

CUNI at the WMT26 Automated MT Evaluation Shared Task

Miroslav Hrabal, Ondřej Bojar

Charles University, Faculty of Mathematics and Physics
{hrabal,bojar}@ufal.mff.cuni.cz

Abstract

We describe the CUNI submissions to the WMT26 Automated MT Evaluation Shared Task. We take part in error-span prediction (Task 1) and segment-level quality score prediction (Task 2). Our systems use a single open-weight Gemma 4 31B instruction-tuned model (Team et al., 2026) as a backbone. For Task 1, we decompose error prediction into semantic error detection, self-review, and a separate span-alignment step that maps free-form error descriptions to source and target character spans. We submit two variants of this approach, one additionally predicting MQM-style error categories. For Task 2, we explore two complementary approaches: regression over features extracted from the Task 1 predictions and LLM-based direct assessment on a 0–100 scale.

1 Introduction

A particular difficulty of using large language models for span prediction is that identifying an error and expressing its exact location are not necessarily the same problem. Even when a model is able to describe translation errors in natural language, producing exact offsets or reproducing the corresponding substring can still be a challenge due to failures to copy spans exactly or difficulty counting characters directly (Semin et al., 2026).

We therefore try to explicitly separate these operations. Our Task 1 systems identify and describe translation errors without imposing strict constraints on representing spans, allow the model to review and modify its analysis, and finally align the resulting errors to a tokenized representation of source and translation. The predicted token spans are algorithmically converted to the character offsets required by the shared task.

For Task 2, we investigate two different approaches. The first derives numerical features from Task 1 error predictions and learns a regression

model over them. The second directly asks the language model to assess translation quality on a 0 to 100 integer scale using few-shot demonstrations.

In addition to evaluating the resulting systems on subsets of WMT25 data (Kocmi et al., 2025), we examine how well the same prompting architectures transfer to alternative backbone LLMs without model-specific adaptation.

2 Related Work

The difficulty of obtaining exact input spans from generative language models is not specific to MT evaluation. Semin et al. (2026) categorize LLM span-labeling strategies into input tagging, numerical position indexing, and span-content matching. They show that seemingly simple content-matching approaches suffer from alignment failures and propose constrained decoding as one solution. Our Task 1 pipeline uses a different strategy: the model is first allowed to describe errors without localization, after which a separate model call maps the detected errors to a token-indexed representation of the input. Character offsets are then recovered deterministically.

3 Shared Framework

All of our submissions use a single open-weight Gemma 4 31B instruction-tuned model with 4-bit AWQ quantization (Lin et al., 2026)¹ as a backbone. We fixed this model choice early in development based on our preliminary multilingual experiments with models of similar size.

We implement the model interactions using DSPy (Khatab et al., 2022, 2024) for prompting the model. All our systems are reference-free, taking source language, target language, source text, target text, and optionally, translation instructions as input.

¹<https://huggingface.co/cyankiwi/gemma-4-31B-it-AWQ-4bit>

4 Task 1: Error Span Prediction

Our Task 1 approach separates semantic error analysis from predicting error offsets.

Both of our Task 1 submissions consist of three steps:

1. Creating free-form error lists (Section 4.1)
2. Self-review (Section 4.2)
3. Span alignment (Section 4.3)

4.1 Free-form error list

In the first stage, the model produces lists of critical, major, minor errors and omissions. At this stage it is not required to produce substrings or offsets. Table 1 shows an authentic model output from this stage.

4.2 Self-review

In the second stage, the model reviews and modifies its own outputs. An example from this stage is shown in Table 2.

4.3 Span Alignment

In the final stage, we ask the model to align the errors to exact token spans. We apply a simple rule-based tokenizer to the source and translation and present the resulting tokens to the model together with explicit token indices. The tokenizer treats whitespace as a delimiter, tokenizes characters in configured CJK ranges as individual tokens and treats the remaining non-whitespace sequences as tokens. The model is then asked to align previously identified errors with the corresponding token span. These token-level spans are then parsed and programmatically converted to character-level spans required by the shared task.

This stage is similar for both *v14* and *cat-v4* submissions, differing in that *cat-v4* also outputs corresponding MQM-style error categories at the same time. We use the same categories as Kocmi and Federmann (2023). They also differ in their demonstration sets. Model outputs are parsed and validated before token spans are converted to character offsets.

For *cat-v4* system, if the conversion fails due to invalid span range, we attempt one repair call and only accept modifications to the invalid spans. We used this repair only for *cat-v4*.

Localized omissions are represented as source-side spans, while unlocalized omissions receive a

segment-level severity. We convert all these representations to the organizer format only when processing shared task test set.

We also chunked a record if either its source or translation exceeded 1024 model-tokenizer tokens to handle long inputs.

Examples from this stage are shown in Tables 3 to 5

4.4 Inference Parameters

Our Task 1 uses temperature 1.0, top-*p* 0.95, top-*k* 64 and fixed seed 26 during inference. The *cat-v4* repair call uses temperature 0, top-*p* 1 and disables top-*k*.

5 Task 2: Segment Score Prediction

For Task 2, we submit two different approaches. *v14-regression* system converts Task 1 output into 31 features, including error counts and span coverage by severity, and omission features. A constrained nonnegative linear regression is fitted on WMT25 training data: $\hat{y} = 100 - (b + \beta^\top \mathbf{x})$, with $b, \beta \geq 0$, followed by clipping to $[0, 100]$. Failed predictions from Task 1 receive the score mean from the training data: 71.1493.

On the other hand, *da-6d* uses a different approach: directly asking the model for a 0–100 score. For this submission, we drew demonstrations from WMT25 training data. We evaluated two-, four- and six-demonstration candidates on 466-record subset of the development split and selected one of the six-demonstration variants using macro segment-level Spearman score. As a result, one of the six-demonstration candidates was selected. The total pool of candidates was 24 consisting of two demos, 14 consisting of four demos and 16 consisting of six demos.

6 Experimental Setup

6.1 Data Partitioning

We partitioned WMT25 data by their source documents, by removing the language-pair prefix when identifying the source documents. This prevents the same underlying document from appearing in different splits when reused across language pairs. Table 6 summarizes the resulting data. Internally, we also used two smaller subsets of dev set during development for faster evaluation.

Detection field	Authentic value
src_lang	Czech
tgt_lang	German
src	Ina T. hodnotí sexy pokusy hvězd: Myslivcová přestřelila!
tgt	Ina T. bewertet die sexy Versuche der Stars: Die Jägerin hat daneben gefeuert!
prompt_instruction	""
critical_errors	[]
major_errors	[["Die Jägerin—mistranslation of the proper name 'Myslivcová'."]]
minor_errors	[]
source_omissions	[]
segment_omission_severity	none

Table 1: Authentic input and output fields for error detection.

Self-review field	Authentic value
draft_detection	major_errors: [["Die Jägerin—mistranslation of the proper name 'Myslivcová'."]]
critical_errors	[]
major_errors	[["Die Jägerin' is a mistranslation of the proper name 'Myslivcová'."]]
minor_errors	[]
source_omissions	[]
segment_omission_severity	none

Table 2: Authentic self-review fields. The unchanged language, text, and instruction fields are omitted.

6.2 Evaluation

For Task 1, we report span precision, span recall, and the official metric: micro averaged MPP with optimal one-to-one span matching and severity weights 1.0 for major and 0.25 for minor errors (Perrella et al., 2026). Errors we internally tag as critical are mapped to major.

For Task 2, we evaluate segment-level predictions using Pearson and Spearman correlation together with mean absolute error (MAE) and root mean squared error (RMSE). We also report system-level Pearson and Spearman correlations in portability experiments.

7 Development Results

7.1 Task 1

Table 8 reports our development results. In Table 7 we compare our systems on the intersection of the development set with the official system results from the previous year (Lavie et al., 2025).

Our systems achieve the highest observed MPP on this common subset, with cat-v4 reaching 0.2490 and v14 0.2460. The external systems show substantially different precision-recall profiles. AIP secondary obtains the highest precision, 0.2863, but only 0.1152 recall, while XCOMET-XXL reaches the highest recall, 0.2821, with a precision of 0.0590. Our systems achieve a more balanced trade-off between the two quantities, resulting in higher

MPP.

7.2 Task 2

Table 9 shows our development results for Task 2. The six-demo version gives the strongest point estimate for every reported metric. Because its selection subset is part of the development data, we treat these results with caution and defer final conclusions to the WMT26 official results.

8 Cross-Model Portability

Table 10 shows Task 1 v14 when used with different model backbones. Gemma 4 26B reaches MPP 0.2271, close to the 31B result of 0.2463, while the other four transfers fall between 0.1165 and 0.1381. Gemma 3 27B (Team et al., 2025) scores 0.1328 despite its similar parameter count. However, this experiment is not intended as a general model comparison, as the complete system was tweaked for the instruction-tuned version of Gemma 4 31B during development and few-shot optimization and deliberately left unchanged.

For the Task 2 direct assessment submission, Table 11 compares the submitted six-demo version with its zero-demo control across different models.

9 WMT26 Submissions

For Task 1, we submit v14 and repaired cat-v4 on the eight priority language pairs, covering 3 164

v14 alignment field	Authentic value
detection_response	major_errors: [““Die Jägerin” is a mistranslation of the proper name ‘Myslivcová’.”]
target_tokens	(0) Ina (1) T. ... (8) Die (9) Jägerin (10) hat ...
target_spans	[(8,10,major)]
source_omission_spans	[]
segment_omission_severity	none

Table 3: Authentic v14 alignment fields. The token span becomes target-character span [45,56).

Categorized alignment field	Authentic value
detection_response	major_errors: [““Die Jägerin” is a mistranslation of the proper name ‘Myslivcová’.”]
target_tokens	(0) Ina (1) T. ... (8) Die (9) Jägerin (10) hat ...
target_spans	[(8,10,major, accuracy/mistranslation)]
source_omission_spans	[]
segment_omission_severity	none

Table 4: Authentic fields for the version with additional categories for the same record.

items and 95 323 hypotheses. For Task 2, we submit six-demo DA and v14 regression on priority language pairs plus several additional language pairs: English–Czech, Czech–Ukrainian, and the released English–Chinese variants. Each Task 2 submission covers 8 252 items, 149 574 hypothesis occurrences, and 141 006 unique effective inputs. Six-demo DA has no prediction fallback; v14 regression uses the frozen train-mean as fallback for 747 unique failed v14 predictions.

10 Conclusion

In this paper, we presented CUNI’s WMT26 submissions based on an open-weight Gemma 4 31B model. Our Task 1 pipeline separates error analysis from span alignment, while our Task 2 systems compare regression on spans with direct assessment with few-shot demonstrations. Both approaches show promising development results, but we defer broader conclusions to the official WMT26 evaluation because repair and few-shot demonstrations selection used development data.

Acknowledgments

This work has been supported by the grants CZ.02.01.01/00/23_020/0008518 (Jazyková, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím), TQ28000003 (NCUI, Národní centrum umělé inteligence), also by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

AI Use Declaration

We used OpenAI GPT models as code assistants during the development of the experiment code and to refine the phrasing of portions of the text.

References

- Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Drach, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Repair field	Authentic value
target_tokens	(0) sidebar.invites:Pakkumised
failed_alignment	[(1,2,critical,accuracy/mistranslation), (0,1,major,fluency/punctuation)]
conversion_error	Invalid categorized target token span [1,2,...]
target_spans	[(0,1,critical,accuracy/mistranslation), (0,1,major,fluency/punctuation)]
source_omission_spans	[]
segment_omission_severity	none

Table 5: Example of authentic failure-only repair. Only the invalid boundary changes.

Split	Records	Annotations	Documents
Train	4 533	9 066	30
Development	3 058	6 116	34
Test	46 828	93 656	181
Mini-test	4 868	9 736	40

Table 6: Dataset split statistics. Original data comes from WMT25 human annotations and source document-aware splitting was used to ensure that a given source document only appears in a single split. Mini-test is a subset of test.

System	Precision	Recall	MPP
cat-v4 (w/o repair)	0.2338	0.2662	0.2490
v14	0.2287	0.2661	0.2460
AIP primary	0.2860	0.1358	0.1842
AIP secondary	0.2863	0.1152	0.1643
GemSpanEval-QE	0.0888	0.2702	0.1336
GemSpanEval	0.0830	0.2715	0.1271
AutoLQA	0.1505	0.1007	0.1207
AutoLQAESA	0.1736	0.0851	0.1142
XCOMET-XXL	0.0590	0.2821	0.0975
AutoLQA41	0.2586	0.0592	0.0963
XCOMET-XL	0.0537	0.2445	0.0880

Table 7: MPP on the 2 887-record development intersection shared with official WMT25 Task 2 inputs. Since we recalculated these results on the intersection of the systems with selected development data, these results differ from the official results.

System	Precision	Recall	MPP
v14	0.2261	0.2641	0.2436
cat-v4 (w/o repair)	0.2287	0.2639	0.2451
cat-v4	0.2293	0.2657	0.2462

Table 8: Task 1 development results.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. [Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2026. [Awq: Activation-aware weight quantization for llm compression and acceleration](#). *Preprint*, arXiv:2306.00978.

Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. [Span-level machine translation meta-evaluation](#). *Preprint*, arXiv:2603.19921.

Danil Semin, Ondřej Dušek, and Zdeněk Kasner. 2026. [Strategies for span labeling with large language models](#). *Preprint*, arXiv:2601.16946.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.

Method	Pearson	Spearman	MAE	RMSE	Sys. Spearman	Sys. Kendall
V14 regression	0.7631	0.5886	15.20	20.41	0.6624	0.5128
Instruction-off DA	0.7425	0.6421	16.82	24.90	–	–
Instruction-aware DA	0.7723	0.6507	15.94	22.23	0.7605	0.5980
Six-demo DA	0.8204	0.6935	14.14	19.50	0.7708	0.6057

Table 9: Task 2 full-development results. System metrics are macro-averaged over 14 language pairs.

Backbone	MPP
Gemma 4 31B AWQ	0.2463
Gemma 4 26B A4B AWQ	0.2271
Gemma 4 E4B BF16	0.1381
Ministral 3 14B FP8	0.1339
Gemma 3 27B AWQ	0.1328
Gemma 4 E2B BF16	0.1165

Table 10: Task 1 v14 results on the 4 868-record score-blind test portability subset. The v14 system is used with different model backbones without model-specific adaptation.

Backbone	DA program	Pearson	Spearman	MAE	RMSE	Sys. r	Sys. ρ
Gemma 4 31B	Zero-demo	0.6710	0.5866	19.83	26.66	0.8124	0.7260
Gemma 4 31B	Six-demo	0.7181	0.6225	17.79	23.66	0.8627	0.7298
Gemma 4 26B	Zero-demo	0.7256	0.6137	18.44	24.79	0.8761	0.7409
Gemma 4 26B	Six-demo	0.7095	0.6214	17.22	23.24	0.8614	0.7245
Gemma 4 E4B	Zero-demo	0.5401	0.4825	26.60	35.53	0.6953	0.6840
Gemma 4 E4B	Six-demo	0.5836	0.5134	24.81	33.52	0.7125	0.7041
Gemma 4 E2B	Zero-demo	0.5475	0.4402	23.87	32.41	0.7121	0.6488
Gemma 4 E2B	Six-demo	0.5567	0.4216	23.98	32.35	0.6826	0.4723
Gemma 3 27B	Zero-demo	0.3951	0.3779	27.00	36.92	0.5736	0.6148
Gemma 3 27B	Six-demo	0.5199	0.4768	27.66	36.84	0.6703	0.6507
Ministral 3 14B	Zero-demo	0.4079	0.2275	23.53	31.95	0.5956	0.4270

Table 11: Direct-assessment portability on 4 868 WMT25 mini-test records.