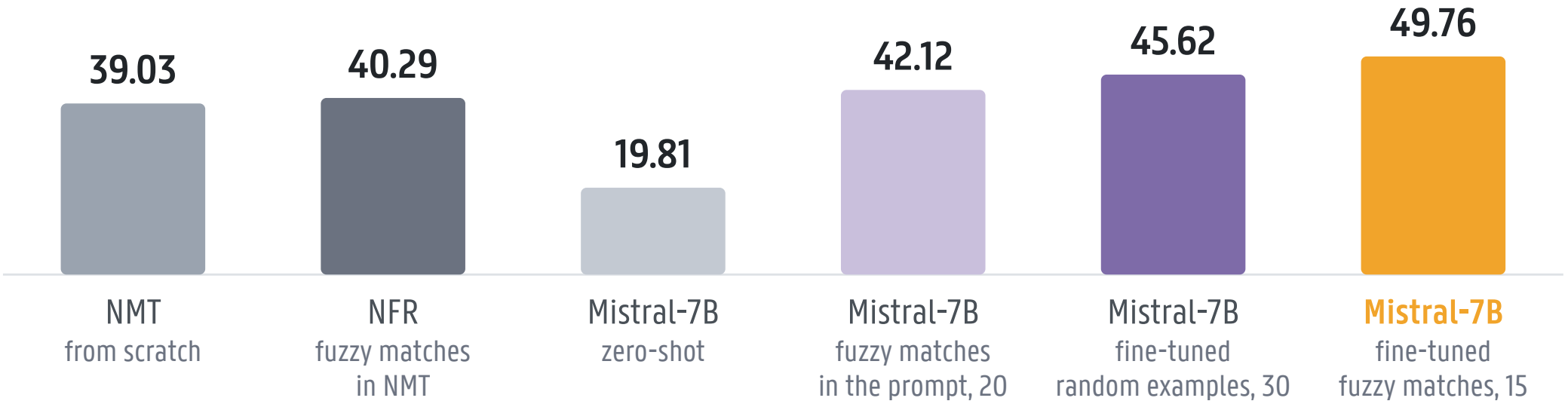
[Tezcan, Skidanova & Moerman 2024]

THE SAME RECIPE FOR LLMS, UNDER REVIEW

Configuration

Fine-tuning and in-context examples are complementary

BLEU, legal EN to UK, 50k bilingual pairs. The number is the examples in the prompt at inference



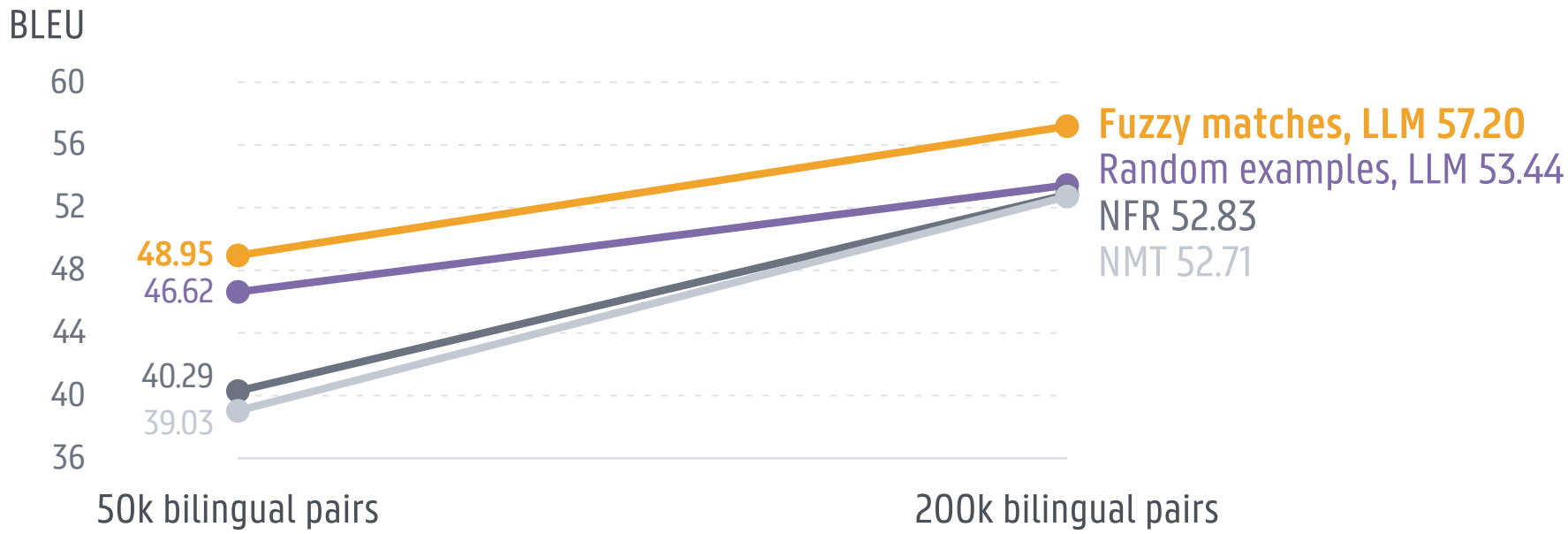
At 50k pairs the LLM beats NFR by up to 11.62 BLEU

Fuzzy matches do not yet help NMT at this size

[Moerman, Lefever & Tezcan, under review]

More bilingual data

NMT catches up on the 7B LLM as the bilingual data grows



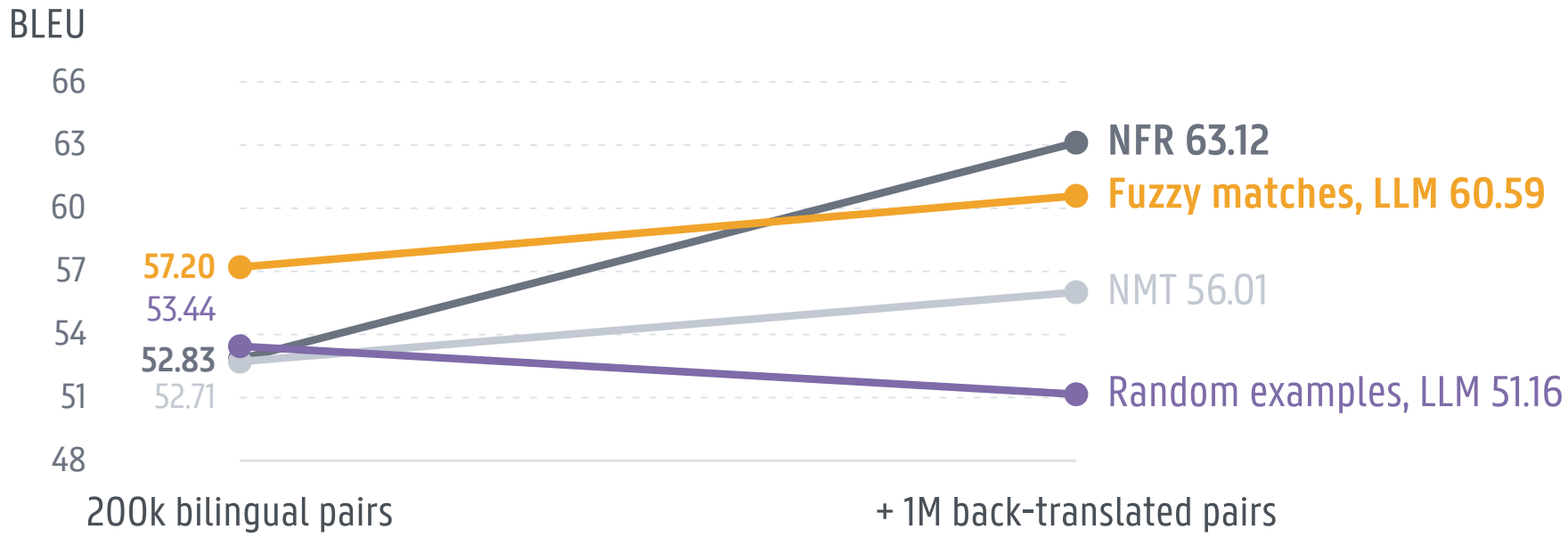
Legal EN to UK

The COMET advantage of the LLM shrinks from 10.27 at 50k to 4.17 at 100k and 2.67 at 200k

[Moerman, Lefever & Tezcan, under review]

Adding back-translation

NFR with back-translation overtakes the 7B LLM for Ukrainian



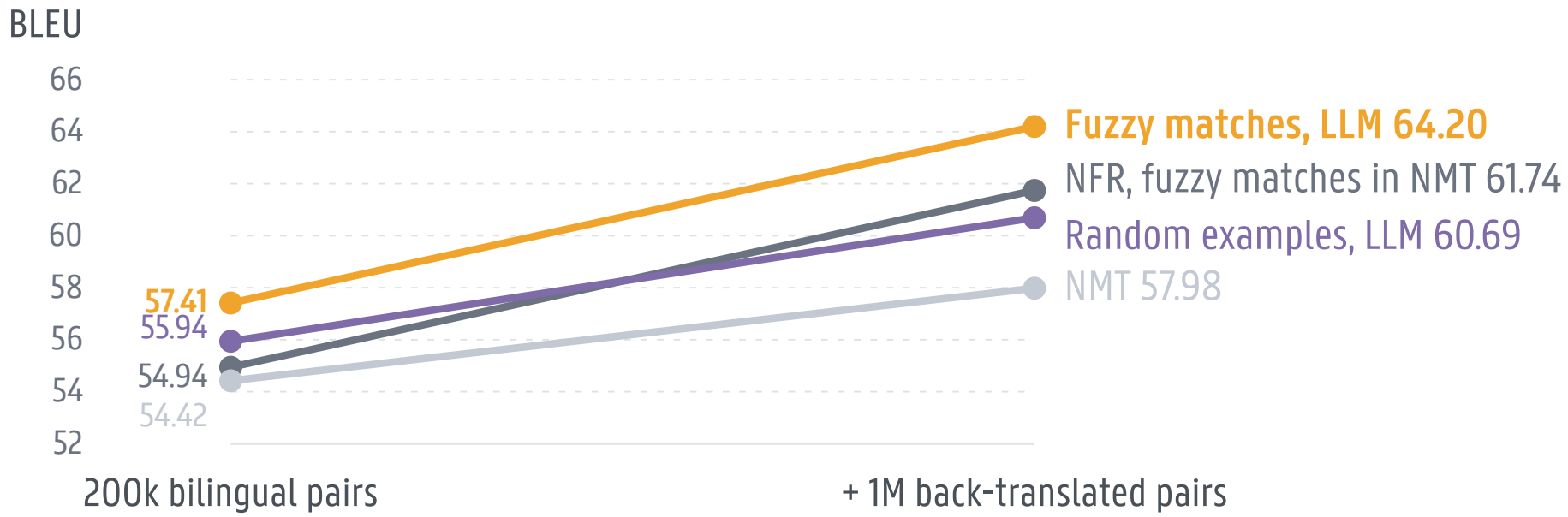
The COMET of the LLM stays at a ceiling of 92.7 to 93.0

Random examples degrade with synthetic data

[Moerman, Lefever & Tezcan, under review]

French directions

The LLM with back-translation beats NFR for French



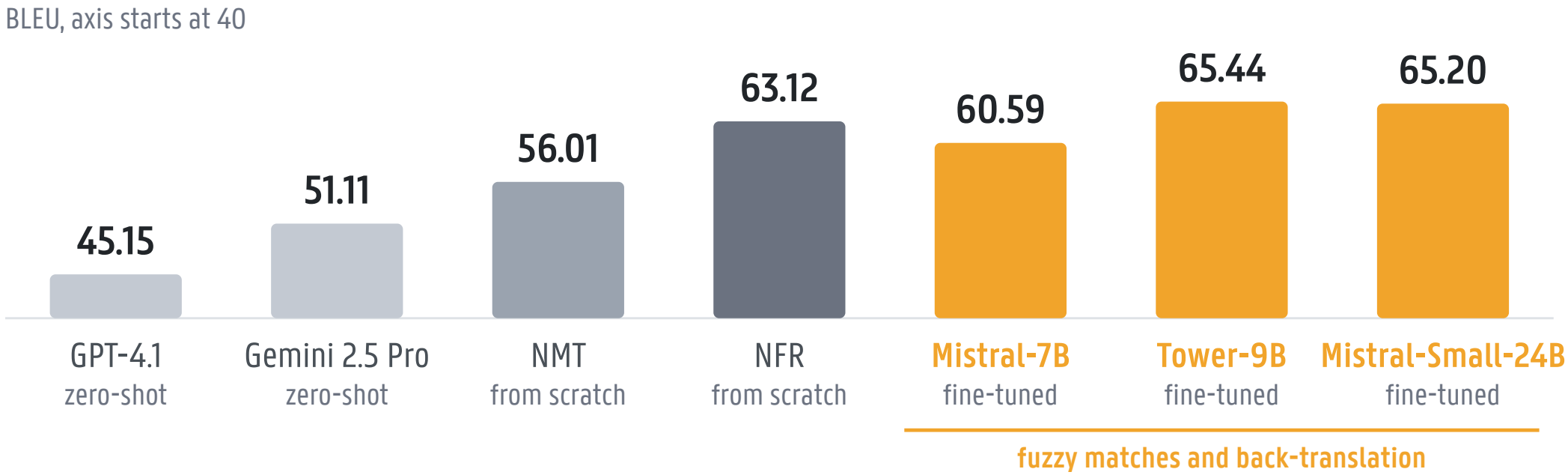
Patent EN to FR with Mistral-7B

In legal EN to FR and FR to EN the LLM with BT also ends ahead of NFR

[Moerman, Lefever & Tezcan, under review]

Model type and size

A 9B translation-specialised model matches a 24B general one



Legal EN to UK

The differences between Tower-9B and Mistral-Small-24B are not significant on any metric

[Moerman, Lefever & Tezcan, under review]

The MaTIAS project

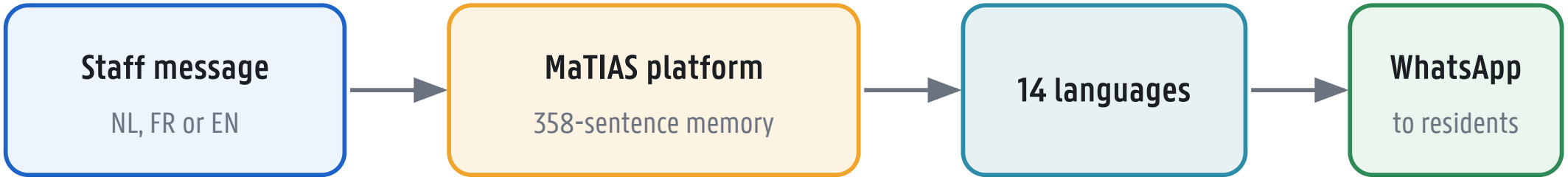
Machine translation to inform asylum seekers in Belgian reception centres

[Macken et al. 2025]

[Moerman, Tezcan & Macken 2026]

The MaTIAS platform

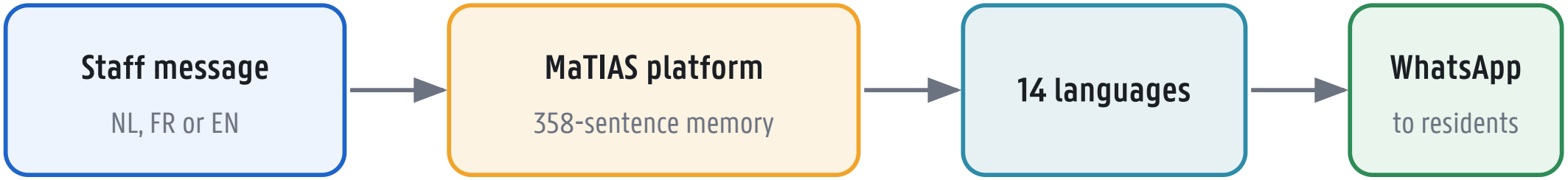
The platform translates reception messages into 14 languages



[Macken et al. 2025]

The MaTIAS platform

The platform translates reception messages into 14 languages



The content is medical and legal and personal. The translation has to stay data-sovereign.

Memory compiled from fieldwork in Belgian reception centres

[Macken et al. 2025]

Dataset and setup

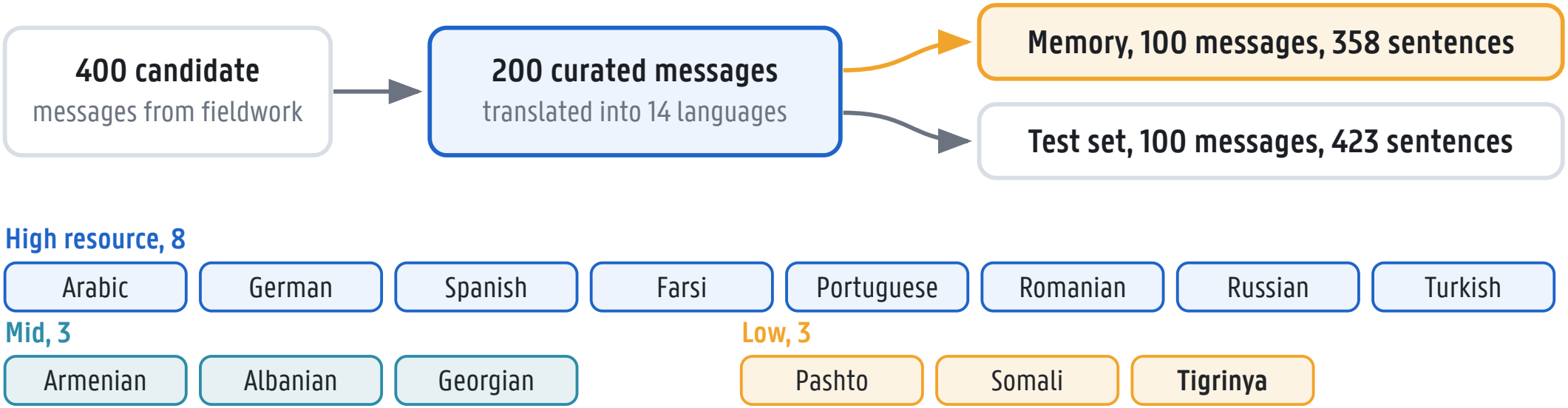
A 358-sentence memory and 14 languages in three tiers



[Moerman, Tezcan & Macken 2026]

Dataset and setup

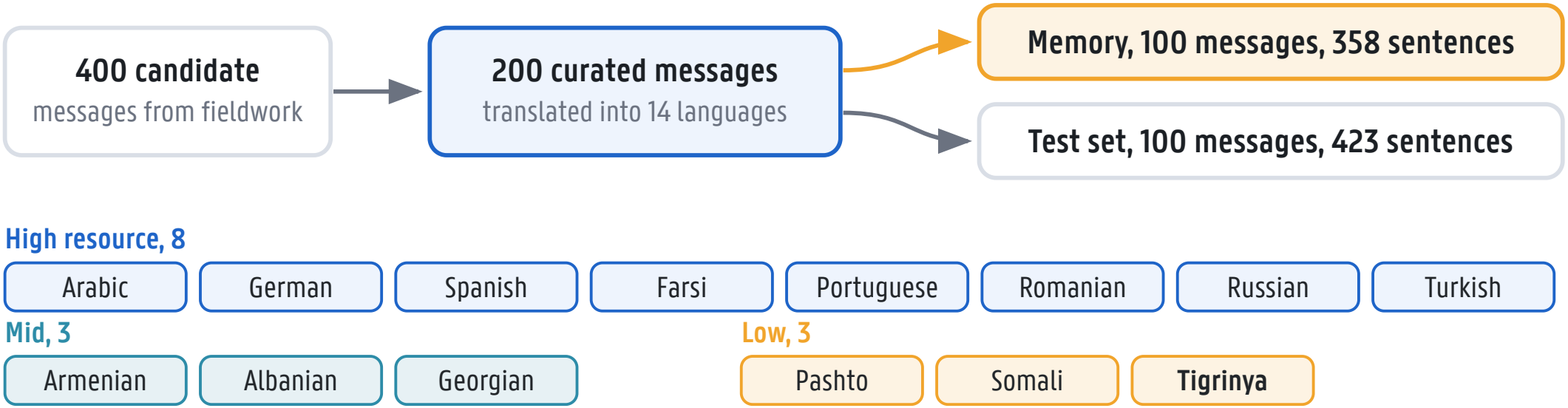
A 358-sentence memory and 14 languages in three tiers



[Moerman, Tezcan & Macken 2026]

Dataset and setup

A 358-sentence memory and 14 languages in three tiers



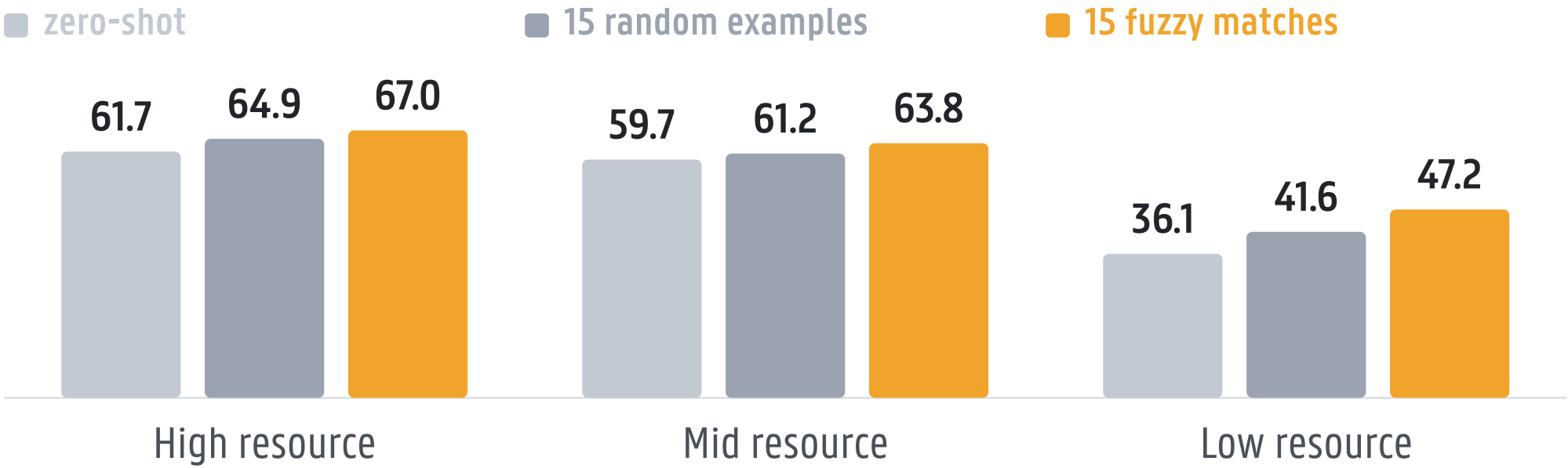
Three conditions per model. Zero-shot and 15 random examples and 15 fuzzy matches

Five systems. ModernMT with and without the memory, TranslateGemma, TowerPlus and Gemini 2.5 Pro

[Moerman, Tezcan & Macken 2026]

Fuzzy matches against random examples

Fuzzy matches beat random examples for 13 of 14 languages



TranslateGemma 27B with the 358-sentence memory

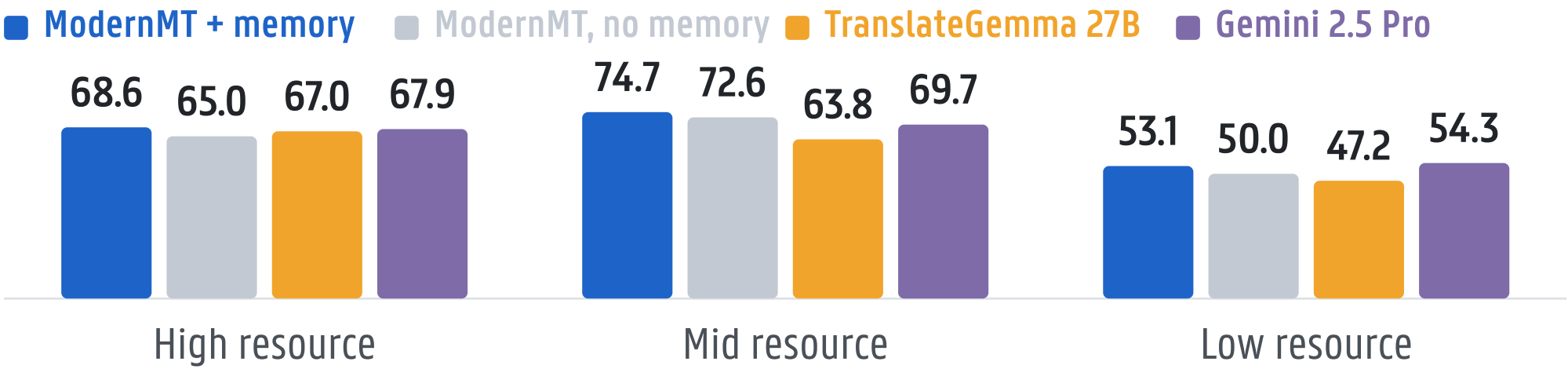
+6.29 chrF over zero-shot on average

Significant over random for 11 of 14

[Moerman, Tezcan & Macken 2026]

Adaptive NMT against retrieval-augmented LLMs

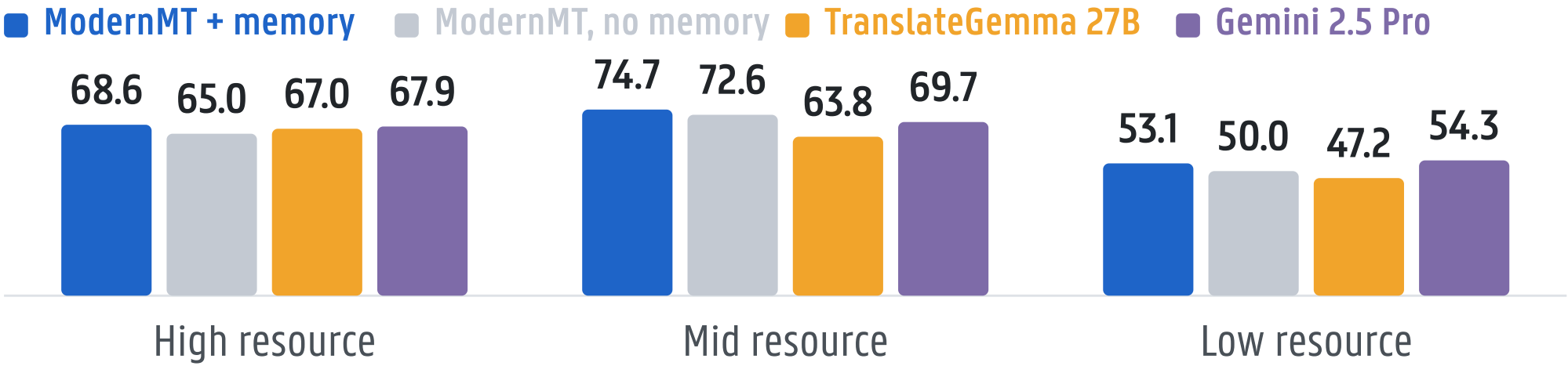
Adaptive NMT still wins overall and Gemini Pro comes within 1.22 chrF



[Moerman, Tezcan & Macken 2026]

Adaptive NMT against retrieval-augmented LLMs

Adaptive NMT still wins overall and Gemini Pro comes within 1.22 chrF



ModernMT with the memory averages 66.60 chrF and wins 8 of 14 languages

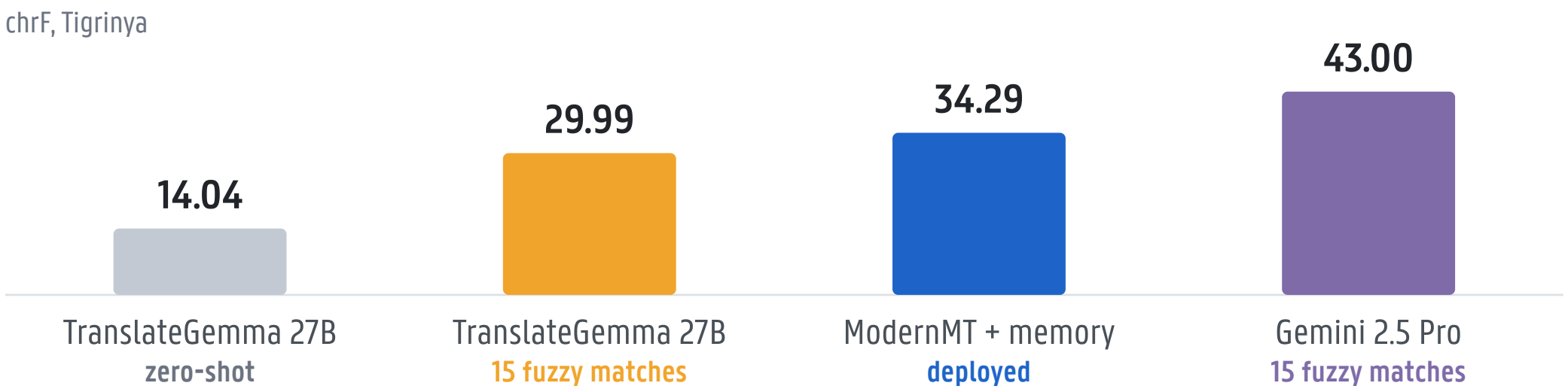
Gemini Pro averages 65.38 and wins 6 of 14

Gemini Pro sends sensitive data to a commercial API

[Moerman, Tezcan & Macken 2026]

Tigrinya

Gemini Pro beats the deployed system by 8.7 chrF



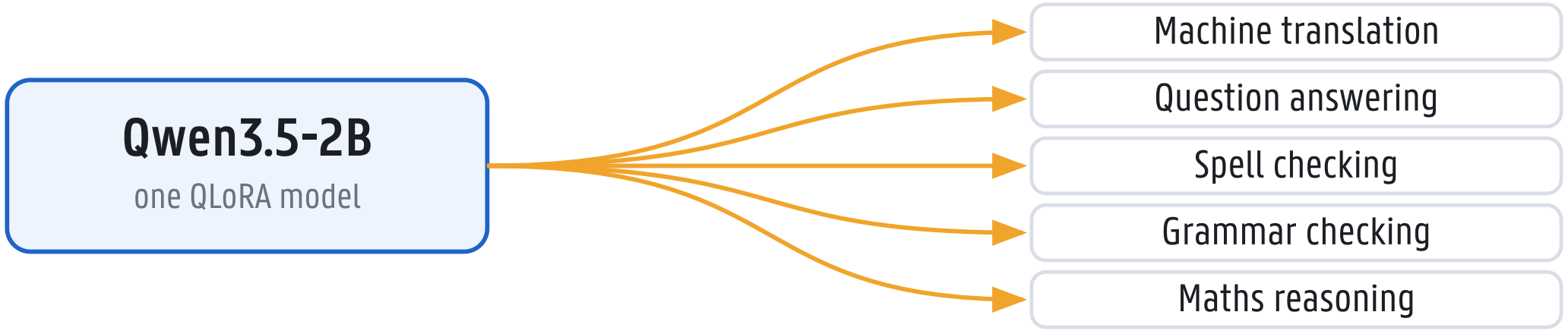
Fuzzy matches more than double the zero-shot score

An earlier human evaluation rated the deployed Tigrinya output too low for end users

[Moerman, Tezcan & Macken 2026]
[Macken et al. 2025]

Task

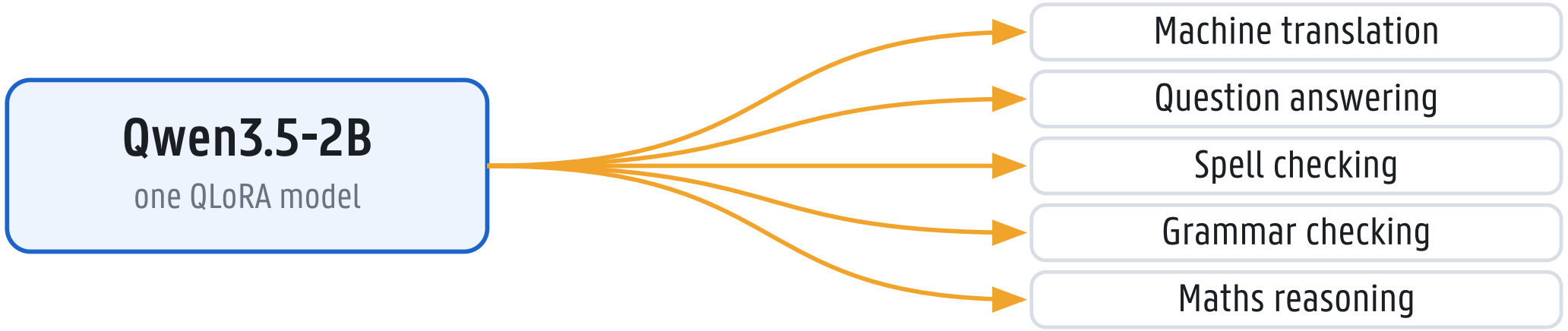
One 2B model serves five Sorbian tasks



[Moerman & Tezcan 2026]

Task

One 2B model serves five Sorbian tasks



FUZZY MATCHES BY STAGE

0 or 1 match in fine-tuning and 3 at inference

SYNTHETIC TASK DATA

Typos, paradigm swaps and cloze items

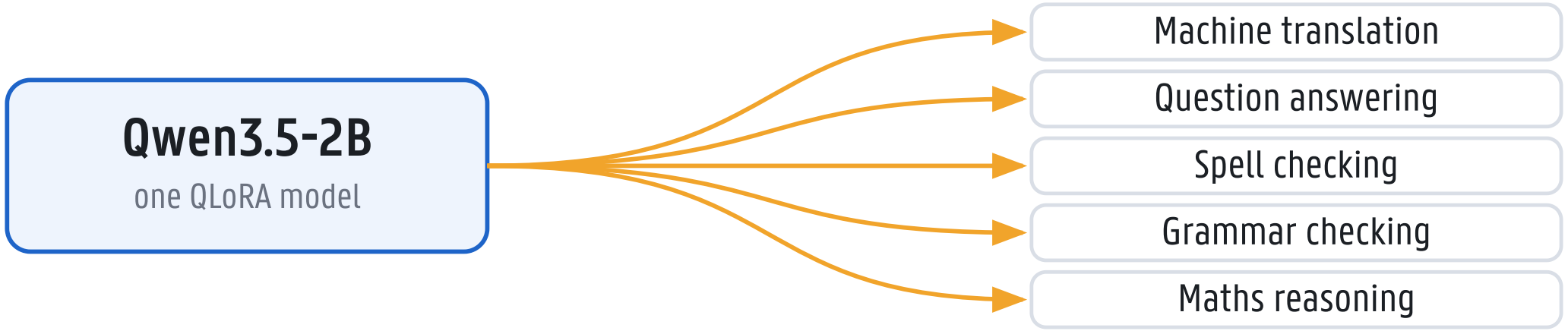
SCORE, DO NOT WRITE

QA, spell and grammar checking rank candidates

[Moerman & Tezcan 2026]

Task

One 2B model serves five Sorbian tasks



FUZZY MATCHES BY STAGE

0 or 1 match in fine-tuning and 3 at inference

SYNTHETIC TASK DATA

Typos, paradigm swaps and cloze items

SCORE, DO NOT WRITE

QA, spell and grammar checking rank candidates

Joint winner of the track on 22 rank points

First on question answering and maths reasoning

[Moerman & Tezcan 2026]

Fuzzy matches by stage

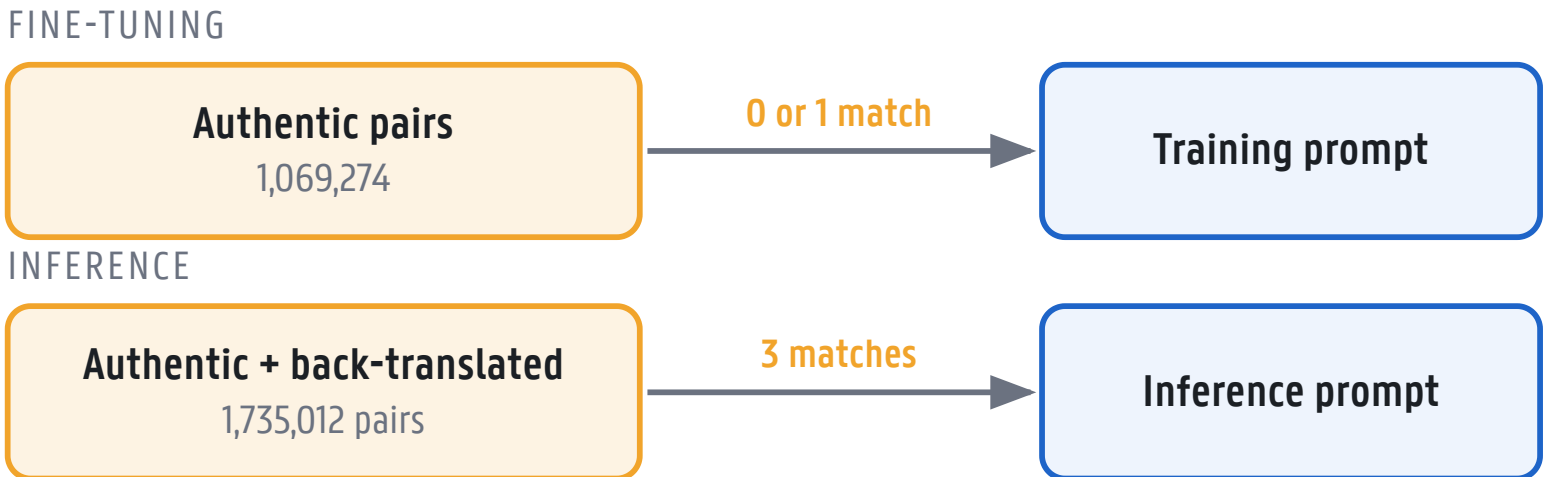
Authentic matches in fine-tuning and the full pool at inference



[Moerman & Tezcan 2026]

Fuzzy matches by stage

Authentic matches in fine-tuning and the full pool at inference



Similarity

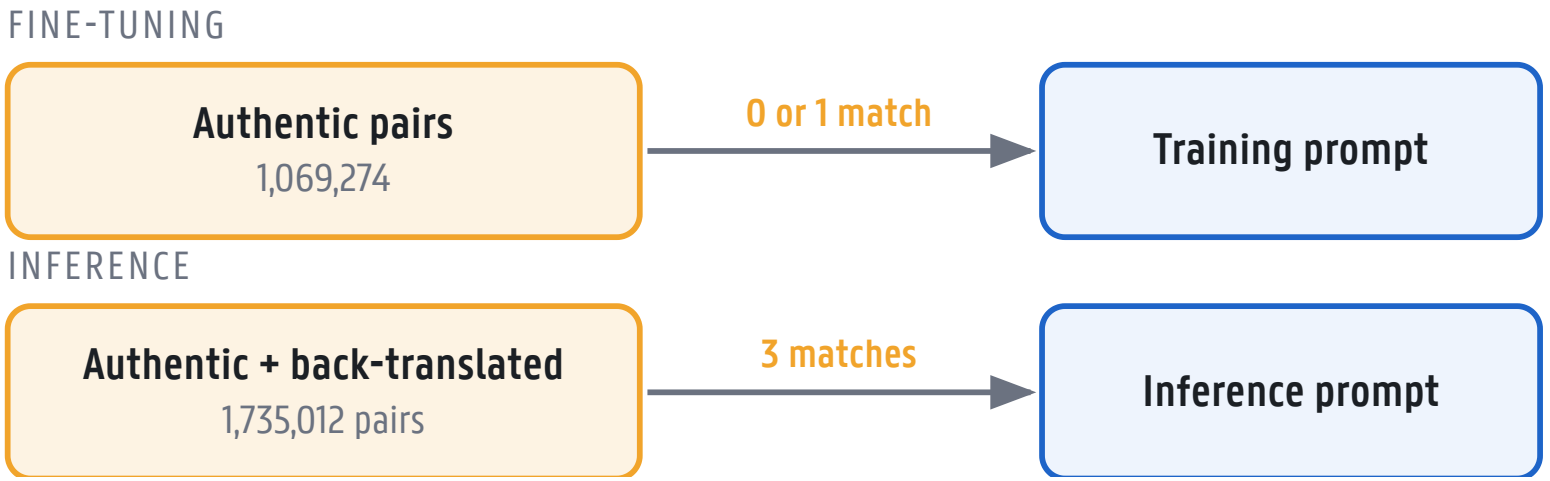
MiniLM cosine
for a German source

character-set Jaccard
for a Sorbian source
since no encoder covers it

[Moerman & Tezcan 2026]

Fuzzy matches by stage

Authentic matches in fine-tuning and the full pool at inference



Similarity

MiniLM cosine
for a German source

character-set Jaccard
for a Sorbian source
since no encoder covers it

Fuzzy matches in fine-tuning +3.63 chrF++ over random

Authentic-only rows +6.26 over rows with back-translation

Back-translated pairs in the inference pool +0.06

[Moerman & Tezcan 2026]

Synthetic task data

We synthesised the training data for the four non-MT tasks

TASK	HOW THE DATA IS MADE	SYNTHESISED	ORGANISERS' ITEMS
Spell checking	character-level typos injected into clean sentences	115,000	4,000
Grammar checking	paradigm swaps from the soblex and dsb-spell lexicons	115,000	3,250
Question answering	cloze items built from monolingual sentences	72,000	158
Maths reasoning	GSM8K and MATH problems translated by our own MT engine	28,147	48

The MT engine was built first because it translates the maths problems

[Moerman & Tezcan 2026]

Scoring instead of writing

A 2B model scores candidates better than it writes them

QUESTION ANSWERING

Each option is placed in the question and scored by likelihood. Shared option banks are solved jointly

SPELL CHECKING

Dictionary candidates for the flagged word are scored in the sentence and the best is kept

GRAMMAR CHECKING

Inflected forms are the candidates. An extra correction needs a scoring margin and a corpus witness

[Moerman & Tezcan 2026]

Scoring instead of writing

A 2B model scores candidates better than it writes them

QUESTION ANSWERING

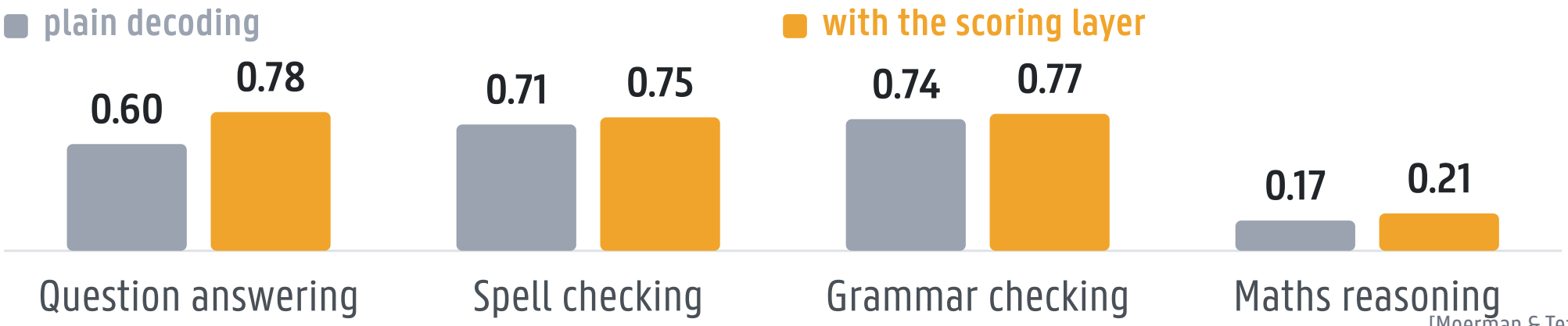
Each option is placed in the question and scored by likelihood. Shared option banks are solved jointly

SPELL CHECKING

Dictionary candidates for the flagged word are scored in the sentence and the best is kept

GRAMMAR CHECKING

Inflected forms are the candidates. An extra correction needs a scoring margin and a corpus witness



[Moerman & Tezcan 2026]

Data scenario

A language model generates the missing in-domain text

Fuzzy matches

plenty of bilingual pairs

bilingual pairs



monolingual text in the domain

not needed

the memory is the data

Back-translation

limited pairs and ample monolingual text

bilingual pairs



monolingual text in the domain



make synthetic pairs from the text

[Moerman, Lefever & Tezcan, in preparation]

Data scenario

A language model generates the missing in-domain text

Fuzzy matches

plenty of bilingual pairs

bilingual pairs



monolingual text in the domain

not needed

the memory is the data

Back-translation

limited pairs and ample monolingual text

bilingual pairs



monolingual text in the domain



make synthetic pairs from the text

This study

limited pairs and no monolingual text

bilingual pairs



monolingual text in the domain

none. We generate it

generate the text and back-translate it

[Moerman, Lefever & Tezcan, in preparation]

Data scenario

A language model generates the missing in-domain text

Fuzzy matches

plenty of bilingual pairs

bilingual pairs



monolingual text in the domain

not needed

the memory is the data

Back-translation

limited pairs and ample monolingual text

bilingual pairs



monolingual text in the domain



make synthetic pairs from the text

This study

limited pairs and no monolingual text

bilingual pairs



monolingual text in the domain

none. We generate it

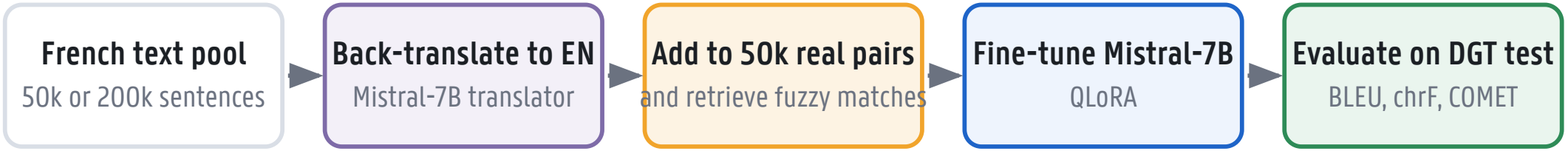
generate the text and back-translate it

The aim throughout is to make a system with limited data as good as possible

[Moerman, Lefever & Tezcan, in preparation]

Setup

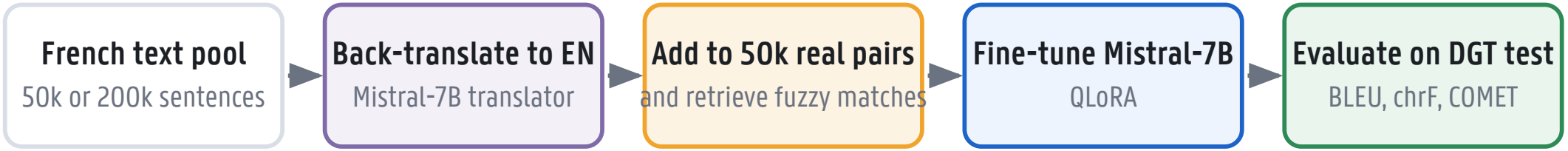
Five text pools go through the same back-translation pipeline



[Moerman, Lefever & Tezcan, in preparation]

Setup

Five text pools go through the same back-translation pipeline



WHAT VARIES IS THE POOL

Real in-domain, DGT French | Real general, news | Narrow synthetic | Extract-based synthetic | Open synthetic

Two pool sizes of 50k and 200k sentences. The generator is Mixtral-8x7B and the student is Mistral-7B.

EN to FR legal.

[Moerman, Lefever & Tezcan, in preparation]

Three prompts

The three synthetic pools differ only in the prompt

Same Mixtral-8x7B generator, same real seeds and same cleaning. Only the prompt differs.

Narrow (low)

least varied and most legal-looking

Persona a distinguished French legal scholar
Examples 5 real DGT sentences
Asks for 3 sentences in the same professional style

Extract-based (middle)

in between on every measure

Persona a senior legal linguist for the Commission
Examples one 50-word DGT extract with a topic, an audience and a format
Asks for a text that expands the extract

Open (high)

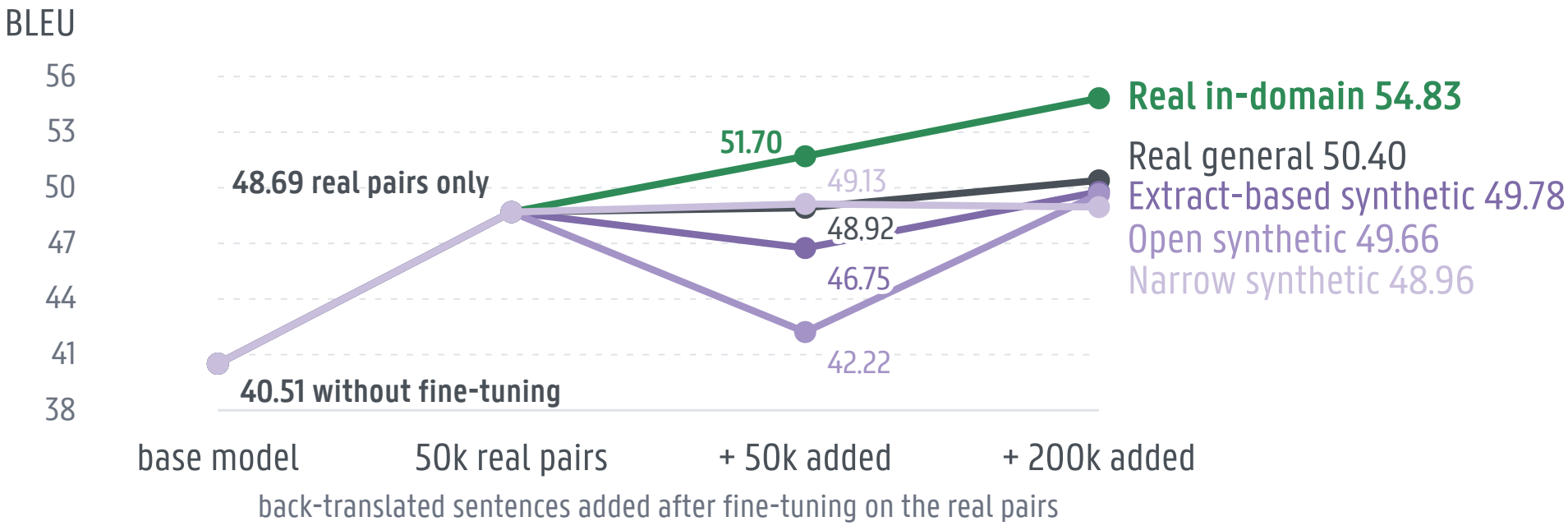
most varied and least legal-looking

Persona a legal writing assistant
Examples 5 real DGT sentences
Asks for 3 similar legal sentences with varied syntax

[Moerman, Lefever & Tezcan, in preparation]

Results

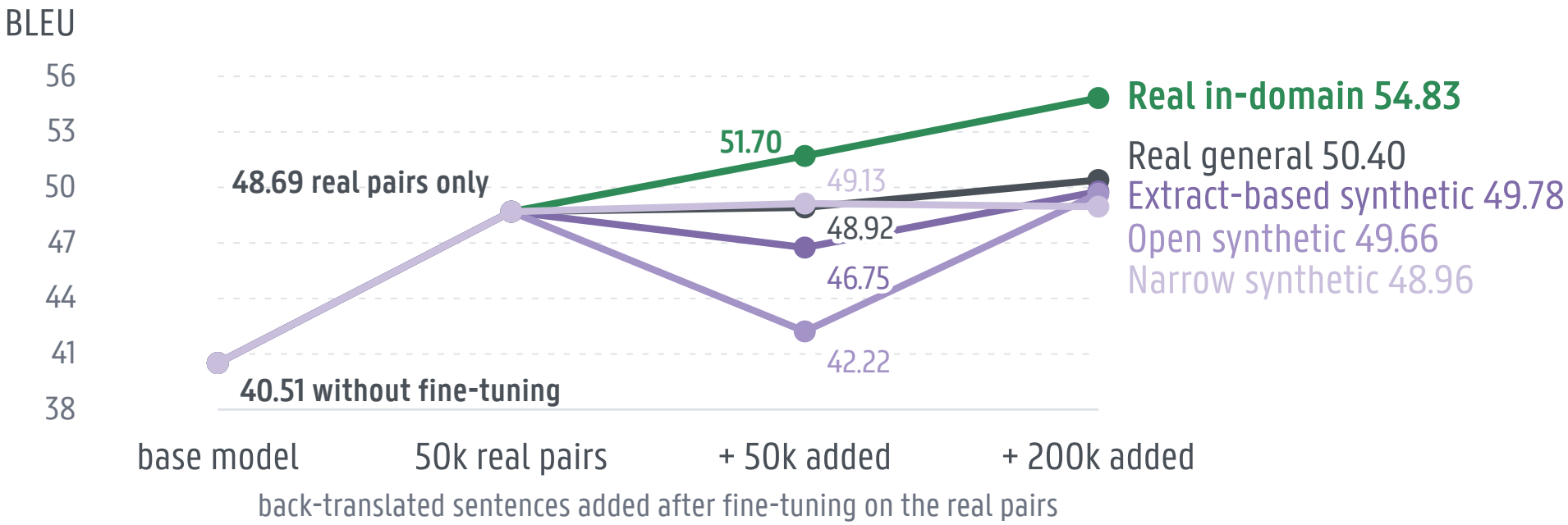
The gap between synthetic and real in-domain data grows with scale



[Moerman, Lefever & Tezcan, in preparation]

Results

The gap between synthetic and real in-domain data grows with scale



Real in-domain is 2.43 BLEU ahead of the best synthetic pool at 50k and 5.49 ahead at 200k

Preliminary

[Moerman, Lefever & Tezcan, in preparation]

Text measures

The measures restate the prompt and do not predict translation quality

POOLS ORDERED BY	LOWEST			HIGHEST
BLEU, 50k added	Open 42.2	Extract-based 46.7	Narrow 49.1	Real in-domain 51.7

[Moerman, Lefever & Tezcan, in preparation]

[Guo et al. 2025]

Text measures

The measures restate the prompt and do not predict translation quality

POOLS ORDERED BY	LOWEST			HIGHEST
BLEU, 50k added	Open 42.2	Extract-based 46.7	Narrow 49.1	Real in-domain 51.7
Lexical diversity	Narrow 0.40	Extract-based 0.47	Open 0.52	Real in-domain 0.53
Surprise for a legal LM	Narrow 6.6	Extract-based 9.3	Open 11.8	Real in-domain 16.2
Test-set coverage	Narrow 0.59	Extract-based 0.66	Open 0.67	Real in-domain 0.70
Surprise for a general LM	Narrow 4.4	Extract-based 6.9	Open 9.3	Real in-domain 19.5

[Moerman, Lefever & Tezcan, in preparation]

[Guo et al. 2025]

Text measures

The measures restate the prompt and do not predict translation quality

POOLS ORDERED BY	LOWEST			HIGHEST
BLEU, 50k added	Open 42.2	Extract-based 46.7	Narrow 49.1	Real in-domain 51.7
Lexical diversity	Narrow 0.40	Extract-based 0.47	Open 0.52	Real in-domain 0.53
Surprise for a legal LM	Narrow 6.6	Extract-based 9.3	Open 11.8	Real in-domain 16.2
Test-set coverage	Narrow 0.59	Extract-based 0.66	Open 0.67	Real in-domain 0.70
Surprise for a general LM	Narrow 4.4	Extract-based 6.9	Open 9.3	Real in-domain 19.5
BLEU, 200k added	Narrow 49.0	Open 49.7	Extract-based 49.8	Real in-domain 54.8

[Moerman, Lefever & Tezcan, in preparation]

[Guo et al. 2025]

Text measures

The measures restate the prompt and do not predict translation quality

POOLS ORDERED BY	LOWEST			HIGHEST
BLEU, 50k added	Open 42.2	Extract-based 46.7	Narrow 49.1	Real in-domain 51.7
Lexical diversity	Narrow 0.40	Extract-based 0.47	Open 0.52	Real in-domain 0.53
Surprise for a legal LM	Narrow 6.6	Extract-based 9.3	Open 11.8	Real in-domain 16.2
Test-set coverage	Narrow 0.59	Extract-based 0.66	Open 0.67	Real in-domain 0.70
Surprise for a general LM	Narrow 4.4	Extract-based 6.9	Open 9.3	Real in-domain 19.5
BLEU, 200k added	Narrow 49.0	Open 49.7	Extract-based 49.8	Real in-domain 54.8

Real text looks least legal-like and least fluent by these measures

Real text wins at both sizes

[Moerman, Lefever & Tezcan, in preparation]

[Guo et al. 2025]

Other projects

The same two ideas in three more lines of work

SMT MEETS LLMS

Phrase tables as examples

A phrase table from statistical MT holds aligned fragments. The LLM gets matching fragments of the source as examples instead of whole sentences.

MECHANISTIC INTERPRETABILITY

Why fuzzy matches work

Which layers and heads copy from the retrieved pair. What the model does with a match inside the forward pass.

OUTSIDE MT

Grading and gender ambiguity

ShAnEL grades short answers in language learning with retrieved student answers as examples. **ContraGAND** detects gender ambiguity with synthetic contrastive variants for fine-tuning and distillation.

Where this leaves us

Fuzzy-match augmentation and synthetic data make small MT systems competitive with much larger ones.

The open question is whether synthetic data can be made good enough to **retrieve** and **train** from. That is the goal of the research stay.

PART 2

Research stay

Two open problems

Which examples to use from a pool that mixes real and synthetic → An in-context example selector trained with reinforcement learning

How to make the synthetic data itself more human-like → Steering vectors on the generator and a machine-translation detector as judge

[García-Romero, PhD thesis]

Example selector and MT-detector

What is already done and what we will add to it

	WHAT IS ALREADY DONE	WHAT WE WILL ADD
Example selector	An agent learns which five examples to put in the prompt and in which order. The translation quality of the result is its reward.	A pool that mixes real and synthetic examples. Which does the selector prefer and does that help?

[García-Romero, PhD thesis]
[García-Romero, Esplà-Gomis & Sánchez-Martínez 2025]

Example selector and MT-detector

What is already done and what we will add to it

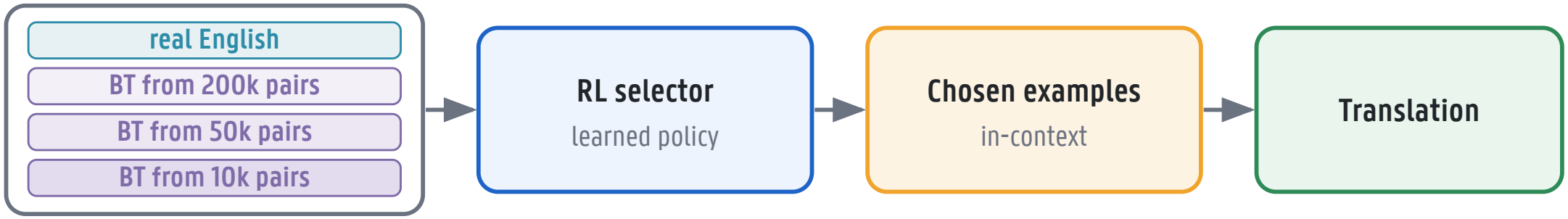
	WHAT IS ALREADY DONE	WHAT WE WILL ADD
Example selector	An agent learns which five examples to put in the prompt and in which order. The translation quality of the result is its reward.	A pool that mixes real and synthetic examples. Which does the selector prefer and does that help?
MT-detector	A classifier reads a source and its translation through a frozen translation model. It says whether a person or a machine translated it.	The detector judges how human-like synthetic text is. A steering vector moves the generator toward what it calls human.

[García-Romero, PhD thesis]

[García-Romero, Esplà-Gomis & Sánchez-Martínez 2025]

Example selection

Does the RL selector prefer real or synthetic examples?

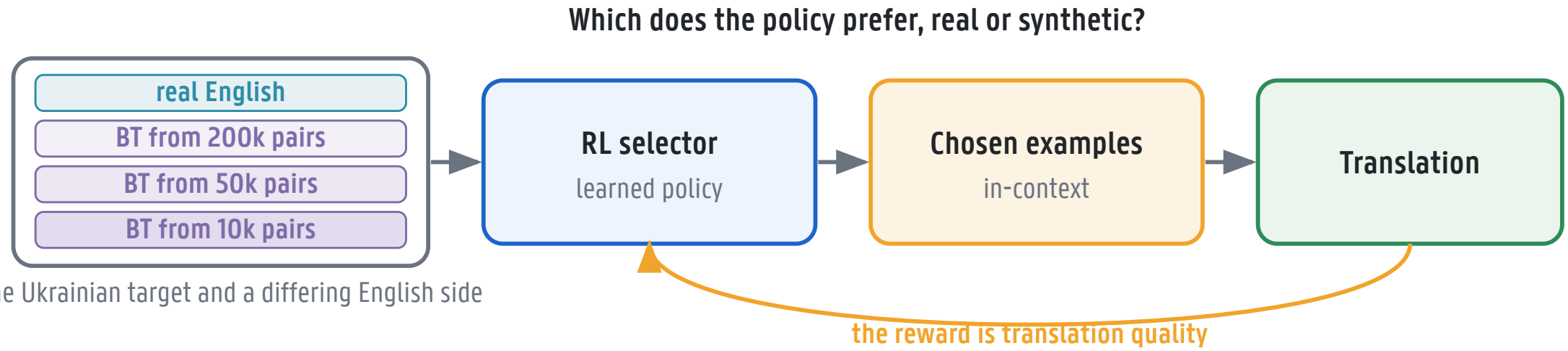


ame Ukrainian target and a differing English side

[García-Romero, PhD thesis]

Example selection

Does the RL selector prefer real or synthetic examples?

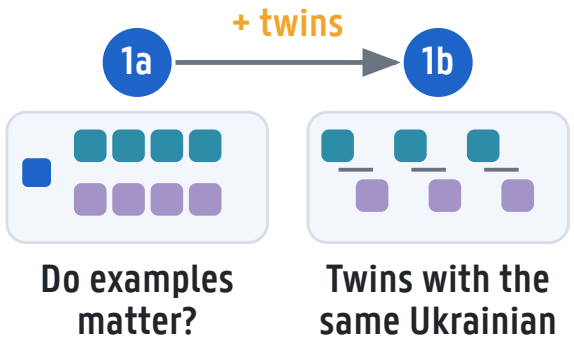


ame Ukrainian target and a differing English side

[García-Romero, PhD thesis]

Research questions for the selector

Does a learned selector pick the best examples from real plus synthetic pairs?



1a Does the choice of examples matter and do similarity heuristics get it right?

Real examples against weak back-translations for the same query. Do BM25 and SONAR notice?

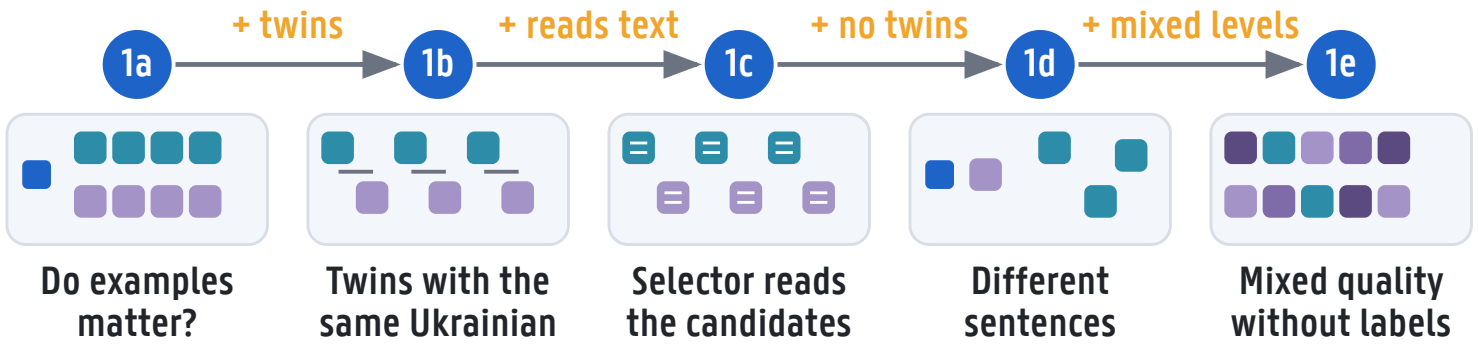
1b Can the selector learn to pick the better version when all else is equal?

Twin versions with the same Ukrainian. How often does it pick the real English side?

[García-Romero, PhD thesis]

Research questions for the selector

Does a learned selector pick the best examples from real plus synthetic pairs?



1c Does a selector that reads its candidates learn the same and carry it to new sentences?

It scores each candidate from its own text. Does the preference hold on unseen sentences?

1d When does a closer synthetic match beat a less similar real one?

Real and synthetic are different sentences. Close synthetic match or less similar real one?

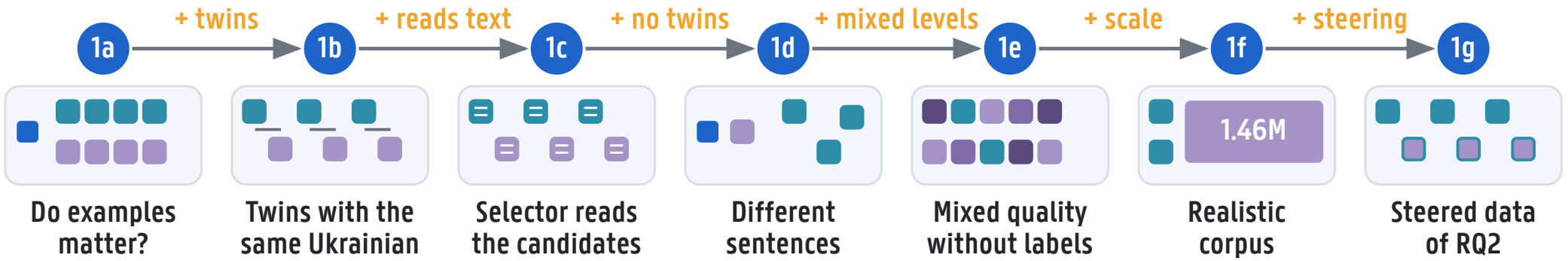
1e Does it rank unmarked quality levels by usefulness rather than origin?

Back-translations of several qualities mixed in unlabelled. Does it sort them by usefulness?

[García-Romero, PhD thesis]

Research questions for the selector

Does a learned selector pick the best examples from real plus synthetic pairs?



1f Does it beat the heuristics and both single corpora on limited real pairs plus 1.46M back-translations?

Few real pairs plus 1.46M back-translations. Better than BM25 and SONAR and either corpus alone?

1g Does the more human-like synthetic data of RQ2 change its preference?

The steered back-translations of RQ2 replace the plain ones. Does the real share of its picks move?

[García-Romero, PhD thesis]

Human-like synthetic text

Machine text carries a signature and we steer the generator away from it

DETECTOR SCORE

HUMAN-WRITTEN ENGLISH
The applicant may lodge an appeal against the decision within fourteen days.

human 0.9

BACK-TRANSLATED ENGLISH
The applicant can submit the appeal against this decision in the term of fourteen days.

machine 0.9

illustrative numbers

[García-Romero, PhD thesis]

Human-like synthetic text

Machine text carries a signature and we steer the generator away from it

HUMAN-WRITTEN ENGLISH
The applicant may lodge an appeal against the decision within fourteen days.

BACK-TRANSLATED ENGLISH
The applicant can submit the appeal against this decision in the term of fourteen days.

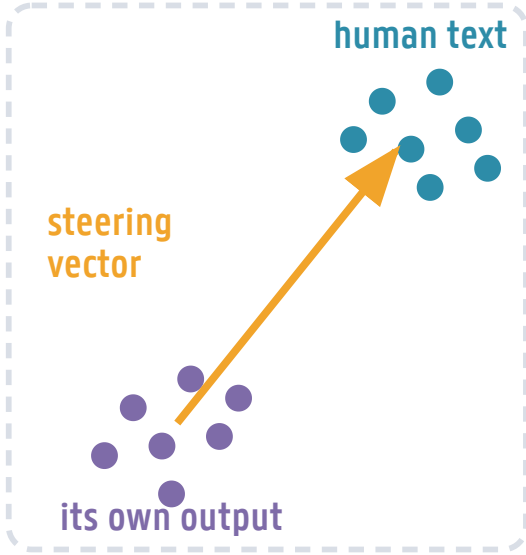
DETECTOR SCORE

human 0.9

machine 0.9

illustrative numbers

INSIDE THE GENERATOR



[García-Romero, PhD thesis]

Human-like synthetic text

Machine text carries a signature and we steer the generator away from it

HUMAN-WRITTEN ENGLISH
The applicant may lodge an appeal against the decision within fourteen days.

BACK-TRANSLATED ENGLISH
The applicant can submit the appeal against this decision in the term of fourteen days.

STEERED BACK-TRANSLATION
The applicant may lodge an appeal against this decision within fourteen days.

DETECTOR SCORE

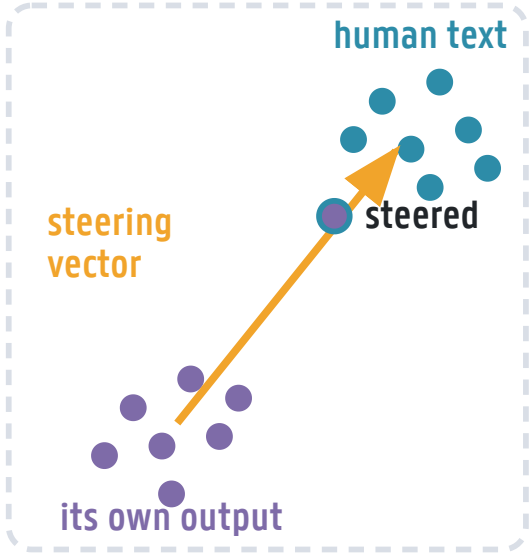
human 0.9

machine 0.9

human 0.7

illustrative numbers

INSIDE THE GENERATOR



[García-Romero, PhD thesis]

Human-like synthetic text

Machine text carries a signature and we steer the generator away from it

HUMAN-WRITTEN ENGLISH
The applicant may lodge an appeal against the decision within fourteen days.

BACK-TRANSLATED ENGLISH
The applicant can submit the appeal against this decision in the term of fourteen days.

STEERED BACK-TRANSLATION
The applicant may lodge an appeal against this decision within fourteen days.

DETECTOR SCORE

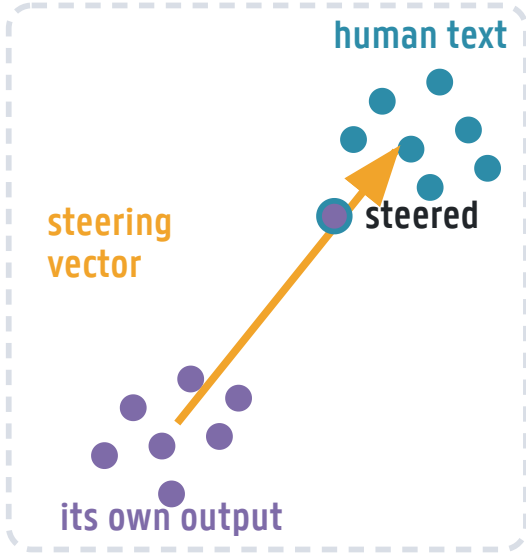
human 0.9

machine 0.9

human 0.7

illustrative numbers

INSIDE THE GENERATOR



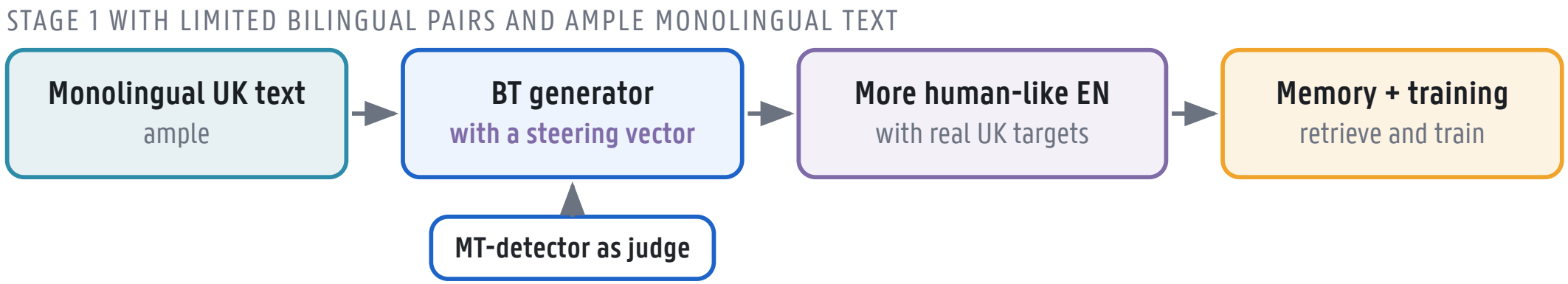
The detector judges and the vector steers

Meaning is checked separately against the Ukrainian

[García-Romero, PhD thesis]

RQ2 in two stages

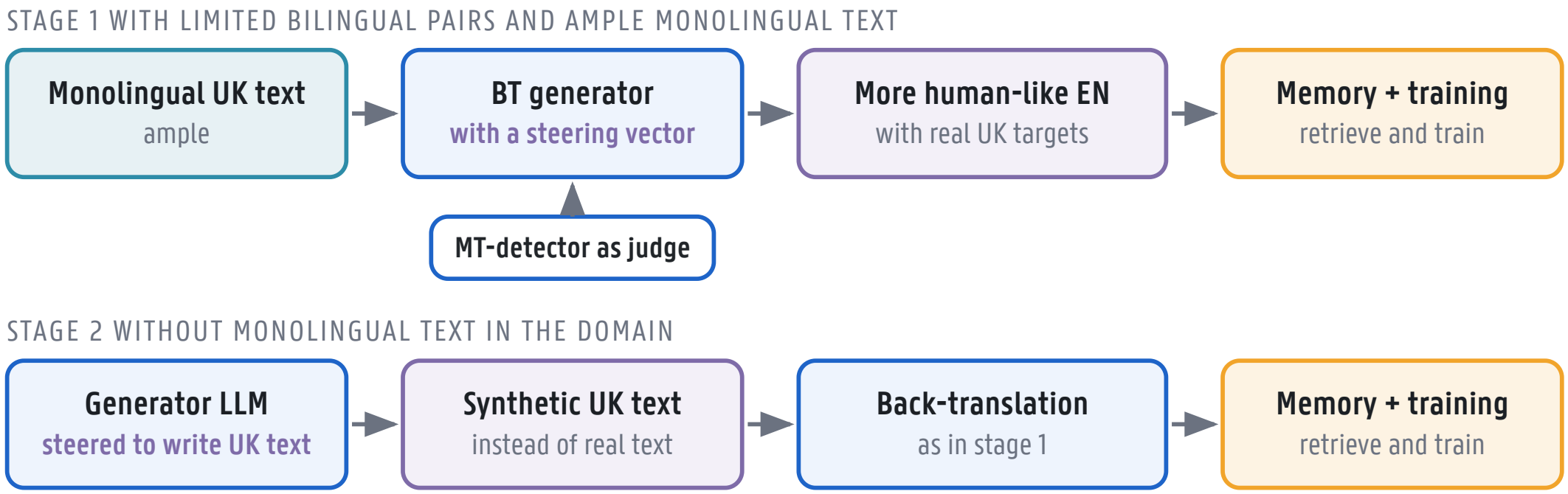
Stage 1 steers the back-translation and stage 2 also steers the text generator



[García-Romero, PhD thesis]

RQ2 in two stages

Stage 1 steers the back-translation and stage 2 also steers the text generator



[García-Romero, PhD thesis]

Combining RQ1 and RQ2

The steered data enters the selection pool



The more human-like synthetic data of RQ2 enters the selection pool of RQ1. Does the selector's preference between real and synthetic examples change?

Thank you

Questions and ideas welcome

Thomas Moerman · thomas.moerman@ugent.be

LT3, Ghent University · research stay at Transducens

lt3.ugent.be · thomasmoerman.dev · huggingface.co/TomMoeras · github.com/TomMoeras

GHENT UNIVERSITY

language and translation technology team

TransTech

BACKUP

Backup slides, for questions

BACKUP

BLEU and COMET per direction in the NMT study

SYSTEM	EN TO UK BLEU	COMET	EN TO FR BLEU	COMET	FR TO EN BLEU	COMET
BASE	54.85	91.23	51.50	84.22	53.95	85.01
BT	56.21	92.18	54.88	85.27	56.99	86.72
NFR	57.73	91.78	52.67	84.50	54.43	85.26
NFR mono	60.39	92.03	52.39	84.45	55.54	85.44
BT + NFR	**66.95**	**92.78**	**61.91**	**87.13**	**64.69**	**87.79**
BASE HQ			59.13	87.40	61.56	88.03
NFR HQ			64.58	88.36	67.75	88.61
GPT-4o	41.66	92.35	43.47	86.75	44.58	86.87
Llama 3.1 70B	27.81	88.70	44.90	86.37	49.37	86.56

[Tezcan, Skidanova & Moerman 2024]

BACKUP

Per-language chrF for four systems in the asylum study

LANGUAGE	MMT+	MMT	TG-27B	GEMINI PRO
Arabic	**66.76**	63.78	63.63	63.76
German	73.67	69.26	70.97	**73.85**
Spanish	72.64	69.77	69.98	**74.30**
Farsi	54.01	51.36	**58.43**	55.03
Portuguese	**80.97**	75.48	75.65	76.54
Romanian	**67.97**	63.10	63.80	67.22
Russian	67.62	65.25	**67.95**	67.01

LANGUAGE	MMT+	MMT	TG-27B	GEMINI PRO
Turkish	65.34	62.20	65.47	**65.60**
Armenian	**74.26**	68.58	68.61	73.42
Albanian	70.01	65.46	61.15	**70.05**
Georgian	79.87	**83.74**	61.70	65.69
Pashto	**66.82**	64.86	57.10	62.53
Somali	**58.19**	55.57	54.47	57.34
Tigrinya	34.29	29.48	29.99	**43.00**
Average	**66.60**	63.42	62.06	65.38

[Moerman, Tezcan & Macken 2026]

BACKUP

Fuzzy against random per language for TranslateGemma 27B

LANGUAGE	RANDOM	FUZZY	Δ	P
Arabic	60.82	63.63	+2.81	< 0.01
German	69.22	70.97	+1.75	< 0.01
Spanish	70.35	69.98	−0.37	n.s.
Farsi	55.44	58.43	+2.99	< 0.001
Portuguese	74.27	75.65	+1.38	< 0.001
Romanian	63.34	63.80	+0.46	n.s.
Russian	63.25	67.95	+4.70	< 0.001

LANGUAGE	RANDOM	FUZZY	Δ	P
Turkish	62.72	65.47	+2.75	< 0.001
Armenian	64.48	68.61	+4.13	< 0.001
Albanian	61.08	61.15	+0.07	n.s.
Georgian	57.92	61.70	+3.78	< 0.001
Pashto	53.21	57.10	+3.89	< 0.001
Somali	50.39	54.47	+4.08	< 0.001
Tigrinya	21.16	29.99	**+8.83**	< 0.001

GHENT UNIVERSITY
language and translation technology team
Trans Tech

[Moerman, Tezcan & Macken 2026]

BACKUP

LLM study in numbers

LEGAL EN TO UK, 200K + 1M BT	BLEU	CHRF	COMET
NMT	56.01	76.82	92.31
NFR	63.12	80.21	92.74
Mistral-7B, FM FT+ICL BT	60.59	78.31	92.72
Tower-9B	**65.44**	**81.89**	**93.94**
Mistral-Small-24B	65.20	81.55	93.87
GPT-4.1, zero-shot	45.15	70.88	92.79
Gemini 2.5 Pro, zero-shot	51.11	74.57	93.29

PATENT EN TO FR	BLEU	CHRF	COMET
NMT, 200k + 1M BT	57.98	77.97	84.81
NFR, 200k + 1M BT	61.74	79.82	85.29
FM FT+ICL, 200k	57.41	77.35	85.38
FM FT+ICL BT, 200k + 1M	**64.20**	**81.08**	**86.34**
Random FT+ICL BT	60.69	79.37	85.87
Tower-9B	65.18	81.77	86.58
Mistral-Small-24B	65.18	81.69	86.64

The optimal number of examples at inference is 2 for legal with BT and 15 to 35 for patents

[Moerman, Lefever & Tezcan, under review]

BACKUP

Full table of the synthetic-data study

POOL	KIND	BLEU 50K	BLEU 200K	CHRF 200K	COMET 200K
No added data	anchor	48.69		69.20	86.31
Real in-domain, DGT	real	51.70	**54.83**	**73.57**	**87.50**
Real general, news	real	48.92	50.40	70.40	86.61
Open (high) synthetic	synthetic	42.22	49.66	70.02	86.70
Extract-based (middle) synthetic	synthetic	46.75	49.78	70.14	86.59
Narrow (low) synthetic	synthetic	49.13	48.96	69.50	86.35
50/50 news and extract-based	combination	48.26	49.90	70.16	86.61
Three synthetic tiers mixed	combination	48.57	49.78	70.12	86.76

[Moerman, Lefever & Tezcan, in preparation]