

PATTRANSBENCH: AN IPC-STRATIFIED, MULTI-DIMENSIONAL BENCHMARK FOR PATENT MACHINE TRANSLATION EVALUATION

Xiaowei Su^{1,2} Xusen Cheng^{1,*}

¹Renmin University of China, Beijing, China

²Beijing Bayuegua Technology Co., Ltd., Beijing, China
{xwsu, xusen.cheng}@ruc.edu.cn

*Corresponding author

ORCID: 0000-0002-7784-7141 (Su), 0000-0002-2028-1337 (Cheng)

ABSTRACT

We present PatTransBench, a 12,000-item Chinese–English/Japanese/Korean patent machine translation (PatMT) benchmark with a four-dimensional evaluation protocol. The benchmark comprises 6,000 short-text (sentence/claim) and 6,000 long-text (paragraph) pairs, evenly split across three language pairs and both translation directions, built via IPC-section-stratified sampling and dual-threshold quality gating from upstream patent-family parallel corpora. The evaluation protocol spans literal n -gram overlap (BLEU, ROUGE-L, chrF), cross-lingual semantic textual similarity (CLSTS), 38-probe-language consistency testing, and serving performance under concurrency. A case study on PatMT2-7B and its fp8/AWQ4/AutoRound-4bit variants reveals direction-dependent quantization effects masked by aggregate scores and non-monotonic language-failure patterns across quantization schemes that are invisible to single-metric evaluation. We plan to release the benchmark, protocol, and case-study results upon publication to support reproducible PatMT evaluation.

Index Terms— patent machine translation, benchmark, evaluation, IPC stratification, quantization

1 INTRODUCTION

Patent claims and specifications are written in dense, legally load-bearing technical prose where a mistranslated hyphen or a dropped negation can change the scope of a legal right, and the same document must often be produced consistently across Chinese, English, Japanese, and Korean for multi-jurisdictional filing. We refer to this task throughout as Patent Machine Translation (PatMT). Despite its practical weight, most reported evaluations of patent-oriented PatMT systems still rely on a single corpus-level BLEU or ROUGE score computed on whatever internal test split happened to be available. A single scalar cannot distinguish a model that paraphrases correctly from one that has quietly started answering in the wrong language, nor can it tell a practitioner whether a quantized deployment variant will hold up under concurrent production load.

PatTransBench addresses these gaps through three design decisions: (1) *IPC-stratified sampling* with honest reporting when stratification fails due to upstream data scarcity; (2) *precise metric naming*, using Cross-lingual Semantic Textual Similarity (CLSTS) computed from a named multilingual sentence-embedding model; (3) *multi-dimensional evaluation* that surfaces findings no single metric reveals. The PatTransBench dataset is currently

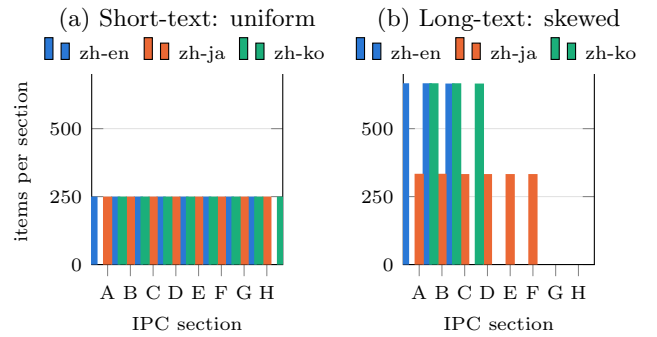


Fig. 1. Per-IPC-section item counts. Short-text (left): exact 250-item uniform coverage across all eight sections. Long-text (right): skewed to A–C (zh-en/zh-ko) and A–F (zh-ja) after dual-threshold quality gating.

being prepared for public release and will be made available at <https://huggingface.co/datasets/SuStarSmile/PatTransBench>. Our contributions are:

- A 12,000-item IPC-stratified PatMT benchmark with dual-threshold quality gating, covering three language pairs with honest reporting of stratification outcomes.
- A four-dimensional evaluation protocol spanning literal overlap, CLSTS, language consistency, and serving performance.
- A case study revealing direction-dependent quantization effects and non-monotonic language-failure patterns across quantization schemes that no single dimension would surface.

Existing MT benchmarks such as WMT shared tasks and FLORES [1] provide broad-coverage evaluation sets drawn from news and Wikipedia, but are not stratified by technical domain and do not evaluate serving performance or output-language consistency under fine-tuning. Standard MT metrics including BLEU [2], ROUGE [3], and chrF [4] capture surface-level overlap but miss semantic fidelity and deployment viability. Learned metrics such as COMET [5] require human-annotated training data; our CLSTS instead scores semantic agreement directly with a fixed multilingual sentence encoder, at no annotation cost and with full reproducibility.

Evaluating a benchmark of this scope also requires a defensible answer to a practical question: on what evidence should a deployment decision rest? We build the four dimensions so that each targets

a failure mode the others cannot detect on their own, and we report the disagreements among them rather than averaging them into a single number. The case study is a demonstration of that principle on one model family and its quantization variants, not a cross-system comparison; construction, coverage statistics, and protocol stand independently of which systems are ultimately scored against them.

2 BENCHMARK CONSTRUCTION

2.1 Composition and IPC Stratification

PatTransBench contains 12,000 items: 6,000 short-text (sentence- and claim-level pairs) and 6,000 long-text (paragraph-level pairs drawn from full patent specifications), evenly split across three language pairs—Chinese-English (zh-en), Chinese-Japanese (zh-ja), Chinese-Korean (zh-ko)—at 2,000 items per pair. Each item carries both translation directions, so counts refer to aligned pairs rather than single-direction instructions.

Both pipelines use IPC-stratified sampling targeting eight sections (A through H). Given a per-language-pair budget $N = 2000$ and section set S with $|S| = 8$, the per-section target is:

$$n_s = \left\lfloor \frac{N}{|S|} \right\rfloor + \mathbb{I}[\text{rank}(s) < N \bmod |S|] = 250 \quad (1)$$

When a section returns fewer rows than requested, the deficit is re-distributed across sections with non-zero supply using the same formula, fetched at an offset past already-retrieved rows. If a second-pass deficit remains, it is logged as a warning. Short-text achieves exact uniform coverage: all 24 first-pass queries returned exactly 250 rows, confirmed in extraction logs. Long-text is skewed: although upstream patent-family data covers all eight IPC sections for every language pair, the dual-threshold quality gate (Section 2.2) filters out a substantial fraction of candidates, and the remaining qualified items are unevenly distributed across sections. After quality gating and top-up rebalancing, zh-en and zh-ko cover only sections A–C (667/667/666), zh-ja covers A–F (334/334/333/333/333/333), with zero items in remaining sections. This reflects the interaction between patent-family density and quality-gate stringency across IPC domains, not a sampling defect. Figure 1 visualizes both outcomes.

2.2 Dual-Threshold Quality Gating

Short-text candidates require `similarity_score` > 0.9 . Long-text candidates must satisfy a conjunction of two independent embedding-model thresholds plus a language-pair-specific word-count-ratio range:

$$\text{sim}_{0.6B} > 0.9 \wedge \text{sim}_{8B} > 0.8 \wedge r_{\min} < \frac{|\text{words}(\text{tgt})|}{|\text{words}(\text{src})|} < r_{\max} \quad (2)$$

where $\text{sim}_{0.6B}$ and sim_{8B} are precomputed scores from a LoRA-adapted 0.6B and an 8B embedding model respectively; requiring agreement from two independently-sized models is stricter than either alone. Word-ratio ranges are (0.5, 0.9) for zh-en, (1.0, 1.5) for zh-ja, and (1.0, 1.6) for zh-ko, reflecting different average expansion/contraction ratios. Both gates are applied jointly with IPC filtering in a single SQL query.

3 EVALUATION PROTOCOL

Dimension 1: Literal Overlap. We report corpus-level BLEU via `sacrebleu` [6], ROUGE-L [3], and chrF in `chrF2` configuration (char-

acter order 6, word order 0, $\beta = 2$) [4, 7]. For Japanese and Korean targets we use MeCab-based tokenizers rather than whitespace segmentation, since whitespace tokenization is not meaningful for these languages.

Dimension 2: Cross-lingual Semantic Textual Similarity. We compute CL-STs as cosine similarity of L2-normalized [CLS] embeddings from paraphrase-multilingual-mpnet-base-v2 [8]:

$$\text{CL-STs}(h, r) = \frac{\mathbf{e}_h \cdot \mathbf{e}_r}{\|\mathbf{e}_h\|_2 \|\mathbf{e}_r\|_2}, \quad \mathbf{e} = \text{CLS}(\text{Enc}(\cdot)) \quad (3)$$

The encoder is the multilingual sentence-embedding model paraphrase-multilingual-mpnet-base-v2, an XLM-RoBERTa network distilled across 50+ languages so that equivalent sentences in different languages land close together in one shared embedding space. We score both hypothesis and reference with this frozen encoder, so the resulting number measures geometric agreement between two vectors rather than a quality estimate produced by a trained regression head; it needs no annotated data and is exactly reproducible from the released script and model checkpoint.

Dimension 3: Language Consistency. We probe 38 target languages covering major world languages, low-resource and non-Latin-script languages, and script variants (e.g., Traditional Chinese, Cantonese). For each, we issue a fixed-source translation instruction and inspect raw outputs for two failure modes: repetition-collapse (a short n -gram repeated far beyond fluent output) and instruction-echo (output translates the instruction rather than the source). Automated non-empty-response checks are insufficient; only direct text inspection reveals these failures.

Dimension 4: Serving Performance. We measure TTFT (`avg/p95/p99`), TPOT, QPS, and TPS under concurrent load at levels 10/20/50, separately for short and long text, since long documents produce qualitatively different latency and throughput profiles.

4 CASE STUDY: PATMT2

We apply the full protocol to PatMT2-7B, a bilingual/multilingual PatMT model fine-tuned in two stages on zh-en/ja/kr patent data, and three quantized variants: fp8, AWQ 4-bit (awq4), AutoRound 4-bit (auto4). This is a demonstration on one model family rather than a cross-system comparison: we hold no independently re-verified numbers for other systems under one frozen test set and matched decoding, so we present the within-family disagreements the four dimensions expose and leave cross-system ranking to the released benchmark.

4.1 Direction-Dependent Quantization Effects

Table 1 reports BLEU/ROUGE-L/chrF across all six directions. Quantization does *not* degrade quality uniformly: on ja→zh, fp8 drops 4.4 BLEU points below base (48.38→43.95) while awq4 stays within 0.1 points; auto4 lands between them. The same direction shows the largest ROUGE-L/chrF spread across variants. This direction-specific effect is masked by cross-direction averaging—a practitioner choosing between fp8 and awq4 from aggregate scores alone would not see it. CL-STs tracks literal-overlap metrics directionally, confirming broad consistency at this quantization stage.

4.2 Persistent Four-Language Failure Mode

34 of 38 probed languages produce fluent, correctly-scripted output under the BF16 and fp8 variants. Four languages fail un-

Table 1. BLEU / ROUGE-L / chrF by variant and direction (4,000 items/direction).

	Model	Dir.	BLEU	ROUGE-L	chrF
base		zh→en	40.31	72.70	68.04
		en→zh	42.96	54.31	42.21
		zh→ja	48.65	13.02	53.34
		ja→zh	48.38	39.63	49.03
		zh→ko	36.44	42.48	44.56
		ko→zh	43.59	40.54	41.49
fp8		zh→en	40.18	72.71	67.93
		en→zh	42.70	54.38	42.13
		zh→ja	48.69	13.20	53.17
		ja→zh	43.95	39.53	48.19
		zh→ko	36.11	42.35	44.25
		ko→zh	43.43	40.75	41.02
awq4		zh→en	40.10	72.57	68.01
		en→zh	43.64	54.03	41.77
		zh→ja	48.65	12.67	53.09
		ja→zh	48.45	40.27	48.40
		zh→ko	36.73	42.35	44.62
		ko→zh	43.86	40.65	40.97
auto4		zh→en	39.71	72.46	67.85
		en→zh	43.10	54.09	42.02
		zh→ja	47.76	12.95	52.42
		ja→zh	45.88	39.78	48.23
		zh→ko	36.41	42.31	44.41
		ko→zh	43.29	40.65	40.80

Table 2. Language-consistency failures (BF16/fp8, representative). Failure sets vary across quantization variants; see text for details.

Lang.	Chars	Time(s)	Failure
Tibetan	18,140	177.4	Phrase repeated 464×
Mongolian	9,798	177.4	Token loop
Tamil	14,020	137.6	“Moreover” repeated 1,949×
Uyghur	45	0.3	Instruction echo

der BF16/fp8 (Table 2): Tibetan, Mongolian, and Tamil exhibit repetition-collapse; Uyghur echoes the instruction. The failure set varies across quantization schemes: AutoRound-4bit recovers three of the four (Tamil, Mongolian, Uyghur) but introduces a new failure on Gujarati (37/38); AWQ-4bit recovers Tamil and Mongolian but fails on Tibetan and Uyghur (36/38). That the core weakness concentrates in low-resource, non-CJK, non-Latin-adjacent scripts across all variants suggests a systematic representational fragility rather than a sampling artifact. Automated non-empty-response checks report 38/38 success for every variant; only direct text inspection reveals these failures.

4.3 Serving Performance Trade-offs

Table 3 reports serving metrics for zh→en. fp8 outperforms awq4 by 25–30% on QPS at every concurrency level for short-text. Long-text QPS is ~2 orders of magnitude lower than short-text, but TPS is much closer between data types because TPS aggregates over many more output tokens per request. Figure 2 shows near-linear scaling with concurrency. Quality tables alone would not predict this throughput ordering; deployment decisions require both quality and

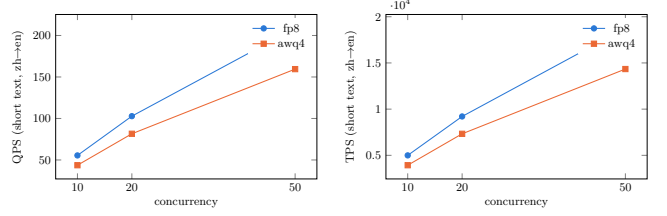


Fig. 2. QPS and TPS vs. concurrency, zh→en short-text, fp8 vs. awq4. Both scale near-linearly; fp8 leads at every concurrency level.

Table 3. Serving performance, zh→en, fp8 vs. awq4 (2,000 items/cell).

	Var.	Type	Conc.	QPS	TPS	TTFT avg
fp8		short	10	55.47	4995	0.018s
		short	20	102.79	9207	0.019s
		short	50	208.79	18760	0.023s
		long	10	1.97	3966	0.073s
		long	20	3.02	6115	0.079s
		long	50	4.51	9077	0.101s
awq4		short	10	43.66	3929	0.024s
		short	20	81.57	7331	0.023s
		short	50	159.47	14339	0.030s
		long	10	1.60	3253	0.117s
		long	20	2.55	5128	0.125s
		long	50	4.03	8138	0.155s

performance dimensions.

4.4 What the Four Dimensions Reveal Together

Taken individually, each dimension supports a plausible but incomplete recommendation. Quality scores alone suggest awq4 is interchangeable with fp8. Serving numbers alone suggest fp8 wins with no quality context. Language-consistency testing alone suggests all variants share the same low-resource language weakness. Only together do we get the actionable picture: fp8 is faster and not meaningfully worse on five of six directions, but meaningfully worse on ja→zh; quantization can both recover and introduce language failures (AutoRound recovers 3/4 BF16 failures but adds Gujarati); and no single variant dominates all dimensions. This is the central argument for multi-dimensional evaluation: the failure modes that matter for deployment do not all appear in the same metric.

5 CONCLUSION

PatTransBench provides a 12,000-item IPC-stratified patent MT benchmark with a four-dimensional evaluation protocol. Our case study demonstrates that multi-dimensional evaluation surfaces actionable findings—direction-dependent quantization effects and non-monotonic language-failure patterns across quantization schemes—that no single metric reveals. Limitations include: (1) skewed long-text IPC coverage (A–C/A–F only, due to quality-gate filtering reducing qualified candidates in other sections despite upstream data availability); (2) absence of human evaluation (planned with ~260 items, 3 annotators/language, adequacy+fluency Likert scales); (3) CL-STS correlation with human

judgments not yet validated; (4) no systematic quantization ablation beyond four variants; (5) case study limited to one model family. Future work includes expanding long-text IPC coverage, validating CL-STS against human judgments and trained COMET, extending the protocol to distilled/pruned models, and incorporating diverse translation systems for cross-system comparison. We plan to release all code, data, and results upon publication.

6 REFERENCES

- [1] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang, “Scaling neural machine translation to 200 languages,” *Nature*, vol. 630, no. 8018, pp. 841–846, 2024.
- [2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2002, pp. 311–318, Association for Computational Linguistics.
- [3] Chin-Yew Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 2004, pp. 74–81, Association for Computational Linguistics.
- [4] Maja Popović, “chrF: character n-gram f-score for automatic MT evaluation,” in *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*. 2015, pp. 392–395, Association for Computational Linguistics.
- [5] Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie, “COMET: A neural framework for MT evaluation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 2685–2702, Association for Computational Linguistics.
- [6] Matt Post, “A call for clarity in reporting BLEU scores,” in *Proceedings of the Third Conference on Machine Translation (WMT): Research Papers*. 2018, pp. 186–191, Association for Computational Linguistics.
- [7] Maja Popović, “chrF++: words helping character n-grams,” in *Proceedings of the Second Conference on Machine Translation (WMT)*. 2017, pp. 612–618, Association for Computational Linguistics.
- [8] Nils Reimers and Iryna Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 4512–4525, Association for Computational Linguistics.

Thai (th), Italian (it), German (de), Vietnamese (vi), Malay (ms), Indonesian (id), Filipino/Tagalog (tl), Hindi (hi), Traditional Chinese (zh-TW), Polish (pl), Czech (cs), Dutch (nl), Khmer (km), Burmese (my), Persian (fa), Gujarati (gu), Urdu (ur), Telugu (te), Marathi (mr), Hebrew (he), Bengali (bn), Tamil (ta), Ukrainian (uk), Tibetan (bo), Kazakh (kk), Mongolian (mn), Uyghur (ug), Cantonese (yue). 34 of these produce fluent, correctly scripted output on direct inspection under the base BF16 variant; Tibetan, Mongolian, Tamil, and Uyghur fail as detailed in Table 2.

A THE 38 PROBE LANGUAGES

Chinese (zh), English (en), French (fr), Portuguese (pt), Spanish (es), Japanese (ja), Turkish (tr), Russian (ru), Arabic (ar), Korean (ko),