MT Evaluation

Wei Xu

(many slides from Greg Durrett)

Mean (Math Review)

- Arithmetic Mean = (P + R) / 2
- Geometric Mean = $\sqrt{P \times R}$
- Harmonic Mean = 2 x P x R / (P + R)

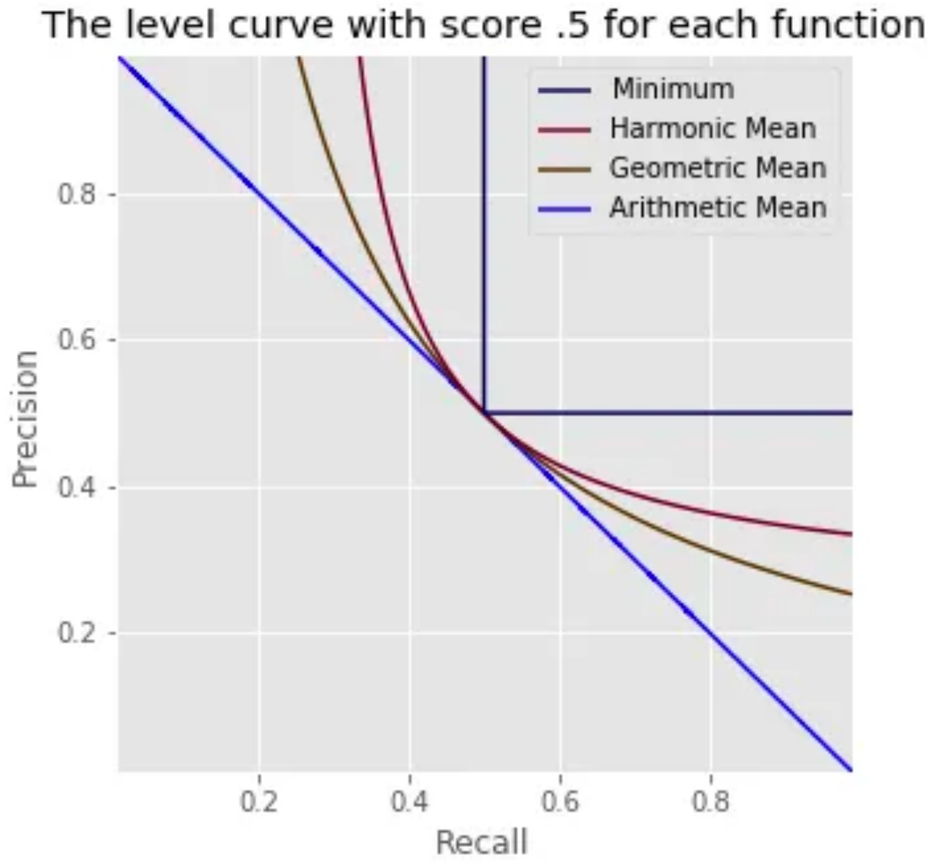


Image credit: Greg Gandenberger

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

		1-gram	2-gram	3-gram
hypothesis 1	I am exhausted	3/3	1/2	0/1
hypothesis 2	Tired is I	1/3	0/2	0/1
hypothesis 3	I I I	1/3	0/2	0/1
reference 1	I am tired			
reference 2	I am ready to sleep now and so exhausted			

Papineni et al. (2002)

Evaluating MT

- Fluency: does it sound good in the target language?
- Fidelity/adequacy: does it capture the meaning of the original?
- BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- Typically $N = 4$, $w_i = 1/4$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- r = length of reference
  c = length of system output

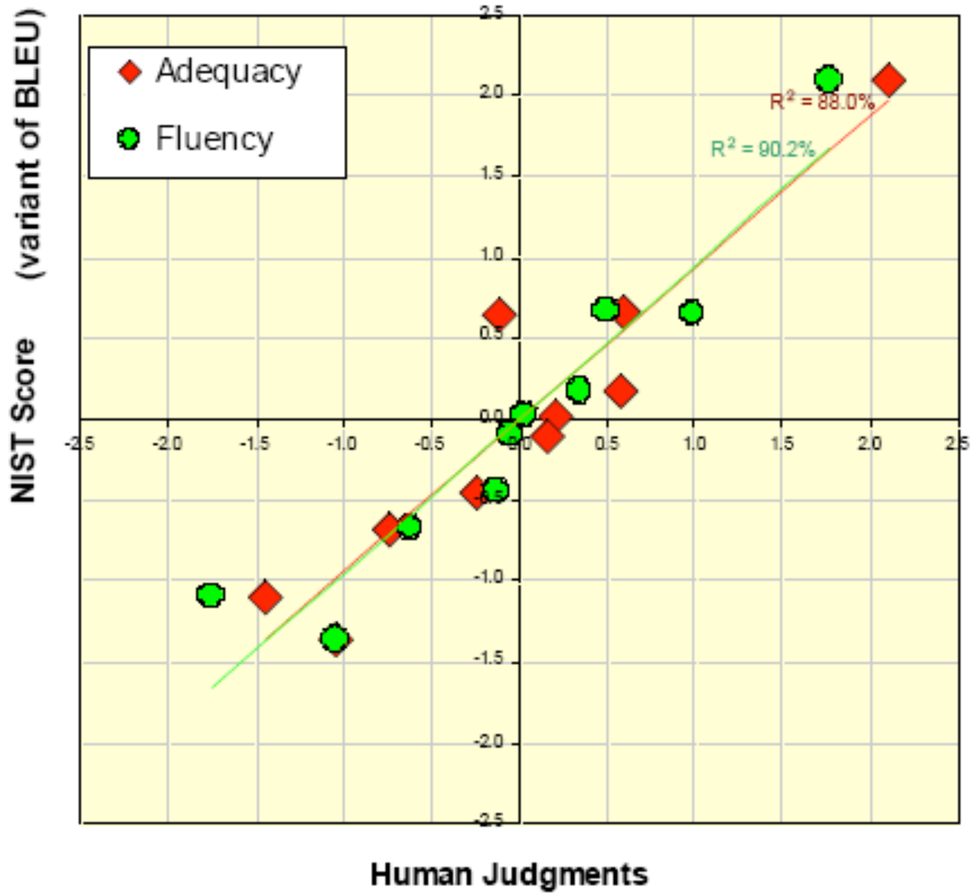
- Does this capture fluency and adequacy?

Papineni et al. (2002)

https://github.com/mjpost/sacrebleu

BLEU Score

- Better methods with human-in-the-loop
- HTER: human-assisted translation error rate
- If you're building real MT systems, you do user studies. In academia, you mostly use BLEU, COMET, etc.



slide from G. Doddington (NIST)

Appraise - Human Evaluation Interface

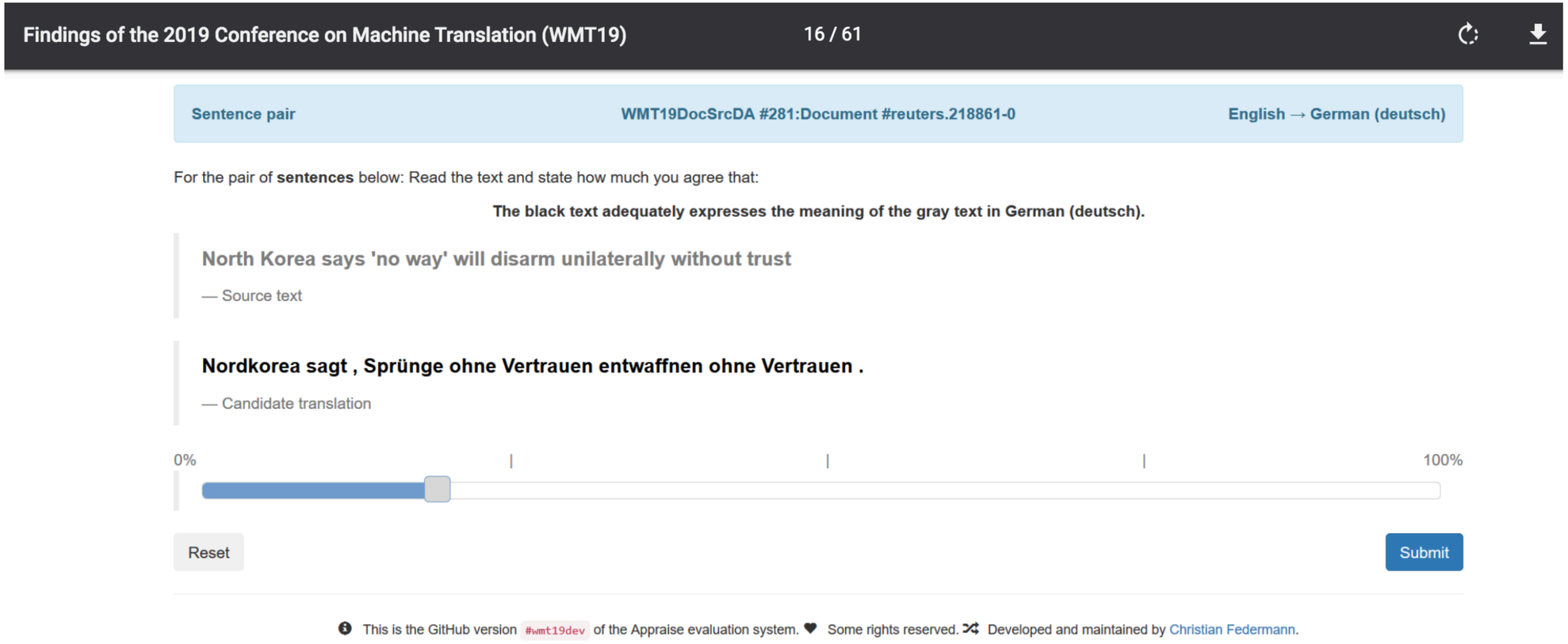
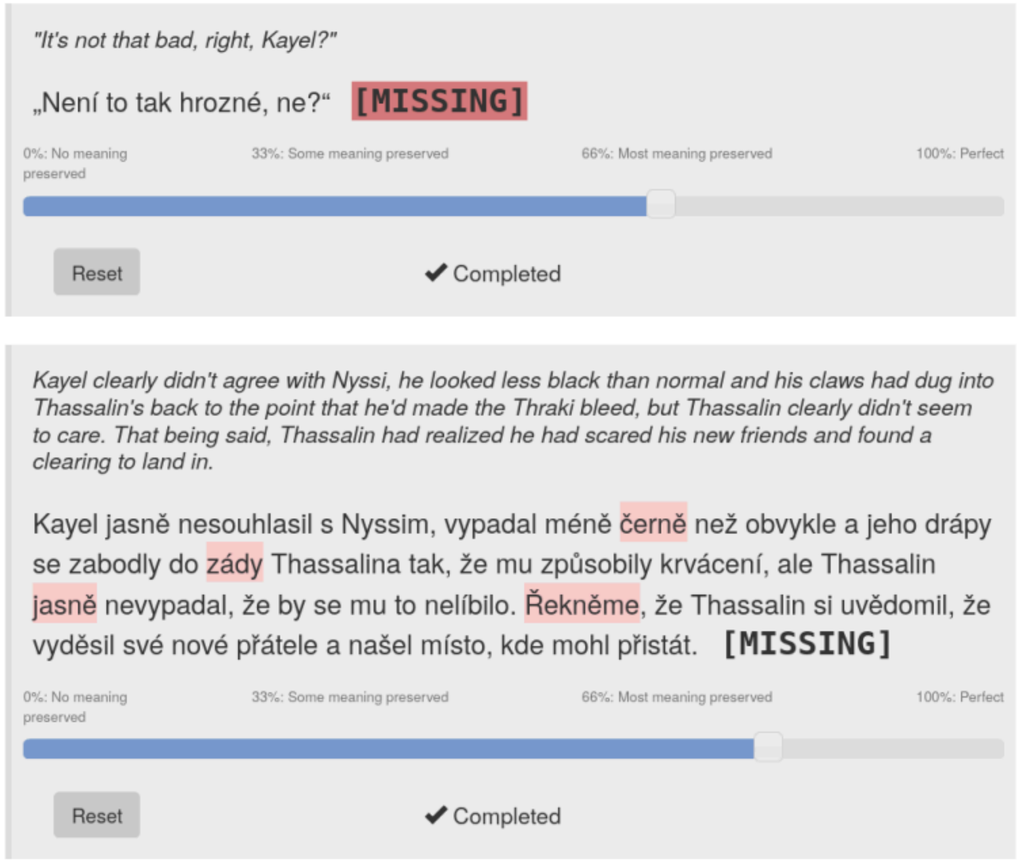


Figure 3: Screen shot of segment-rating portion of document-level direct assessment in the Appraise interface for an example English to German assessment from the human evaluation campaign. The annotator is presented with the machine translation output segment randomly selected from competing systems (anonymized) and is asked to rate the translation on a sliding scale.

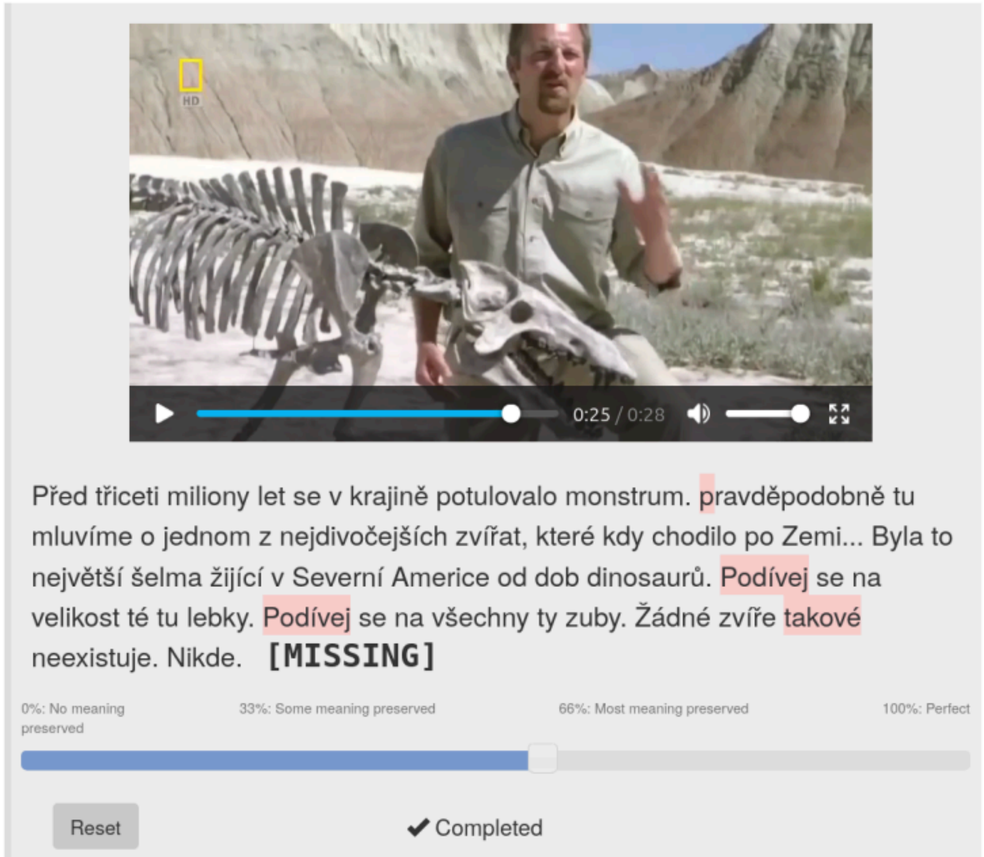
- Direct Assessment (DA) - a subjective quality score on each output

Federmann (2010)

WMT 2024



(a) Excerpt of two segments from a larger document. In the first segment, the name *"Kayel"* is omitted which is a major error. In the second segment, there are many minor errors.



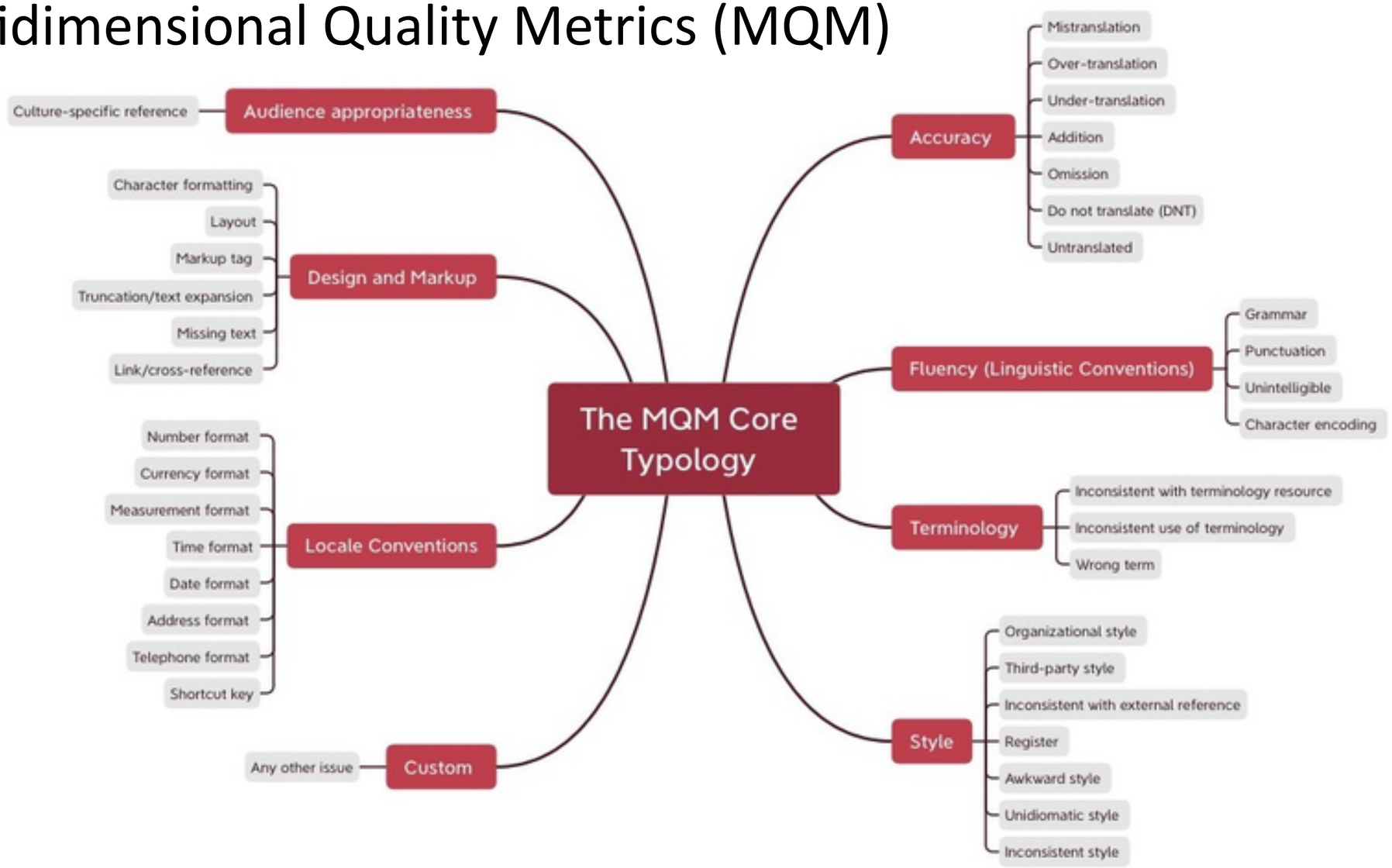
(b) Example of a video to text translation with several minor errors. The annotator can control the video player.

Figure 1: Two screenshots of ESA (Kocmi et al., 2024b) and the annotator instructions. ESA shows multiple segments within a document at once as well as video sources. After marking the individual error spans, the annotator assigns the final segment score from 0 to 100. The tool is implemented in Appraise (Federmann, 2018).

Kocmi et al. (2024)

MQM - word-level Human Eval

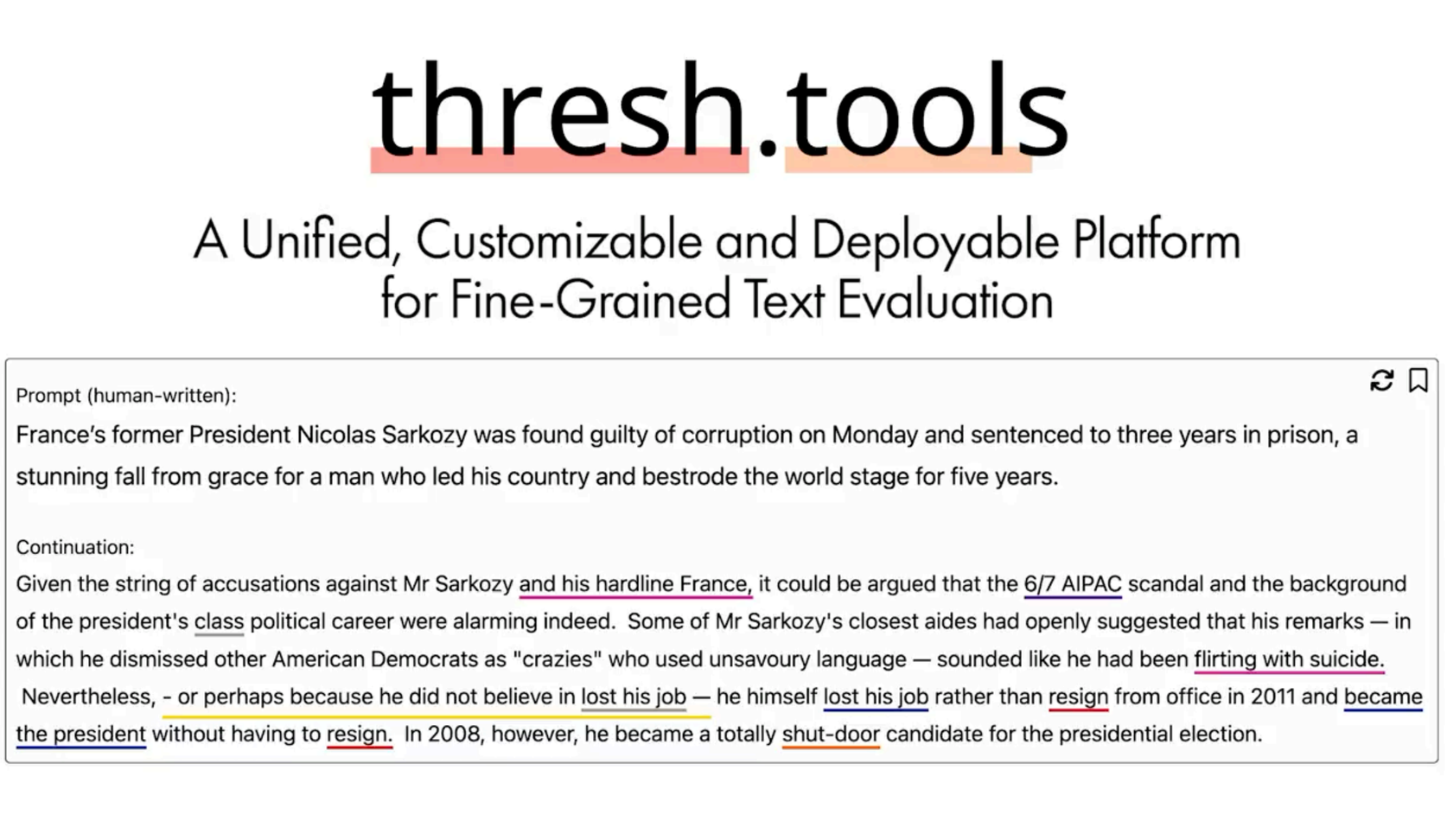
- Multidimensional Quality Metrics (MQM)



Lommel et al. (2014)

Thresh — Fine-grained Text Evaluation Tool

https://github.com/davidheineman/thresh



David Heineman, Yao Dou, Wei Xu. "Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation" (EMNLP 2023 demo)

Automatic MT Evaluation Metrics

- BLEU (2002): n-gram overlap
- METEOR (2005): also take into consideration of synonyms
- HTER (2009): human-assisted translation error rate
- BERTScore (2019): embedding-based
- BLEURT (2020) and COMET (2020): trained neural network model using human evaluation data
- and many more … e.g., CometKiwi-DA-XL (2023), MetricX-23-XL (2023)

WMT 2025

Preliminary Ranking of WMT25 General Machine Translation Systems

Tom Kocmi Eleftherios Avramidis Rachel Bawden Ondřej Bojar Konstantin Dranch
Anton Dvorkovich Sergey Dukanov Natalia Fedorova Mark Fishel Markus Freitag
Thamme Gowda Roman Grundkiewicz Barry Haddow Marzena Karpinska
Philipp Koehn Howard Lakougna Jessica Lundin Kenton Murray Masaaki Nagata
Stefano Perrella Lorenzo Proietti Martin Popel Maja Popović Parker Riley
Mariya Shmatova Steinþór Steingrímsson Lisa Yankovskaya Vilém Zouhar

Introduction

We present the **preliminary** rankings of machine translation (MT) systems submitted to the WMT25 General Machine Translation Shared Task,[1] as determined by automatic evaluation metrics. Because these rankings are derived from automatic evaluation, they may exhibit a bias toward systems that employ re-ranking techniques, such as Quality Estimation or Minimum Bayes Risk decoding. The official WMT25 ranking will be based on human evaluation, which is more reliable and will supersede these results. The official WMT25 ranking will be based on human evaluation, which is more reliable and will supersede these results. The purpose of releasing these findings now is to assist task participants with their system description papers; *not* to provide final findings.

Types of Systems

We distinguish two types of MT systems participating in the shared task:

- **Constrained systems:** must use only publicly available training data and models, be limited to a maximum of 20B parameters, and have their model weights released under an open license.
- **Unconstrained systems:** (marked with gray) are all other systems. They have no restrictions on training data or model size, and there is no requirement to publish the model weights. This category also includes systems for which training information is not public.

[1] www2.statmt.org/wmt25/translation-task.html

Evaluated Systems

Models. Our evaluation includes systems submitted by participants, as well as open-weight and proprietary models. We selected the largest or best-performing version of each model where applicable. All constrained and unconstrained systems are listed in Table 1. Full details for all systems will be available in the upcoming WMT25 finding paper.

Prompts. We prompt all language models using a zero-shot, instruction-following approach, with the specific instructions provided as part of the blind test set. Each model was first tasked with translating the entire document. If this initial attempt failed (e.g., due to producing an incorrect paragraph count or exceeding the token limit), we implemented a fallback strategy of translating the document paragraph by paragraph. This generic setup may disadvantage systems tuned for specific MT instructions, such as TowerLLM or EuroLLM; these are marked with [M].

Additionally, we made two model-specific adjustments: (1) for **Qwen3-235B** reasoning capabilities were disabled, and (2) for **Gemini-2.5-Pro**, no reasoning budget was set, which resulted in a 6.6× increase in output tokens, making it the most expensive model to evaluate.

The code for collecting translations is publicly available at github.com/wmt-conference/wmt-collect-translations and we marked all systems collected by us with ▲.

Evaluation Data

Languages. The evaluation covers 32 language pairs, with each test set containing approximately

WMT’25 Automatic Eval.

Automatic Ranking

This section details our automatic ranking method, which we refer to as AUTORANK. Both the set of automatic metrics and the aggregation procedure have been slightly updated since last year’s shared task.

Metrics. For most language pairs,[3] the AUTORANK is a combination of three distinct families of evaluation methods:

- **LLM-as-a-Judge (reference-less).** We use GEMBA-ESA (Kocmi and Federmann, 2023) with two independent judges: GPT-4.1 (OpenAI, 2025) and Command A (Cohere Team, 2025), both in a reference-less setting.
- **Trained reference-based metrics.** Two supervised metrics trained to approximate human quality judgments with references: MetricX-24-Hybrid-XL (Juraska et al., 2024) and XCOMET-XL (Guerreiro et al., 2024).
- **Trained Quality Estimation (QE).** The reference-less QE metric CometKiwi-XL (Rei et al., 2023), which is also trained to mimic human judgments.

This combination of reference-based and reference-less (or QE) methods is designed to balance their complementary failure modes. Reference-based metrics typically achieve a higher correlation with human judgments when high-quality references are available, while reference-less methods reduce susceptibility to reference

[3] See “Low-resource exception” below.

- LLM-as-a-judge (reference-free): GEMBA-ESA, GPT-4.1, Command A
- Trained reference-based metrics: MetricX, XCOMET
- Trained Quality Estimation (QE): reference-less CometKiwi

GEMBA-ESA

- MQM simplified into Error Span Annotation (ESA), which focuses on the error span severities and not the actual error types.

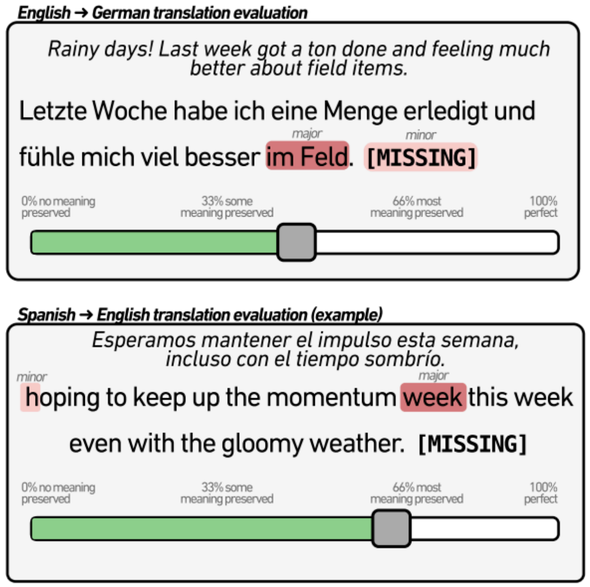


Figure 1: Stylized annotation user interface with Error Span Annotation (ESA). The annotator first marks errors with minor and major severity and then assigns a final score. This is more robust than asking for score directly.[1]

Kocmi et al. (2024)

GEMBA-ESA

- MQM simplified into Error Span Annotation (ESA), which focuses on the error span severities and not the actual error types.

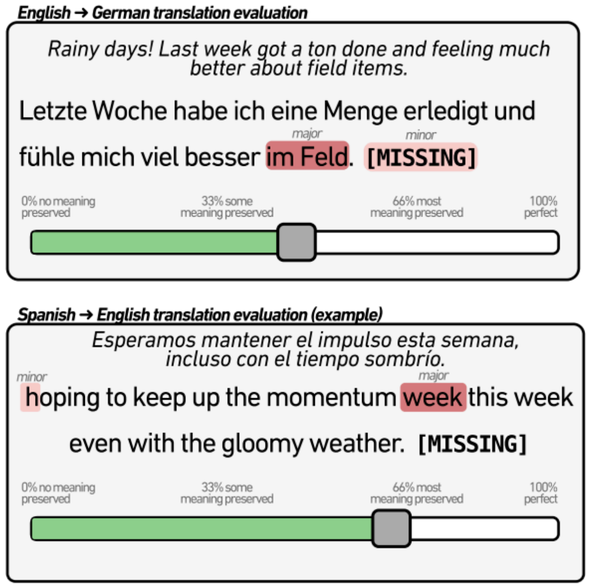


Figure 1: Stylized annotation user interface with Error Span Annotation (ESA). The annotator first marks errors with minor and major severity and then assigns a final score. This is more robust than asking for score directly.[1]

Kocmi et al. (2024)

- Few-shot prompt of GPT-4 as the judge

```
(System) You are an annotator for the quality of machine translation. Your task is to identify
errors and assess the quality of the translation.

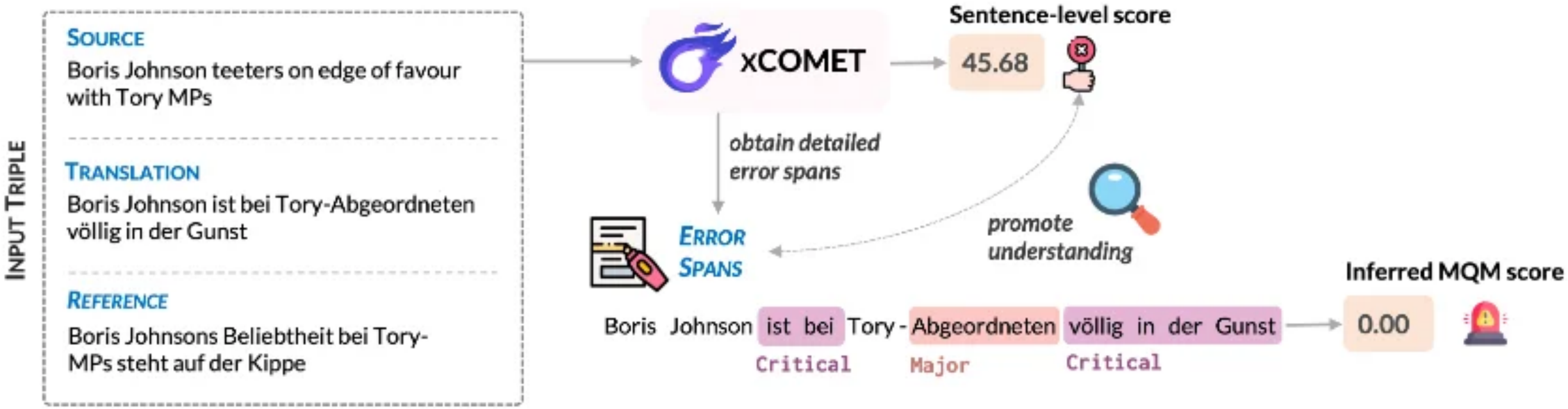
(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify
error types in the translation and classify them. The categories of errors are: accuracy
(addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar,
inconsistency, punctuation, register, spelling),
locale convention (currency, date, name, telephone, or time format)
style (awkward), terminology (inappropriate  for context, inconsistent use), non-translation,
other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what
the text is trying to say is still understandable. Minor errors are technically errors,
but do not disrupt the flow or hinder comprehension.

(assistant) {observed error classes}
```

Figure 1: The general prompt for GEMBA-MQM omits the gray part which performed subpar on internal data (we include it in GEMBA-locale-MQM). The "(user)" and "(assistant)" section is repeated for each few-shot example.

Kocmi & Federmann (2023)

XCOMET



Guerreiro et al. (2024)

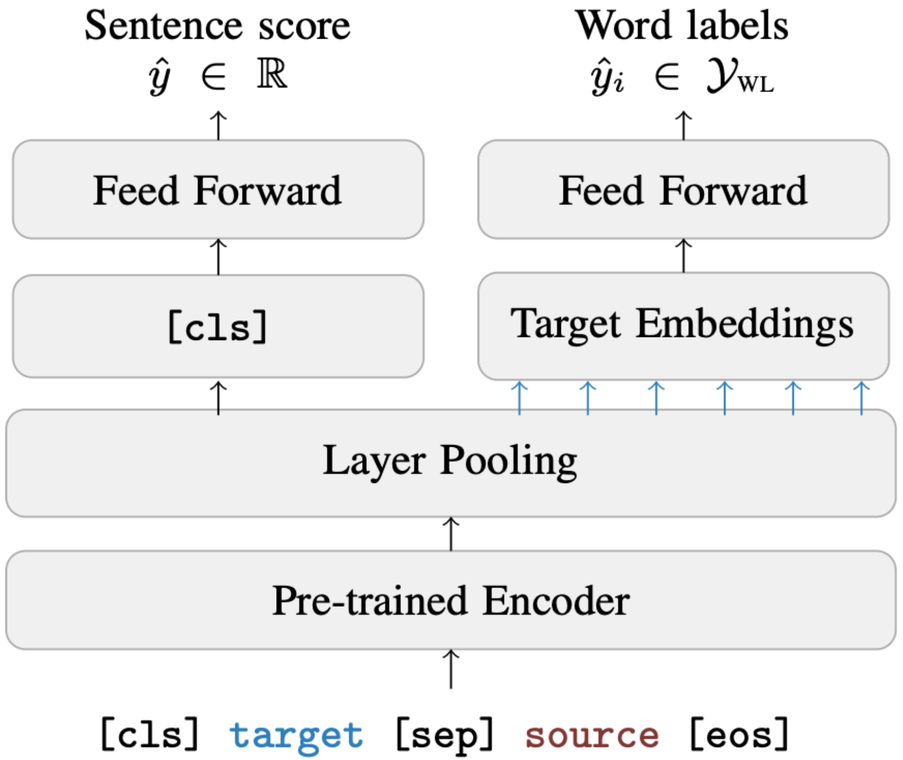
COMETKIWI - Learnt Metric

- Learn from both sentence-level and word-level quality estimations
- Use a (trainable) weighted sum of the hidden states of each layer of the encoder

$$\mathcal{L}_{\text{SL}}(\theta) = \frac{1}{2}(y - \hat{y}(\theta))^2$$

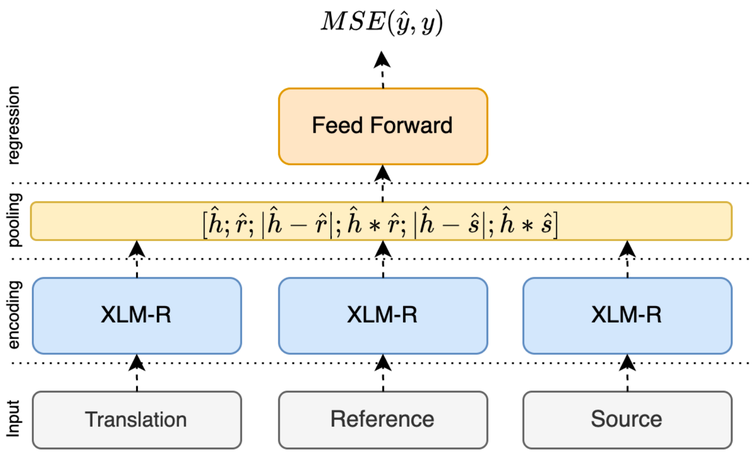
$$\mathcal{L}_{\text{WL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_{y_i} \log p_\theta(y_i)$$

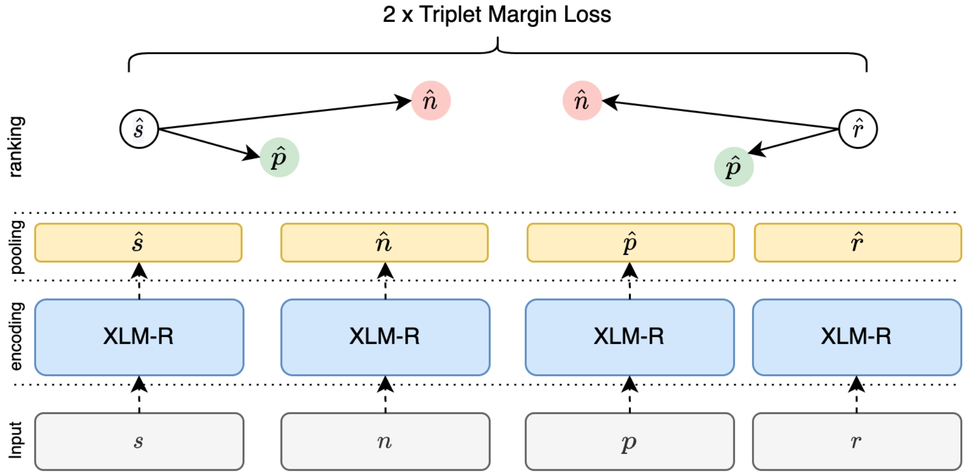
$$\mathcal{L}(\theta) = \lambda_{\text{SL}}\mathcal{L}_{\text{SL}}(\theta) + \lambda_{\text{WL}}\mathcal{L}_{\text{WL}}(\theta),$$

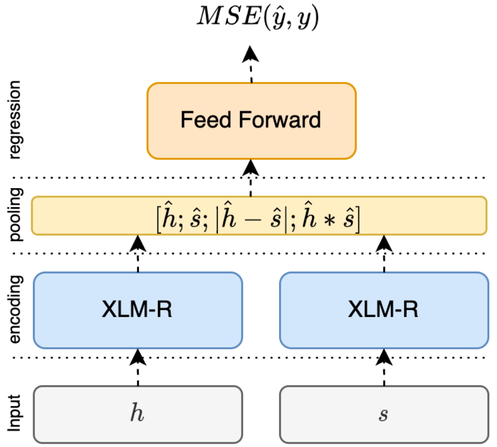


Rei et al. (2022)

COMET - Learnt Metric







- Regression Metric (left): trained on a regression task using source, MT and reference; Ranking Metric (middle): optimize to encode good translations closer to the anchors (source, reference) while pushing bad translations away; Reference-less Metric (right): does not use the reference translation.

Rei et al. (2020)

WMT’25 Preliminary Results

English-Egyptian Arabic

System Name	LP Supported	Params. (B)	Humeval?	AutoRank ↓	CometKiwi-XL ↑	GEMBA-ESA-CMDA ↑	GEMBA-ESA-GPT4.1 ↑	MetricX-24-Hybrid-XL ↑	XCOMET-XL ↑
Shy-hunyuan-MT	✓	7	✓	1.0	0.658	76.3	75.0	-5.7	0.388
Wenyiil	✓	14	✓	2.5	0.65	79.2	73.3	-6.4	0.337
Algharb	✓	14	✓	2.6	0.645	80.0	73.9	-6.5	0.328
GemTrans	✓	27	✓	3.4	0.644	73.0	69.6	-6.0	0.345
CommandA-WMT	✓	111	✓	4.0	0.621	77.8	75.4	-7.0	0.311
UvA-MT	✓	12	✓	4.1	0.637	74.4	73.4	-7.1	0.325
Yolu	✓	14	✓	5.4	0.658	67.8	63.9	-6.6	0.323
▲ Gemini-2.5-Pro	✓	?	✓	5.6	0.552	79.5	84.5	-7.6	0.267
▲ ONLINE-B	✓	?	✓	6.4	0.627	70.4	67.4	-7.1	0.288
▲ GPT-4.1	✓	?	✓	6.5	0.534	78.4	84.1	-7.8	0.265
▲ DeepSeek-V3	?	671	✓	6.9	0.573	74.2	75.7	-7.7	0.273
▲ Mistral-Medium	✓	?	✓	7.5	0.586	71.7	71.0	-7.8	0.274
▲ Claude-4	✓	?	✓	7.6	0.552	76.5	80.0	-8.5	0.246
SRPOL	✗	12	✓	7.9	0.641	65.7	61.7	-7.8	0.286
▲ CommandA	✓	111	✓	8.3	0.533	75.8	80.0	-8.5	0.238
▲ AyaExpanse-32B	✓	32		8.4	0.585	70.7	68.8	-8.1	0.261
▲ ONLINE-W	?	?		9.0	0.607	67.7	64.0	-8.2	0.258
▲ AyaExpanse-8B	✓	8	✓	9.7	0.596	66.1	61.6	-8.2	0.259
▲ Qwen3-235B	✓	235		10.7	0.571	66.1	64.1	-8.7	0.247
▲ Gemma-3-27B	✓	27		10.7	0.549	64.8	63.3	-8.6	0.281
▲ EuroLLM-22B-pre.[M]	✓	22		10.7	0.592	64.0	60.5	-8.5	0.246
IRB-MT	✓	12	✓	10.8	0.532	69.0	67.5	-8.5	0.236
▲ Llama-4-Maverick	✓	400		11.1	0.526	67.9	70.0	-8.8	0.234
IR-MultiagentMT	✗	?		11.3	0.543	66.0	64.2	-8.7	0.247
▲ CommandR7B	✓	7	✓	11.3	0.588	62.7	59.0	-8.8	0.248
▲ Gemma-3-12B	✓	12		11.7	0.529	67.9	67.6	-9.0	0.22
▲ EuroLLM-9B[M]	✓	9		14.0	0.548	58.7	54.5	-9.3	0.233
▲ TowerPlus-72B[M]	✗	72		15.5	0.534	58.2	54.0	-10.5	0.224
TranssionTranslate	?	?		15.8	0.501	59.0	57.4	-9.9	0.2
TranssionMT	✓	1		16.9	0.488	58.5	56.1	-10.4	0.194
▲ NLLB	✓	1		18.0	0.499	53.9	51.2	-10.8	0.201
SalamandraTA	✓	8		20.1	0.492	50.0	44.4	-11.4	0.195
▲ ONLINE-G	✓	?		22.6	0.445	53.5	48.3	-13.5	0.152
▲ Llama-3.1-8B	✗	8		22.8	0.458	45.5	41.8	-12.3	0.18
▲ Qwen2.5-7B	✓	7		24.0	0.436	44.5	39.3	-12.6	0.176
▲ TowerPlus-9B[M]	✗	9		31.9	0.337	31.1	26.9	-15.2	0.162
▲ Mistral-7B	✗	7		37.0	0.262	27.9	23.2	-18.4	0.157

English-Bhojpuri

System Name	LP Supported	Params. (B)	Humeval?	AutoRank ↓	chrF++ ↑
▲ Gemini-2.5-Pro	✓	?	✓	1.0	40.6
Wenyiil	✓	14	✓	2.5	38.9
Algharb	✓	14	✓	2.8	38.6
▲ ONLINE-B	✓	?	✓	4.1	37.1
TranssionTranslate	?	?	✓	4.4	36.9
▲ Claude-4	?	?	✓	4.5	36.7
▲ DeepSeek-V3	?	671	✓	5.1	36.0
▲ GPT-4.1	?	?	✓	5.5	35.6
Yolu	✓	14	✓	5.6	35.4
TranssionMT	✓	1	✓	6.2	34.8
▲ Llama-4-Maverick	✓	400	✓	6.5	34.4
▲ CommandA	✗	111	✓	6.5	34.4
▲ NLLB	✓	1	✓	6.6	34.3
▲ Gemma-3-27B	?	27	✓	8.3	32.4
CommandA-WMT	✗	111		8.8	31.8
COILD-BHO	✓	7	✓	8.9	31.8
▲ Mistral-Medium	?	?		9.0	31.6
▲ Qwen3-235B	✗	235		11.1	29.2
IRB-MT	✓	12	✓	11.4	28.9
▲ AyaExpanse-32B	✗	32		11.4	28.9
Shy-hunyuan-MT	✓	7	✓	11.5	28.8
GemTrans	✓	27		11.9	28.3
SalamandraTA	✓	8	✓	12.1	28.2
▲ Gemma-3-12B	?	12		12.3	27.9
▲ TowerPlus-9B[M]	✗	9		12.7	27.4
▲ TowerPlus-72B[M]	✗	72		12.8	27.3
▲ EuroLLM-22B-pre.[M]	✗	22		13.6	26.4
▲ EuroLLM-9B[M]	✗	9		14.7	25.2
IR-MultiagentMT	✗	?		15.9	23.9
▲ CommandR7B	✗	7		16.7	22.9
▲ AyaExpanse-8B	✗	8		16.7	22.9
▲ Qwen2.5-7B	?	7		17.7	21.8
▲ Mistral-7B	✗	7		20.9	18.2
UvA-MT	✓	12		28.4	9.7
▲ Llama-3.1-8B	✗	8		35.0	2.3

Hunyuan-MT

Tencent Hunyuan

2025-09-10

Hunyuan-MT Technical Report

Tencent Hunyuan Team

Abstract

In this report, we introduce **Hunyuan-MT-7B**, our first open-source multilingual translation model, which supports bidirectional translation across 33 major languages and places a special emphasis on translation between Mandarin and several ethnic minority languages as well as dialects. Furthermore, to serve and address diverse translation scenarios and enhance model performance at test time, we introduce **Hunyuan-MT-Chimera-7B**, a translation model inspired by the *slow thinking* mode. This model integrates multiple outputs generated by the **Hunyuan-MT-7B** model under varying parameter settings, thereby achieving performance superior to that of conventional *slow-thinking* models based on Chain-of-Thought (CoT). The development of our models follows a holistic training process specifically engineered for multilingual translation, which begins with general and MT-oriented pre-training to build foundational capabilities, proceeds to Supervised Fine-Tuning (SFT) for task-specific adaptation, and culminates in advanced alignment through Reinforcement Learning (RL) and weak-to-strong RL. Through comprehensive experimentation, we demonstrate that both **Hunyuan-MT-7B** and **Hunyuan-MT-Chimera-7B** significantly outperform all translation-specific models of comparable parameter size and most of the SOTA large models, particularly on the task of translation between Mandarin and minority languages as well as dialects.

In the WMT2025 shared task (General Machine Translation), our models demonstrate state-of-the-art performance, ranking first in 30 out of 31 language pairs. This result highlights the robustness of our models across a diverse linguistic spectrum, encompassing high-resource languages such as Chinese, English, and Japanese, as well as low-resource languages including Czech, Marathi, Estonian, and Icelandic.

Hunyuan-MT-7B: https://huggingface.co/tencent/Hunyuan-MT-7B

Hunyuan-MT-Chimera-7B: https://huggingface.co/tencent/Hunyuan-MT-Chimera-7B

Code Repository: https://github.com/Tencent-Hunyuan/Hunyuan-MT

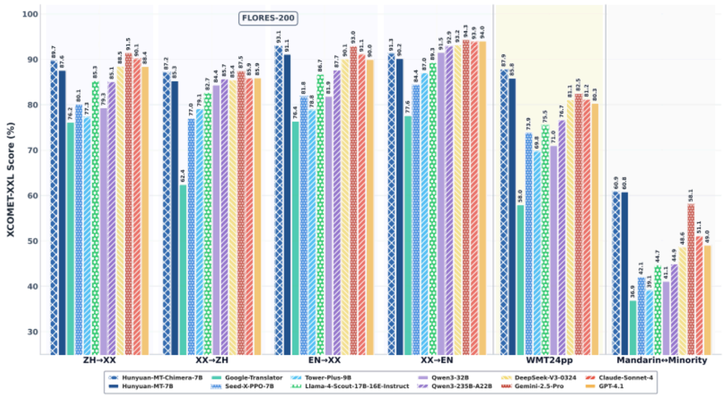


Figure 1: Benchmark performance of Hunyuan-MT models and state-of-the-art baselines.

Hunyuan-MT

Tencent Hunyuan

2025-09-10

Hunyuan-MT Technical Report

Tencent Hunyuan Team

Abstract

In this report, we introduce **Hunyuan-MT-7B**, our first open-source multilingual translation model, which supports bidirectional translation across 33 major languages and places a special emphasis on translation between Mandarin and several ethnic minority languages as well as dialects. Furthermore, to serve and address diverse translation scenarios and enhance model performance at test time, we introduce **Hunyuan-MT-Chimera-7B**, a translation model inspired by the *slow thinking* mode. This model integrates multiple outputs generated by the **Hunyuan-MT-7B** model under varying parameter settings, thereby achieving performance superior to that of conventional *slow-thinking* models based on Chain-of-Thought (CoT). The development of our models follows a holistic training process specifically engineered for multilingual translation, which begins with general and MT-oriented pre-training to build foundational capabilities, proceeds to Supervised Fine-Tuning (SFT) for task-specific adaptation, and culminates in advanced alignment through Reinforcement Learning (RL) and weak-to-strong RL. Through comprehensive experimentation, we demonstrate that both **Hunyuan-MT-7B** and **Hunyuan-MT-Chimera-7B** significantly outperform all translation-specific models of comparable parameter size and most of the SOTA large models, particularly on the task of translation between Mandarin and minority languages as well as dialects.

In the WMT2025 shared task (General Machine Translation), our models demonstrate state-of-the-art performance, ranking first in 30 out of 31 language pairs. This result highlights the robustness of our models across a diverse linguistic spectrum, encompassing high-resource languages such as Chinese, English, and Japanese, as well as low-resource languages including Czech, Marathi, Estonian, and Icelandic.

Hunyuan-MT-7B: https://huggingface.co/tencent/Hunyuan-MT-7B

Hunyuan-MT-Chimera-7B: https://huggingface.co/tencent/Hunyuan-MT-Chimera-7B

Code Repository: https://github.com/Tencent-Hunyuan/Hunyuan-MT

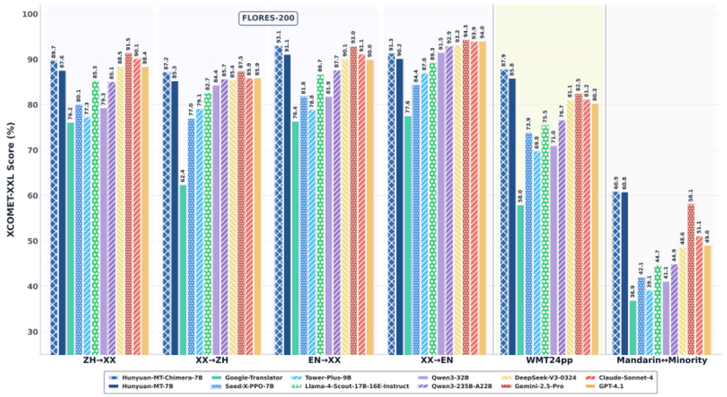


Figure 1: Benchmark performance of Hunyuan-MT models and state-of-the-art baselines.

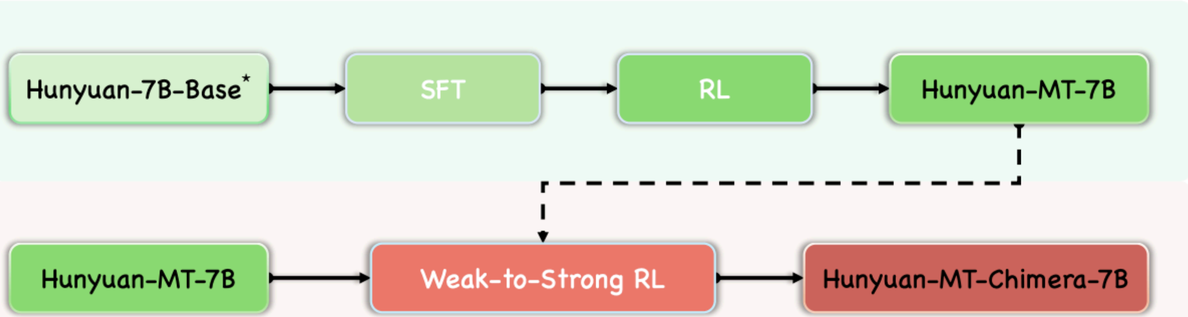


Figure 2: Post-training pipeline of the Hunyuan-MT-7B and Hunyuan-MT-Chimera-7B models.

Prompt Template for Hunyuan-MT-Chimera-7B.

Analyze the following multiple <target_language> translations of the <source_language> segment surrounded in triple backticks and generate a single refined <target_language> translation. Only output the refined translation, do not explain.

The <source_language> segment:
```<source_text>```

The multiple <target_language> translations:
1. ```<translated_text1>```
2. ```<translated_text2>```
3. ```<translated_text3>```
4. ```<translated_text4>```
5. ```<translated_text5>```
6. ```<translated_text6>```
```

Evaluating Robustness of Large Language Models with Neologisms (NeoBench)

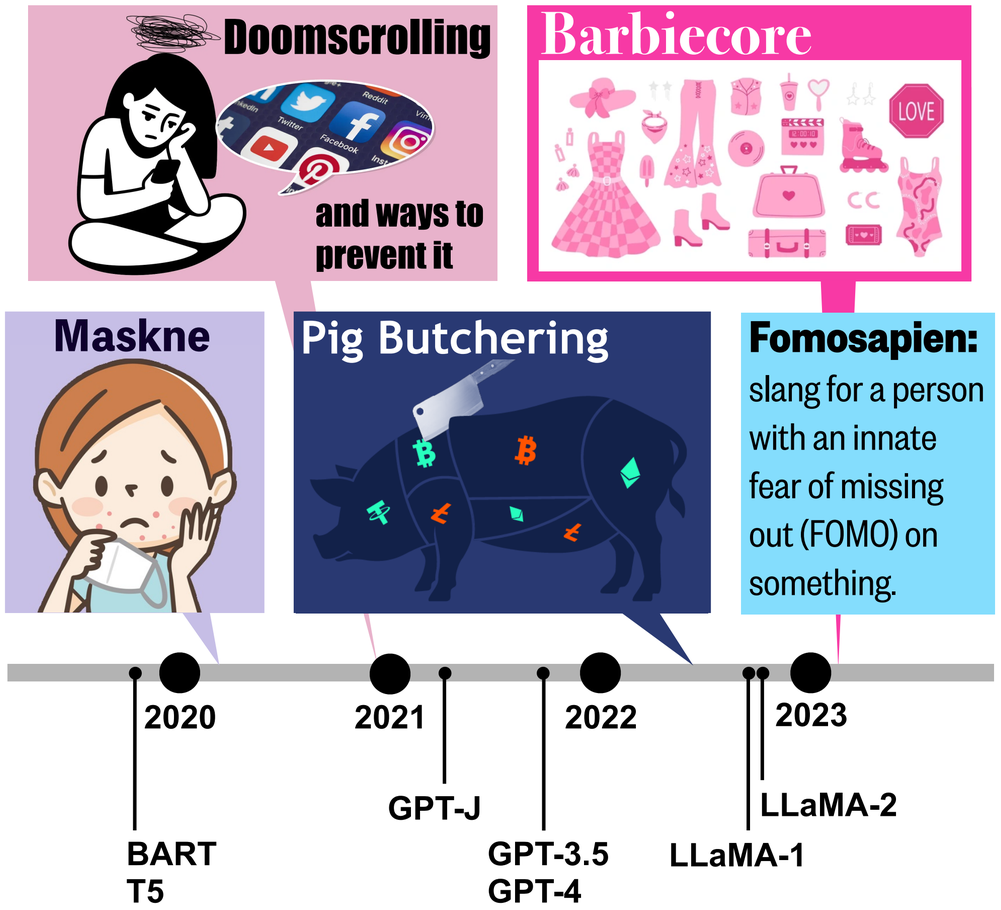
Jonathan Zheng

Alan Ritter

Wei Xu

NeoBench — evolving human languages

Data contamination, long-tail low-frequency words, tokenization, …



We used 3 different methods to obtain 2,505 single- and multi-word neologisms.

NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.

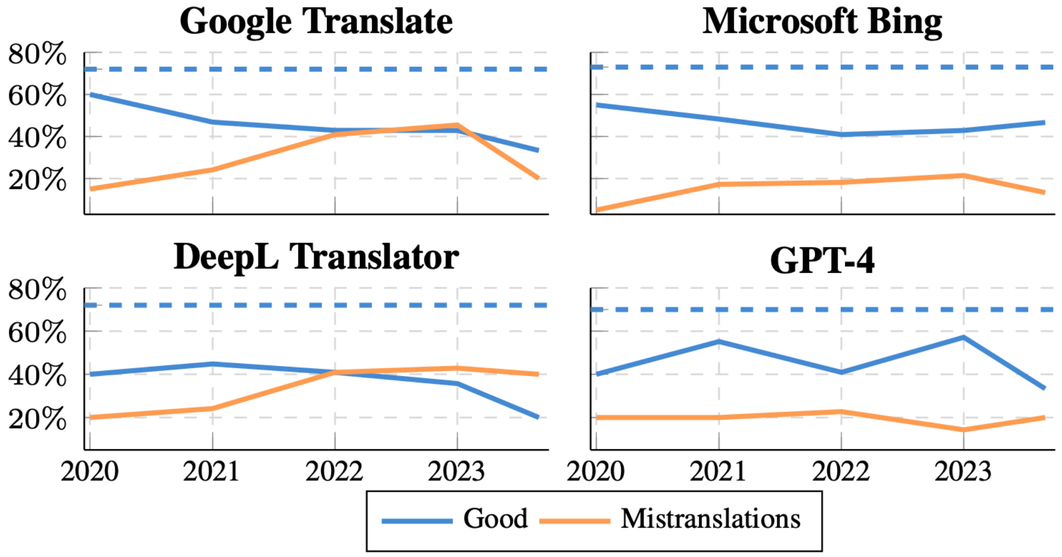


Figure 3: Percentage of good translations and mistranslations of neologism sentences over time. The dashed line represents the percentage of good translations achieved on non-neologism sentences.

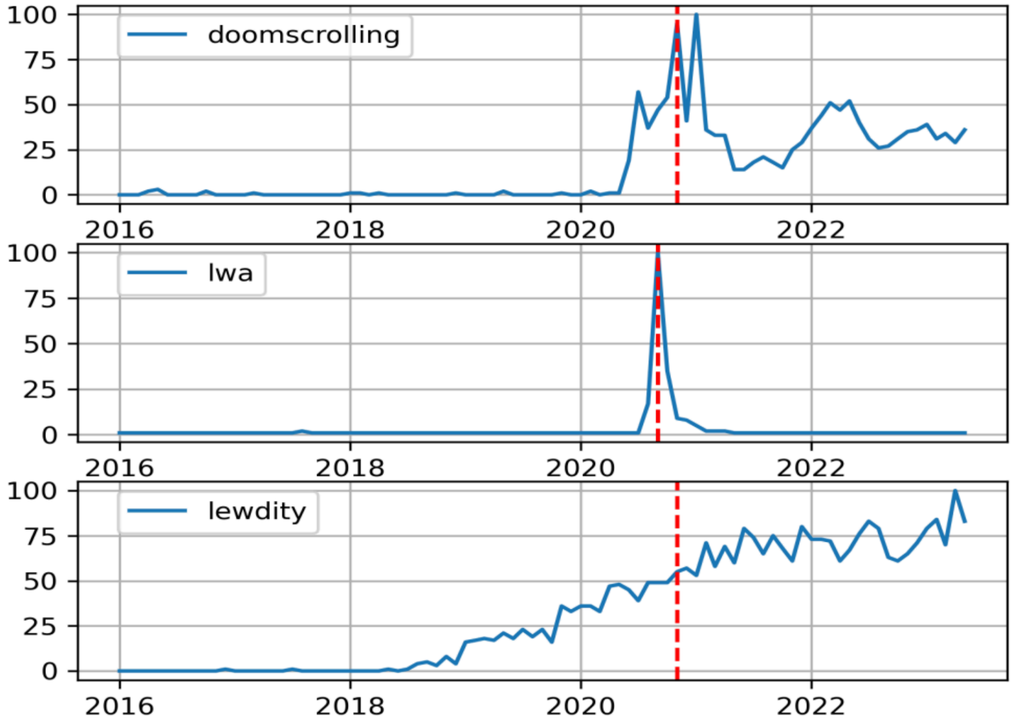


Figure 4: Example Google Trend lines measuring neologism prevalence. The dashed line estimates the date a neologism becomes popular while not yet conventional.

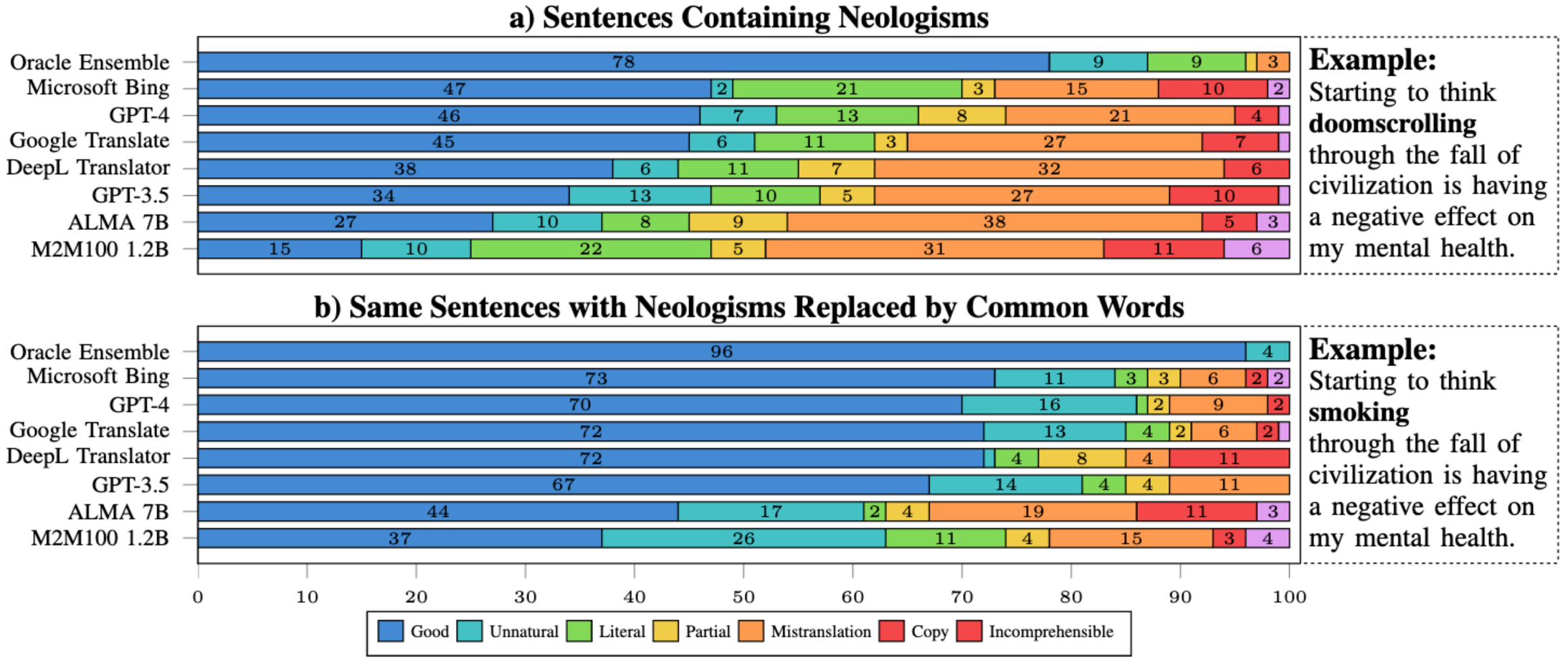
NeoBench — perplexity, Cloze QA, definition

a) Definition Generation Output Examples	
Stablecoin	**Reference Definition:** *A stablecoin is a type of cryptocurrency where the value of the digital asset is supposed to be pegged to a reference asset, which is either fiat money, exchange-traded commodities, or another cryptocurrency.*
	Model Output (Correct): *Stablecoins are cryptocurrencies designed to maintain a stable value, typically by pegging their value to a specific asset or basket of assets, such as the US dollar, gold, or a combination of assets.*
Angel Shot	**Reference Definition:** *An angel shot is a code to inform a bartender that a customer is not safe and needs assistance.*
	Model Output (Incorrect): *An angel shot is a cocktail made with whiskey and cream, served in a shot glass.*
b) Machine Translation Output Examples	
Longcovid	**Input:** Each reinfection increases the risk of **longcovid**, hospitalization, & death.
	Model Output (Correct): 每次再感染都会增加长新冠病毒、住院和死亡的风险。 *(Every reinfection increases the risk of **long COVID**, hospitalization, and death.)*
	Human Translation: 每一次新冠感染都会提高出现后遗症、住院治疗，甚至死亡的风险。 *(Each COVID-19 infection increases the risk of developing **sequelae**, hospitalization, and even death.)*
Doomscrolling	**Input:** Starting to think **doomscrolling** through the fall of civilization is having a negative effect on my mental health.
	Model Output (Incorrect): 开始认为在文明的衰落中滚动的厄运对我的心理健康产生了负面影响。 *(Start to think that the **doom rolling** in the decline of civilization is having a negative impact on my mental health.)*
	Human Translation: 开始觉得, 刷关于文明衰败的负能量新闻对我的心理健康产生了负面影响。 *(Starting to feel that **scrolling** through negative news about the decline of civilization is having a negative impact on my mental health.)*

Table 3: Example model definitions and translations for NEO-BENCH tasks. *"Doomscrolling"* is the act of spending an excessive amount of time reading negative news online. (English translations are shown for information only.)

NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.



NeoBench — automatic eval on translation

Automatic evaluation metrics do not show good system-level correlations with human evaluation.

Model (human rank)		**Reference-Based Metrics**				**Reference-Free Metrics**		
		BLEU↑	**COMET**↑	**MX-23**$_{XXL}$↓	**MX-23**$_{XL}$↓	**COMET**$_{KIWI}$↑	**MX-QE**$_{XXL}$↓	**MX-QE**$_{XL}$↓
Bing Translator	(1)	0.452 (2)	0.825 (5)	2.419 (6)	2.343 (6)	0.788 (5)	1.679 (5)	2.246 (5)
GPT-4	(2)	0.446 (3)	**0.854** (1)	**1.550** (1)	**1.793** (1)	0.793 (3)	1.432 (3)	2.089 (3)
Google Translate	(3)	**0.507 (1)**	0.853 (2)	1.825 (4)	1.945 (4)	0.800 (2)	1.429 (2)	**1.940** (1)
DeepL Translator	(4)	0.406 (4)	0.842 (3)	1.775 (3)	1.901 (3)	**0.807** (1)	**1.260** (1)	1.944 (2)
GPT-3.5	(5)	0.399 (5)	0.841 (4)	1.705 (2)	1.796 (2)	0.792 (4)	1.467 (4)	2.157 (4)
ALMA 7B (LLaMA-2)	(6)	0.285 (7)	0.801 (6)	2.382 (5)	2.251 (5)	0.746 (6)	2.038 (6)	2.462 (6)
M2M100 1.2B	(7)	0.337 (6)	0.776 (7)	3.454 (7)	3.142 (7)	0.745 (7)	2.821 (7)	2.976 (7)
Spearman's ρ		0.244	0.445	0.457	0.380	**0.491**	0.451	0.445

Evaluating Robustness of Large Language Models with Neologisms (NeoBench)

Jonathan Zheng

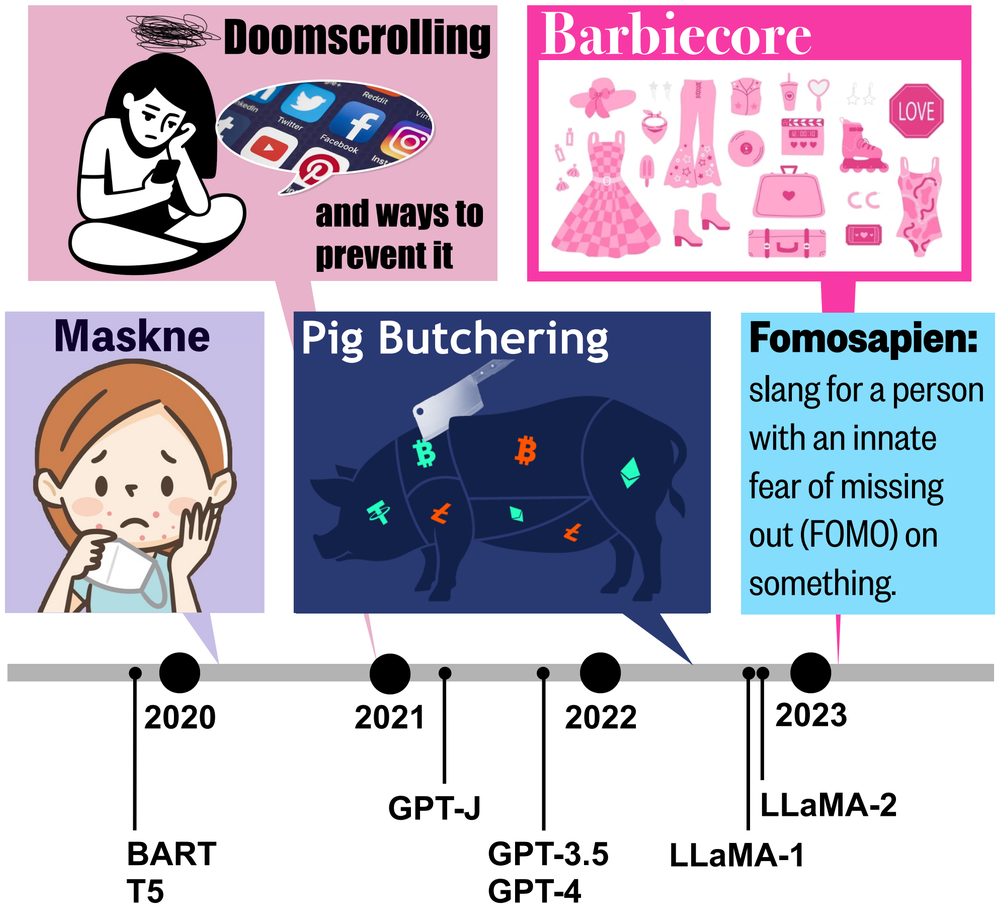
Alan Ritter

Wei Xu

A better technical solution for marker-based label projection

NeoBench — evolving human languages

Data contamination, long-tail low-frequency words, tokenization, …



We used 3 different methods to obtain 2,505 single- and multi-word neologisms.

NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.

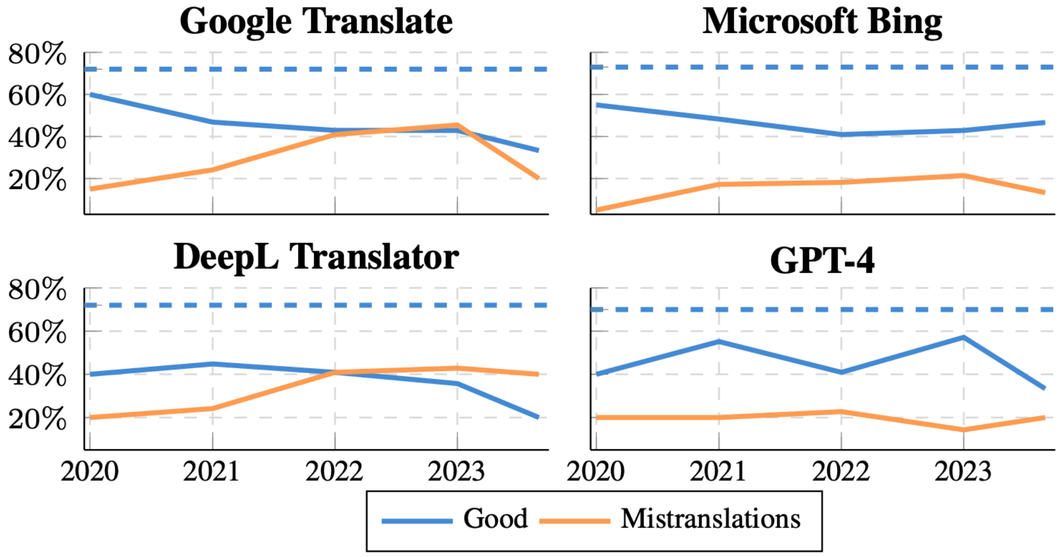


Figure 3: Percentage of good translations and mistranslations of neologism sentences over time. The dashed line represents the percentage of good translations achieved on non-neologism sentences.

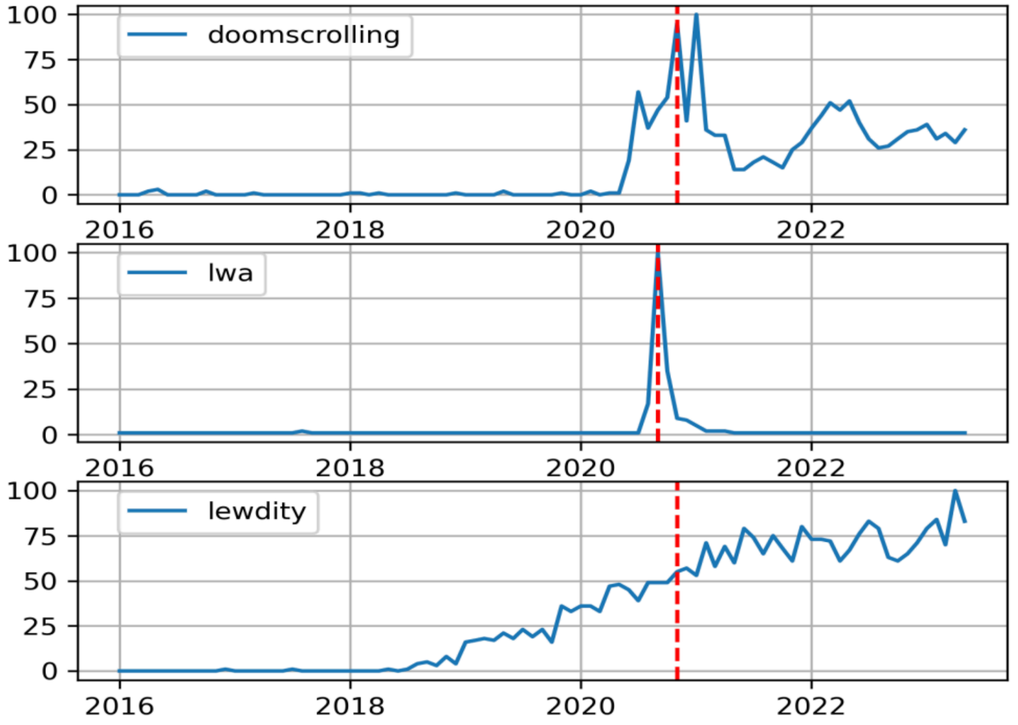
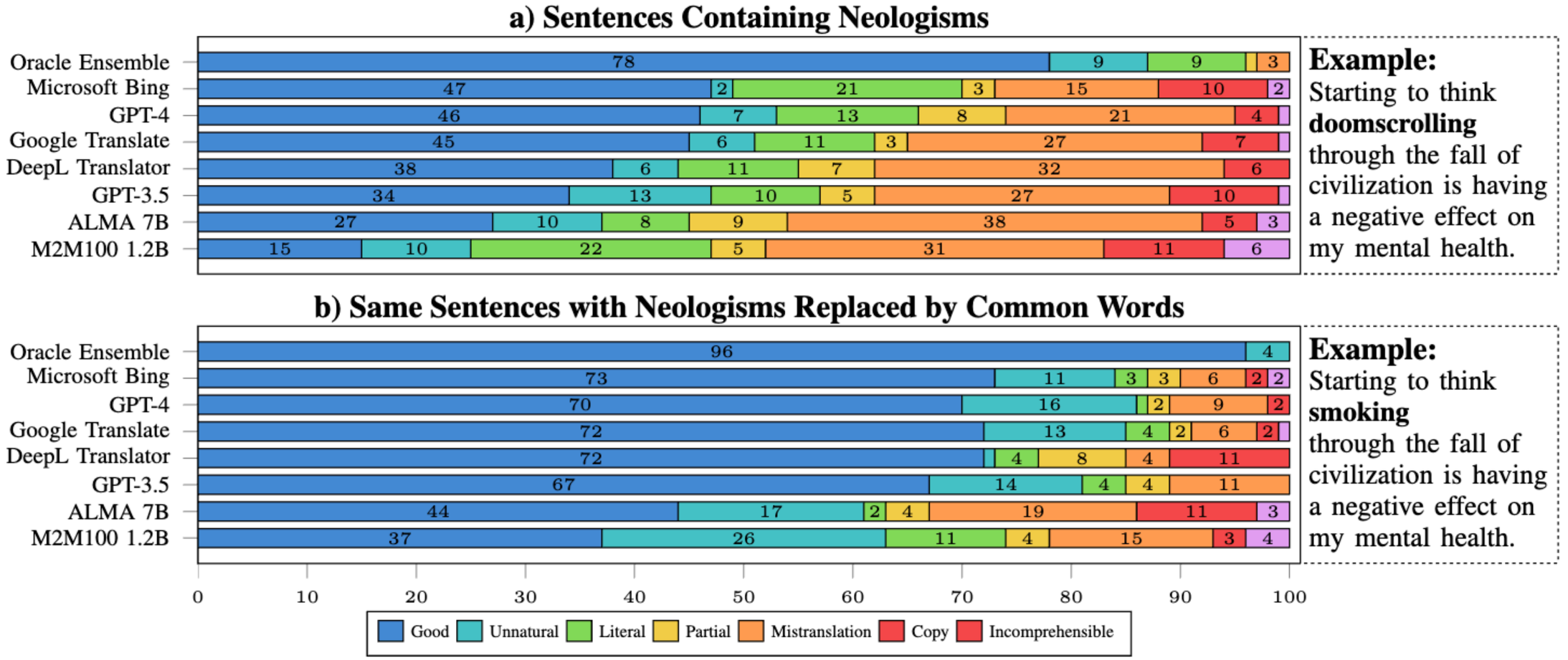


Figure 4: Example Google Trend lines measuring neologism prevalence. The dashed line estimates the date a neologism becomes popular while not yet conventional.

NeoBench — human evaluation on translation

Models struggle to translate sentences that contains neologism (vs. non-neologism) word.



NeoBench — perplexity, Cloze QA, definition

a) Definition Generation Output Examples	
Stablecoin	**Reference Definition:** *A stablecoin is a type of cryptocurrency where the value of the digital asset is supposed to be pegged to a reference asset, which is either fiat money, exchange-traded commodities, or another cryptocurrency.*
	Model Output (Correct): *Stablecoins are cryptocurrencies designed to maintain a stable value, typically by pegging their value to a specific asset or basket of assets, such as the US dollar, gold, or a combination of assets.*
Angel Shot	**Reference Definition:** *An angel shot is a code to inform a bartender that a customer is not safe and needs assistance.*
	Model Output (Incorrect): *An angel shot is a cocktail made with whiskey and cream, served in a shot glass.*
b) Machine Translation Output Examples	
Longcovid	**Input:** Each reinfection increases the risk of **longcovid**, hospitalization, & death.
	Model Output (Correct): 每次再感染都会增加长新冠病毒、住院和死亡的风险。 *(Every reinfection increases the risk of **long COVID**, hospitalization, and death.)*
	Human Translation: 每一次新冠感染都会提高出现后遗症、住院治疗，甚至死亡的风险。 *(Each COVID-19 infection increases the risk of developing **sequelae**, hospitalization, and even death.)*
Doomscrolling	**Input:** Starting to think **doomscrolling** through the fall of civilization is having a negative effect on my mental health.
	Model Output (Incorrect): 开始认为在文明的衰落中滚动的厄运对我的心理健康产生了负面影响。 *(Start to think that the **doom rolling** in the decline of civilization is having a negative impact on my mental health.)*
	Human Translation: 开始觉得, 刷关于文明衰败的负能量新闻对我的心理健康产生了负面影响。 *(Starting to feel that **scrolling** through negative news about the decline of civilization is having a negative impact on my mental health.)*

Table 3: Example model definitions and translations for NEO-BENCH tasks. *"Doomscrolling"* is the act of spending an excessive amount of time reading negative news online. (English translations are shown for information only.)

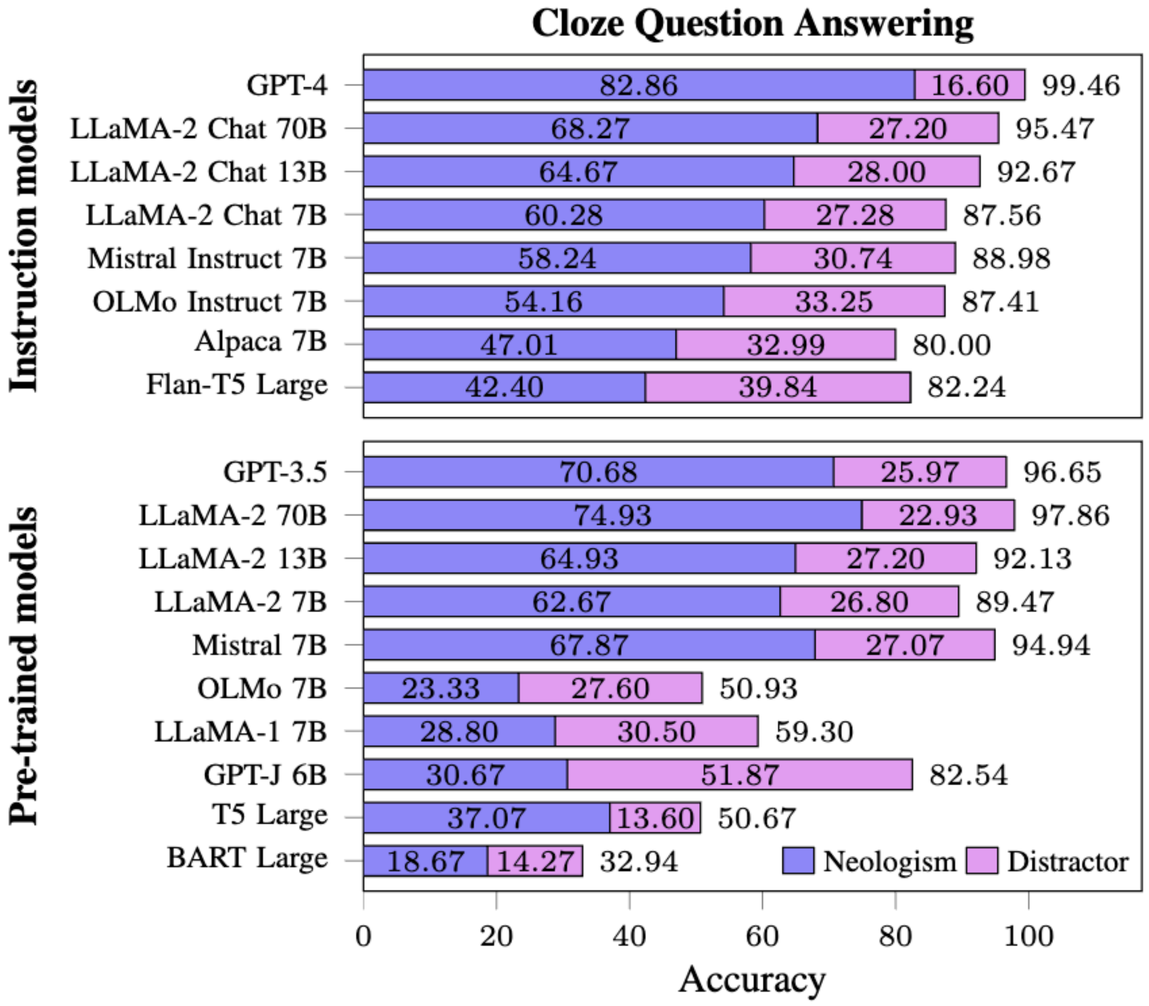
NeoBench — automatic eval on translation

Automatic evaluation metrics do not show good system-level correlations with human evaluation.

Model (human rank)		**Reference-Based Metrics**				**Reference-Free Metrics**		
		BLEU↑	**COMET**↑	**MX-23**$_{XXL}$↓	**MX-23**$_{XL}$↓	**COMET**$_{KIWI}$↑	**MX-QE**$_{XXL}$↓	**MX-QE**$_{XL}$↓
Bing Translator	(1)	0.452 (2)	0.825 (5)	2.419 (6)	2.343 (6)	0.788 (5)	1.679 (5)	2.246 (5)
GPT-4	(2)	0.446 (3)	**0.854** (1)	**1.550** (1)	**1.793** (1)	0.793 (3)	1.432 (3)	2.089 (3)
Google Translate	(3)	**0.507 (1)**	0.853 (2)	1.825 (4)	1.945 (4)	0.800 (2)	1.429 (2)	**1.940** (1)
DeepL Translator	(4)	0.406 (4)	0.842 (3)	1.775 (3)	1.901 (3)	**0.807** (1)	**1.260** (1)	1.944 (2)
GPT-3.5	(5)	0.399 (5)	0.841 (4)	1.705 (2)	1.796 (2)	0.792 (4)	1.467 (4)	2.157 (4)
ALMA 7B (LLaMA-2)	(6)	0.285 (7)	0.801 (6)	2.382 (5)	2.251 (5)	0.746 (6)	2.038 (6)	2.462 (6)
M2M100 1.2B	(7)	0.337 (6)	0.776 (7)	3.454 (7)	3.142 (7)	0.745 (7)	2.821 (7)	2.976 (7)
Spearman's ρ		0.244	0.445	0.457	0.380	**0.491**	0.451	0.445

NeoBench — perplexity, Cloze QA, definition

Newer, larger LLMs work better; but, perplexity becomes worse after instruction tuning.



Neologism: *doomscrolling*	
The silver lining of this website no longer functioning as an even vaguely reliable information source is that ___ has basically been completely undermined. It wouldn't even work now since everything is too geared to outrage clickbait and actual reporting has disappeared, so there is no point staying on the app.	
a) misinformation	*b) surfing*
c) doomscrolling	*d) lying*
e) gaming	***f) anti-productivity (distractor)***

Table 2: Example passage in NEO-BENCH for multiple-choice Cloze Question Answering with correct neologism answers and partially correct distractor answers.

NeoBench — perplexity, Cloze QA, definition

Newer, larger LLMs work better; but, perplexity becomes worse after instruction tuning.