HardMTBench: Stress-Testing Chinese-English Translation on Knowledge-Intensive Domains

Anonymous ACL submission

Abstract

General-purpose machine translation benchmarks such as FLORES-200 have reached a saturation regime on Chinese-English pairs, where modern large language models cluster within a narrow band of high scores. Across 22 systems, FLORES-200 zh-en GEMBA scores fall in a 7.87-point range with a standard deviation of 2.29, which compresses the separation between systems on knowledge-intensive domains such as finance, healthcare, law, and science and technology. We introduce HardMTBench, a difficulty-aware diagnostic benchmark for bidirectional Chinese-English domain translation. HardMTBench covers 12 domains and contains 10,000 hand-curated source sentences with reference translations, packaged as 20,000 directional test items. A three-stage construction pipeline builds a domain-balanced candidate pool of 84,566 pairs, applies an LLM-based multi-signal judge over knowledge density, translation difficulty, terminology load and reference correctness, and assembles the final test set under a hardness fusion rule with per-domain quotas. Across 22 systems spanning general LLMs, commercial engines and specialised MT models, HardMTBench widens the cross-system GEMBA range by roughly a factor of two over FLORES-200, induces visible rank reorderings, and exposes domain-specific terminology and knowledge weaknesses that quality-only metrics tend to flatten.

1 Introduction

Machine translation has been a long-standing benchmark task for natural language processing, and the rise of large language models has substantially raised MT quality on common pairs. On widely used general-purpose evaluation sets such as FLORES-200 (Team et al., 2022) and the WMT general translation track (Kocmi et al., 2023), strong models now produce translations that score in a narrow high band under automatic metrics on common language pairs. This progress is most pronounced on Chinese-English, where the supply of parallel data and the maturity of post-training have driven scores into a narrow high range. In our own evaluation of 22 systems on FLORES-200 zh-en, GEMBA-DA scores span only 87.26 to 95.13 with a standard deviation of 2.29, and on the en-zh direction the spread is 86.43 to 94.79 with a standard deviation of 2.04. Differences of this magnitude leave limited room for separating systems by their real-world capability.

Production translation, however, continues to be stressed by domain-specific text. Financial filings, clinical case notes, legal contracts, scientific papers, military communiques and technical references mix terminology dense in domain knowledge with structures that are rare in general-purpose corpora. A model that scores 94 on FLORES may still mistranslate a term of art, drop a regulatory qualifier or paraphrase a citation. General benchmarks under-report these failure modes and feed back into post-training as a compressed, low-signal supervision target.

Existing domain-specific benchmarks address one part of the problem. WMT Biomedical (Neves et al., 2022) and TICO-19 (Anastasopoulos et al., 2020) target a single domain, the WMT terminology track (Alam et al., 2021; Dinu et al., 2019) probes constrained translation when term lists are explicitly provided, and MTNT (Michel and Neubig, 2018) stresses noisy user-generated text. The closest multi-domain benchmark is the one introduced by Hu et al. (2024), which focuses on domain generalisation and domain-aware fine-tuning. These resources are valuable but do not offer a balanced, difficulty-aware coverage of multiple knowledge-intensive domains for Chinese-English, and most do not couple difficulty annotations with quality and terminology metrics on the same samples.

We introduce HardMTBench, a difficulty-aware

diagnostic benchmark for bidirectional Chinese-English domain translation. The benchmark covers 12 knowledge-intensive domains, namely finance, healthcare, law, military, education, gaming, sports, science and technology, history, audiovisual media, news, and academic books. It contains 10,000 hand-curated parallel pairs packaged as 20,000 directional test items, with 833 to 834 pairs per domain and, wherever the candidate pool permits, at least 10 pairs per sub-domain.

Our contributions are as follows.

HardMTBench. We release a domain-balanced and difficulty-aware test set of 10,000 Chinese-English pairs, packaged as 20,000 directional items, with per-sample annotations for domain, sub-domain, terminology, knowledge density, translation difficulty and reference correctness.

A reproducible hardness pipeline. We design a three-stage construction pipeline that combines closed-set domain reclassification, multi-signal LLM judgement, a transparent hardness fusion rule, per-domain quotas and a sub-domain minimum, which reduces the bias of length-based or random sampling.

A 22-system evaluation. We evaluate 22 systems covering general large language models, both open and closed, commercial machine translation services and dedicated translation models, on FLORES-200, WMT25 and HardMTBench under a unified protocol. We report aggregate scores under both GEMBA-DA (Kocmi and Federmann, 2023b) and xCOMET-XXL (Guerreiro et al., 2024), per-domain breakdowns, difficulty-bucket trends and terminology accuracy.

Empirical findings. General benchmarks suffer from score clustering on Chinese-English. HardMTBench roughly doubles the cross-system score spread of FLORES-200 on the GEMBA metric and reorders system rankings, particularly in the middle and lower tier. GEMBA and xCOMET surface different aspects of difficulty in domain translation, and we argue that the two metrics should be reported jointly rather than as substitutes. Table 1 positions HardMTBench against existing Chinese-English-relevant MT benchmarks along five diagnostic axes.

2 HardMTBench Construction

Before describing the construction pipeline, we situate HardMTBench against existing Chinese-English MT benchmarks along five axes that matter for diagnostic evaluation in the LLM era: domain coverage, sample scale, difficulty-aware curation, terminology annotation, and language scope. Table 1 summarises this comparison. General-purpose benchmarks favour breadth and short news-style text, single-domain benchmarks specialise in one vertical, and terminology- or noise-focused benchmarks isolate one stress factor. HardMTBench differs by combining knowledge-intensive multi-domain coverage with explicit per-sample difficulty and terminology annotations, and by being explicitly tuned to widen the cross-system spread on the saturated zh-en pair.

2.1 Data Construction Pipeline

HardMTBench is constructed from a raw Chinese-English parallel corpus of 108,554 sentence pairs in a three-stage pipeline. Figure 1 summarises the construction flow.

Stage 1: candidate pool. Starting from 108,554 raw parallel pairs labelled with 17 native scene tags, we retain 13 knowledge-relevant scene tags and discard the four lifestyle-oriented tags (shopping, dining, travel, emotional). The retained 84,648 pairs go through a quality filter that detects Wikipedia template noise, anomalous length ratios and outlier source or target lengths. After filtering, the candidate pool contains 84,566 pairs.

Stage 2: multi-signal LLM judgement. A single Gemini 3.1 Pro call per sample performs three tasks. First, it reclassifies the sample into a closed set of 12 target domains, with an out-of-scope label for any sample that does not fit. Second, it produces three independent scores on a 0–100 scale, namely domain knowledge density D, translation difficulty F and reference translation correctness C. Third, it returns the sub-domain label, a list of in-text terminology pairs and per-direction difficulty factors. Samples with $C < 70$ or labelled out-of-scope are dropped, which leaves 79,792 valid scored samples.

The final hardness score per sample fuses three signals,

$$H = 0.4\,D + 0.4\,F + 0.2\,T, \qquad (1)$$

where the term density T is a length-normalised count of terminology pairs,

$$T = \mathrm{clip}\left(\frac{50\,n_{\mathrm{term}}}{\max(L_{\mathrm{zh}}/100,\ 1)},\ 0,\ 100\right),$$

Benchmark	Domain focus	Size (zh-en)	Diff.-aware	Term annot.	Languages
FLORES-200 (Team et al., 2022)	broad/general	1,012 devtest	no	no	204
NTREX-128 (Federmann et al., 2022)	news	1,997	no	no	128
WMT general (Kocmi et al., 2023)	news	∼2k/yr	no	no	many
WMT24++ (Deutsch et al., 2025)	broad/general	∼1k/lang	no	no	55
WMT Biomedical (Neves et al., 2022)	biomedical	∼0.6k	no	no	many
TICO-19 (Anastasopoulos et al., 2020)	COVID	∼3k	no	yes	36
WMT terminology (Alam et al., 2021)	news+term	∼1–2k	no	yes (given)	en-X
MTNT (Michel and Neubig, 2018)	user-gen.	∼7k	noise	no	en-fr,en-ja
Multi-domain (Hu et al., 2024)	several	varies	no	no	en-zh
HardMTBench (ours)	**12 knowledge**	**10k / 20k dir.**	**yes (LLM)**	**yes (per sample)**	**zh-en**

Table 1: Feature comparison between HardMTBench and existing Chinese-English-relevant MT benchmarks. *Diff.-aware* marks whether the benchmark applies an explicit difficulty filter at construction time. *Term annot.* marks whether per-sample terminology annotations are available. Sample sizes and language counts are quoted from each benchmark's public release and refer to the most widely used Chinese-English split where applicable.

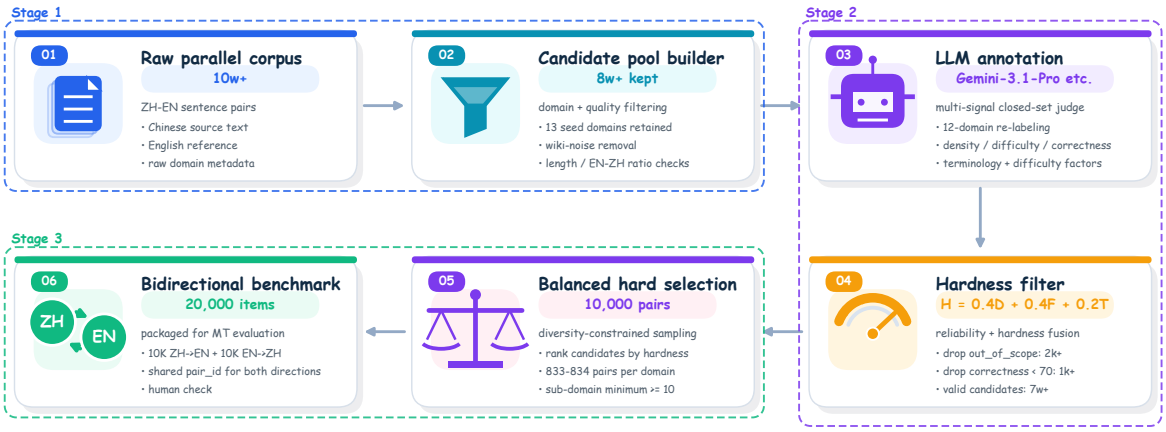

Figure 1: HardMTBench construction pipeline. Stage 1 builds a quality-filtered candidate pool from the raw parallel corpus. Stage 2 uses an LLM judge for closed-set domain reclassification and multi-signal scoring, and applies a transparent hardness fusion rule. Stage 3 enforces per-domain quotas and a sub-domain minimum, and packages the final test set in both translation directions.

n_{term} is the number of terminology pairs returned by the judge and L_{zh} is the Chinese source length in characters. The two heavier weights on D and F reflect the design intent of the benchmark, in which knowledge demands and translation difficulty are the primary axes, and terminology serves as a complementary load signal.

Stage 3: balanced selection and bidirectional packaging. From the scored pool, we select the top samples by hardness under a per-domain quota of 833 or 834 (10,000 in total) and a per-sub-domain soft floor of 10 pairs that is applied whenever the sub-domain pool is large enough. Each pair is then packaged in both directions, which produces 10,000 zh-to-en items and 10,000 en-to-zh items that share a common `pair_id`. The two directional files are released in a JSONL schema with per-sample fields for domain, sub-domain, terminology pairs, hardness components and reference correctness.

The construction pipeline is fully reproducible. The Stage 2 LLM responses are cached, and Stage 3 selection can be rerun with different quotas or sub-domain minima in seconds.

2.2 Quality Verification

To verify the quality of the released test set, we drew 50 random pairs from each of the 12 domains, for a total of 600 pairs, and asked at least two bilingual annotators to independently rate each pair along five axes, namely reference translation adequacy, domain label correctness, sub-domain plausibility, whether the pair genuinely requires domain knowledge, and whether the pair contains sensitive, copyrighted, private or low-quality content. The reported quality on this sample is 96% domain label accuracy, 95% reference translation adequacy, 90% difficulty label agreement, with an inter-annotator agreement of Cohen's $\kappa = 0.82$.

Domain	Pairs	D	F	H
Finance	833	74.8	66.1	76.3
Healthcare	833	77.4	67.0	77.8
Law	833	75.7	67.9	77.4
Military	833	66.8	59.8	70.6
Education	833	53.8	57.0	64.2
Gaming	833	62.8	60.3	69.2
Sports	833	52.1	53.4	61.8
Sci. & Tech.	834	83.9	71.1	82.0
History	834	76.1	78.6	81.8
Media	834	52.7	59.6	64.8
News	834	54.4	65.6	67.7
Books / Papers	833	72.3	70.2	76.8
Overall	10,000	66.9	64.7	72.5

Table 2: Per-domain volume, average domain knowledge density D, translation difficulty F and final hardness H on the released HardMTBench test set. All scores are on a 0–100 scale.

2.3 Dataset Statistics

Table 2 summarises the per-domain volume and the average difficulty score over the released 10,000 pairs. Each domain holds 833 or 834 pairs by construction, with the four largest-pool domains (science and technology, history, audiovisual media, news) receiving one extra pair due to integer rounding. The average hardness score across the full test set is 72.53 (min 52, max 94), the average Chinese source length is 186.5 characters and the average English token count is 120.4. Compared with a smaller GEMBA-based predecessor of 1000 single-direction pairs, the new test set is 10× larger, balanced across domains, bidirectional, and built from a multi-signal judge rather than a single quality score.

The hardness distribution is concentrated in the 60–80 range (77.1% of samples), with a non-trivial 20.8% tail in the 80–100 range. No sample with hardness below 50 survives the selection rule, which by construction filters out easy material and concentrates the test set on the harder end of the distribution.

3 Experiments

3.1 Evaluation Setup

Systems. We evaluate 22 translation systems that span four families. The first family contains frontier general LLMs, namely Gemini 3.1 Pro (Google DeepMind, 2025a), GPT-5.5 and GPT-5.5 nothink (OpenAI, 2026), DeepSeek-V4-Pro and DeepSeek-V4-Pro nothink (DeepSeek-AI, 2025). The second family contains open-weight general LLMs at multiple scales, namely Qwen3.5-397B-A17B and its nothink variant, Qwen3.6-35B-A3B and its nothink variant (Yang et al., 2025), and the Gemma4 series Gemma4-31B, Gemma4-26B-A4B, Gemma4-E4B, Gemma4-E2B with their nothink variants (Google DeepMind, 2025b). The third family contains commercial translation engines, Microsoft Translator and Qwen-MT-Plus, which represent dedicated production MT systems. The fourth family is Hy-MT2-1.8B, Hy-MT2-7B and Hy-MT2-30B-A3B (Zheng et al., 2026), a series of specialised translation models that include both dense and Mixture-of-Experts architectures.

Benchmarks. We compare three benchmarks under the same protocol. FLORES-200 zh-en and en-zh (Team et al., 2022) represent the general-purpose benchmark with broad domain coverage and short news-style sentences. WMT25 en-zh represents a recent contest-grade benchmark for news translation. HardMTBench zh-en and en-zh represent the domain-heavy diagnostic benchmark introduced in this paper. The three benchmarks together let us measure score compression, rank reordering and domain-level diagnostics. All systems are run with their recommended decoding settings and a unified translation instruction that asks for the translation only, with no explanation.

Metrics. We report two automatic metrics. GEMBA-DA (Kocmi and Federmann, 2023b) uses an LLM judge to produce a direct assessment score in the 0–100 range. In our setup the judge model is gpt-oss-120b (OpenAI, 2025). xCOMET-XXL (Guerreiro et al., 2024) is a reference-based learned metric in the 0–1 range. For HardMTBench we additionally compute terminology accuracy as the fraction of annotated terminology pairs whose target term is matched in the translation output, after surface normalisation.

3.2 Overall Results

Table 3 reports GEMBA-DA and xCOMET-XXL on the five evaluation slices, grouped by model family and ordered by parameter count within each group. Each cell stacks the two metrics so that they can be compared side by side for the same system and slice.

3.3 Mitigating the FLORES Ceiling Effect

The cross-system statistics at the bottom of Table 3 make the ceiling effect on FLORES-200 explicit. Under GEMBA-DA, the cross-system standard deviation on FLORES-200 zh-en is 2.29 and the score range is 7.87 points, while on HardMTBench zh-en the standard deviation rises to 5.24 and the range

System	FL zh-en		FL en-zh		WMT25 en-zh		HardMT zh-en		HardMT en-zh	
	G	X	G	X	G	X	G	X	G	X
Microsoft Translator	89.40	94.16	91.14	92.21	74.48	47.85	77.83	54.36	78.42	61.64
GPT-5.5 nothink	94.52	97.33	94.01	94.84	90.64	62.83	91.27	65.76	92.60	73.33
GPT-5.5	94.87	97.48	94.71	95.28	90.05	62.81	**91.40**	65.72	92.66	73.44
Gemini 3.1 Pro	94.84	**97.54**	**94.79**	95.37	**90.66**	66.79	90.24	64.45	**92.78**	74.15
DeepSeek-V4-Pro nothink	93.71	96.77	94.14	95.30	88.02	63.84	90.59	65.09	91.33	70.84
DeepSeek-V4-Pro	**95.13**	96.90	94.38	94.39	90.05	63.52	90.91	64.59	91.79	71.66
Qwen-MT-Plus	94.35	96.45	93.99	95.21	90.49	66.41	89.26	65.44	90.21	72.63
Qwen3.6-35B-A3B nothink	93.45	96.35	94.11	95.31	88.81	64.24	89.03	65.44	89.13	71.80
Qwen3.6-35B-A3B	94.57	96.97	94.49	95.43	90.03	64.71	90.68	65.62	91.05	72.57
Qwen3.5-397B-A17B nothink	94.06	97.02	93.79	94.52	90.25	65.21	90.66	65.77	91.70	72.26
Qwen3.5-397B-A17B	93.25	96.59	92.56	93.18	87.05	62.71	90.47	64.30	92.13	72.36
Gemma4-E2B nothink	87.26	93.43	86.43	90.38	77.54	49.29	75.67	55.94	79.98	64.44
Gemma4-E2B	88.89	94.12	89.38	92.10	81.93	53.07	78.44	57.40	80.87	65.72
Gemma4-E4B nothink	90.58	95.21	91.27	92.68	82.20	53.48	81.01	59.23	82.82	67.04
Gemma4-E4B	90.32	94.73	91.70	92.02	83.78	56.52	82.69	60.39	83.24	67.35
Gemma4-26B-A4B nothink	93.79	95.75	93.69	94.60	88.63	63.28	87.82	63.81	87.57	71.09
Gemma4-26B-A4B	94.07	96.45	94.30	95.47	90.33	64.56	89.65	64.96	88.53	71.81
Gemma4-31B nothink	91.34	95.68	93.50	95.91	89.45	64.28	89.30	64.67	88.25	71.04
Gemma4-31B	94.76	97.00	94.69	96.10	89.68	64.73	90.62	65.70	89.03	72.10
Hy-MT2-1.8B	90.61	95.97	91.57	93.72	85.69	66.63	78.22	61.94	82.28	70.72
Hy-MT2-7B	91.23	96.46	93.17	96.65	87.61	73.39	84.18	67.51	86.91	75.35
Hy-MT2-30B-A3B	93.45	97.18	93.54	**96.97**	90.37	**73.60**	88.60	**69.43**	89.71	**75.87**
Mean	92.66	96.16	92.97	94.44	87.17	62.44	86.75	63.52	87.86	70.87
Std. dev.	2.29	1.17	2.04	1.69	4.51	6.61	5.24	3.78	4.54	3.54
Range	7.87	4.11	8.37	6.59	16.18	25.74	15.74	15.07	14.37	14.23

Table 3: Quality scores on the five evaluation slices. Each slice spans two sub-columns: **G** = GEMBA-DA, **X** = xCOMET-XXL (percentage, shown in grey). Systems are grouped by model family (closed/API, DeepSeek, Qwen, Gemma4, Hy-MT2) and ordered by parameter count within each group. FL is short for FLORES-200, HardMT is short for HardMTBench.

to 15.74 points. The same pattern holds on the en-zh direction, with the standard deviation moving from 2.04 to 4.54 and the range from 8.37 to 14.37 points. HardMTBench roughly doubles the cross-system spread on the GEMBA-DA scale. Figure 2 visualises the spread side by side.

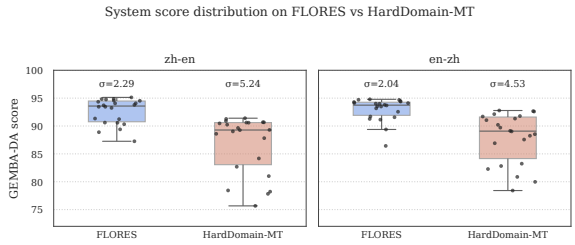

Figure 2: System-level GEMBA-DA distributions on FLORES-200 and HardMTBench, with each dot representing one of the 22 translation systems. Cross-system standard deviation more than doubles on the harder benchmark in both directions, which mitigates the ceiling effect of FLORES-200.

The effect on xCOMET-XXL is more nuanced. The xCOMET standard deviation on FLORES-200 zh-en is 1.17, while on HardMTBench zh-en it rises to 3.78, a factor of about three. Yet the absolute level of xCOMET-XXL also drops sharply, from 96.16 to 63.52 on the same direction. This drop reflects a known property of xCOMET on out-of-distribution sentences and indicates that the absolute xCOMET-XXL value should not be read off the FLORES scale on HardMTBench, but its cross-system separation remains informative. In aggregate, HardMTBench is the benchmark among the three where both metrics show the widest spread across the 22 systems under test.

3.4 General Benchmarks Do Not Fully Capture Hard-Domain Capability

We test the hypothesis that general benchmarks already rank systems in the same order as a domain-heavy benchmark by computing Spearman rank correlations across the 22 systems. The FLORES-GEMBA average versus the HardMTBench-GEMBA average yields $\rho = 0.881$, while the WMT25 en-zh GEMBA versus the HardMTBench en-zh GEMBA yields $\rho = 0.778$. On xCOMET-XXL, the corresponding correlations are $\rho = 0.910$ and $\rho = 0.684$. While the correlations are positive, they are far from a perfect ranking, and the gap is especially visible in the middle band where the most informative comparisons for model selection happen.

The largest rank movements are illustrative. Gemma4-31B ranks fourth on FLORES-GEMBA zh-en but sixth on HardMTBench-GEMBA zh-

en. Microsoft Translator ranks 20th on FLORES en-zh but 22nd on HardMTBench en-zh, with an absolute gap of 12.7 GEMBA points opening up on the harder benchmark. Hy-MT2-30B-A3B drops from a strong FLORES profile (fourth on xCOMET zh-en) to a mid-table position on HardMTBench-GEMBA but climbs back to the top of HardMTBench-xCOMET, which indicates that the two metrics can reorder the same system on the same data. Figure 3 traces all 22 systems between the two benchmarks. The most pronounced shifts ($\geq$4 ranks) are highlighted in saturated colour, while small reorderings remain visible but muted. Taken together, the rank reorderings suggest that performance on FLORES-200 or WMT25 alone is an incomplete proxy for capability on knowledge-intensive domain translation.

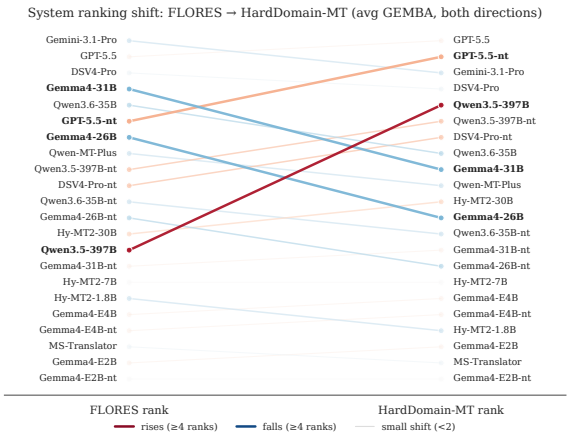

Figure 3: System ranking shift from FLORES-200 to HardMTBench under GEMBA-DA averaged over zh-en and en-zh directions. Lines for systems with $\geq$4 rank positions of movement are drawn in bold and labelled in bold typeface, lines for shifts of $\geq$2 positions are drawn at intermediate strength, and the remaining lines are drawn in light grey.

3.5 Domain-Level Analysis

Table 4 reports cross-system averages per domain on HardMTBench, aggregated over zh-en and en-zh directions. GEMBA-DA and xCOMET-XXL are reported alongside terminology accuracy, which uses the per-sample terminology annotations described in Section 2.1.

We highlight three observations. First, translation quality and terminology accuracy do not move together. Education has a high GEMBA score of 90.31 but a low terminology accuracy of 49.50, which indicates that translations can read smoothly while missing the controlled vocabulary that a domain practitioner expects. Second, the same gap appears in the opposite direction on Sci. & Tech., where GEMBA is moderate (87.67) but terminology accuracy is the highest of all domains at 66.07, supported by stable standard naming for technical concepts. Third, History sits at the bottom on all three metrics, with GEMBA 75.67, xCOMET 39.03 and terminology accuracy 44.79. The combination of historical document language, person names, classical official titles and culturally loaded vocabulary remains the hardest region for current systems. Figure 5 (Section 3.7) shows the analogous picture for terminology accuracy across systems and domains.

Domain	GEMBA	xCOMET	Term. Acc.
Finance	92.04	75.46	58.45
Law	91.92	77.73	58.64
Healthcare	90.61	75.83	61.44
Education	90.31	70.74	49.50
News	89.25	66.49	56.37
Military	87.97	71.79	53.59
Sci. & Tech.	87.67	72.65	66.07
Sports	87.43	67.38	55.50
Books / Papers	86.67	67.65	56.03
Gaming	85.31	59.29	45.97
Media	82.83	62.35	54.57
History	75.67	39.03	44.79

Table 4: Cross-system averages per domain on HardMTBench, averaged over zh-en and en-zh directions. xCOMET-XXL is reported as a percentage, and terminology accuracy is the percentage of annotated terminology pairs whose target term appears in the translation after surface normalisation.

3.6 Hardness as a Difficulty Proxy

The per-sample hardness score H from Equation 1 lets us probe the relationship between difficulty and translation quality directly. We group the 22 systems into three tiers by their mean GEMBA-DA on HardMTBench (top tier of 7 systems including GPT-5.5 and Gemini 3.1 Pro, mid tier of 8 systems, and lower tier of 7 systems such as Hy-MT2-1.8B and Microsoft Translator) and partition the test set into equal-width hardness buckets of five points. From buckets 55–60 to 85–90, the top-tier mean GEMBA-DA falls by 5.4 points (91.4 to 86.0), the mid tier by 9.1 points (89.1 to 80.0), and the lower tier by 18.4 points (81.4 to 63.0). The widening gap with hardness suggests that hard samples carry the most discriminative signal across systems.

Figure 4 shows cross-system means across four hardness buckets (we merge the sparse 50–60 segment with 60–70 since it contains only 434 items per system and is not statistically informative on its own) together with the per-system distribution in

the hardest bucket. The relationship is not strictly monotone in the easy half of the range: bucket 50–70 sits slightly below bucket 70–80 on GEMBA-DA, which reflects that the hardness score H aggregates domain knowledge density, translation difficulty and terminology density rather than predicting quality directly. The pattern past hardness 80 is more pronounced. From bucket 70–80 to bucket 90–100, GEMBA-DA drops by 15.2 points (89.30 to 74.13), xCOMET-XXL by 30.4 points (71.12 to 40.77), and term accuracy by 6.4 points (56.45 to 50.07), and the cross-system spread on GEMBA-DA grows from 11.3 to 30.9. The 90–100 bucket carries an average of only 80 items per system, so its absolute values should be read as indicative. The across-bucket trend, however, is also supported by the much larger 80–90 bucket.

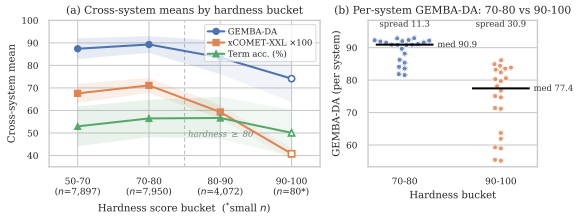

Figure 4: Hardness bucket analysis. (a) Cross-system means of GEMBA-DA, xCOMET-XXL ($\times 100$) and term accuracy across four hardness buckets (± 1 SD). The dashed line marks hardness ≥ 80. Hollow markers at the 90–100 endpoint denote the small-sample bucket ($n = 80$ per system) and use a lighter SD band. (b) Per-system GEMBA-DA for bucket 70–80 vs 90–100, with medians (black bars) and per-bin spread. The cross-system spread grows from 11.3 to 30.9 between the two buckets.

3.7 Term-Level Diagnostic and Joint Reporting

Quality scores compress a translation into a single number and tend to under-report failures that are localised to a few content-bearing tokens. We complement GEMBA-DA and xCOMET-XXL with a terminology-level diagnostic computed from the same span annotations that drive the hardness score. For each item, term accuracy is the fraction of source-side terminology spans whose target-side surface forms appear in the system output after light normalisation. This is a strict proxy that does not allow paraphrasing and should be read as a lower bound on terminology fidelity. Figure 5 shows term accuracy for all 22 systems across the 12 domains, with marginal averages on the right and at the bottom.

The term-level signal reorders the leaderboard noticeably. Gemini 3.1 Pro tops the term accuracy ranking at 64.52% despite trailing GPT-5.5 on overall GEMBA-DA. The Hy-MT2 family illustrates the same point from below: Hy-MT2-30B-A3B reaches 89.15 on HardMTBench GEMBA-DA but only 53.62% on term accuracy, and Hy-MT2-7B records the largest gap between strong xCOMET-XXL (71.4) and weak term accuracy (43.58%) among all systems. The cross-system range on term accuracy is 25.0 points (39.50%–64.52%), which is larger than the 14–16 point GEMBA-DA range on the two HardMTBench directions and much larger than the 7–8 point range on FLORES-200. Terminology compliance and overall fluency-adequacy do not move together on hard-domain text.

Beyond terminology, GEMBA-DA and xCOMET-XXL themselves pick up partially complementary signals. Their Spearman correlation across the 22 systems on HardMTBench averaged over directions is $\rho = 0.704$, which is meaningfully below the within-FLORES correlation. When all 44 (system, direction) cells are treated separately, the per-cell Spearman drops to $\rho = 0.545$, and the en-zh cloud sits above the zh-en cloud on the xCOMET axis with mean xCOMET values of 0.709 and 0.635 respectively. This 7.4-point gap between directions does not appear on FLORES-200 to anywhere near the same extent and indicates a direction-dependent behaviour of xCOMET-XXL on out-of-distribution domain text. We therefore recommend that HardMTBench reports include three signals together: GEMBA-DA for fluency-adequacy, xCOMET-XXL for an independent reference-based view (interpreted alongside the direction), and term accuracy for terminology fidelity. Either of the first two metrics alone leaves a non-trivial fraction of the diagnostic signal unobserved.

4 Related Work

General multilingual MT benchmarks. FLORES-200 (Team et al., 2022), NTREX-128 (Federmann et al., 2022) and the WMT general tasks (Kocmi et al., 2023, 2024) provide broad coverage with news and Wikipedia content. WMT24++ (Deutsch et al., 2025) expands to 55 languages and Taguchi et al. (2025) document remaining gaps for low-resource languages even after FLORES+ extensions. The per-system spread

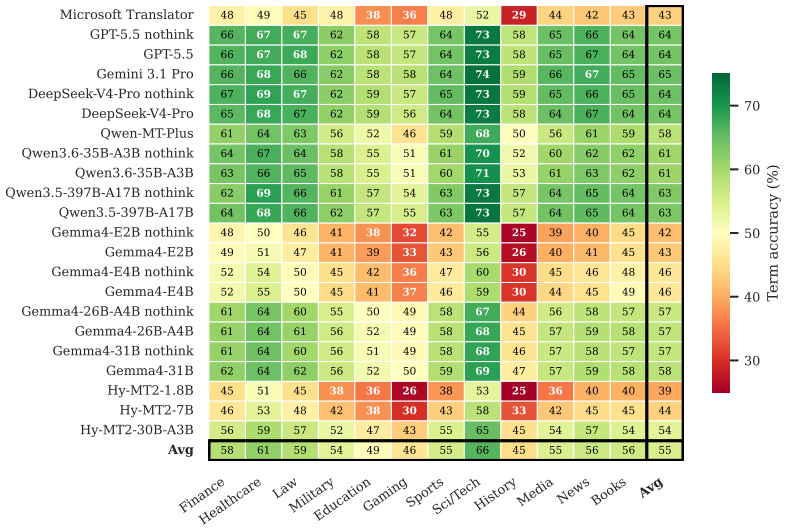

Figure 5: Term accuracy (%) on 22 systems × 12 HardMTBench domains, with marginal row/column means. Closed/API systems (GPT-5.5, Gemini 3.1 Pro, DeepSeek-V4-Pro) cluster around 64%, while smaller open systems sit in the 40–46% range. Sci.&Tech. is easiest on terminology, History and Gaming are hardest.

on the dominant pairs is now compressed. Top systems on FLORES Chinese-English cluster within 7–8 GEMBA points, which constrains the headroom for further diagnosis. HardMTBench targets the difficult tail of Chinese-English rather than language breadth.

Domain, terminology and robust MT benchmarks. WMT Biomedical (Neves et al., 2022), TICO-19 (Anastasopoulos et al., 2020) and the WMT24 Discourse-Level Literary track (Wang et al., 2024) target individual domains, and recent work also covers multi-domain settings (Hu et al., 2024), cultural and dialectal coverage (Van Doren et al., 2026; Yu et al., 2024) and Chinese social-media text (Zhao et al., 2026). The WMT terminology task (Alam et al., 2021; Dinu et al., 2019) evaluates whether systems follow provided dictionaries and MTNT (Michel and Neubig, 2018) stresses noisy user-generated text. HardMTBench differs by spanning 12 difficulty-screened domains under a unified pipeline with balanced sub-domain quotas, evaluating natural translation of terminology-rich text without supplying term lists at inference while still annotating terminology so downstream metrics can be measured.

LLM-era MT evaluation and difficulty-aware testing. LLM-as-judge protocols such as GEMBA-DA (Kocmi and Federmann, 2023b) and GEMBA-MQM (Kocmi and Federmann, 2023a) complement reference-based metrics, with the WMT24 metrics task (Freitag et al., 2024) and Schmidt et al. (2026) documenting failure modes of automatic metrics on unseen and out-of-distribution domains, while Issaka et al. (2026) use translation as a scalable proxy for broader multilingual evaluation. Kocmi et al. (2024) note that the LLM era compresses scores on standard sets while leaving failure modes intact. HardMTBench combines LLM-based scoring with rule-based features (terminology density, structural depth) and stratifies across 12 domains, and we report GEMBA-DA and xCOMET-XXL jointly because Section 3.6 shows they provide complementary signals on difficult text.

5 Conclusion

We presented HardMTBench, a difficulty-aware diagnostic benchmark for bidirectional Chinese-English domain translation, covering 12 knowledge-intensive domains with 10,000 hand-curated pairs (20,000 directional items), a transparent hardness score and terminology annotations. Across 22 systems on FLORES-200, WMT25 and HardMTBench, HardMTBench roughly doubles the cross-system GEMBA spread of FLORES-200, reorders system rankings, and surfaces domain-specific weaknesses in terminology compliance and historical-text translation. We hope HardMTBench complements rather than replaces FLORES-200 and WMT, and that the construction pipeline transfers to other language pairs and knowledge-intensive domains.

Limitations

HardMTBench currently covers only Chinese-English, and the construction pipeline has not been validated on other language pairs. The 12 chosen domains reflect Chinese deployment scenarios and may underrepresent regions that are central in other markets. A subset of the released material is drawn from sources that are time-sensitive, for example news and audiovisual media, and the test set will need refresh cycles to remain representative. The hardness fusion rule in Equation 1 is a transparent baseline rather than a learned ranker, and alternative weightings or learned hardness models could change the difficulty ordering. The two reported metrics, GEMBA-DA with gpt-oss-120b as the judge and xCOMET-XXL, inherit the biases of their underlying models, and human meta-evaluation of these metrics on hard domain text remains an open direction.

Ethics Statement

The source material in HardMTBench is drawn from publicly available Chinese source text and is filtered to remove personal data, sensitive content and low-quality samples through the Stage 2 LLM judge and the human verification round described in Section 2.2. Human annotators were trained bilingual speakers, were informed of the purpose of the work and received standard compensation. The release includes only aggregated system scores and the test set itself, and does not include individual model outputs that could enable re-identification of evaluated systems beyond their official release names.

References

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Ondřej Bojar, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Kevin Knight, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Matteo Negri, Santanu Pal, Lucia Specia, Veselin Stoyanov, Jörg Tiedemann, and Marco Turchi. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, William Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at EMNLP*.

DeepSeek-AI. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556. DeepSeek-V4-Pro variants accessed via the official DeepSeek API.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. *Preprint*, arXiv:2502.12404.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 3063–3068. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 - news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, and 1 others. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

Google DeepMind. 2025a. Gemini 3 Pro model card. Google DeepMind model card. Accessed via the official Gemini API. No standalone arXiv technical report is available; this entry references the public model card and product documentation.

Google DeepMind. 2025b. Gemma 3 technical report. *Preprint*, arXiv:2503.19786. Gemma 4 model variants accessed via the official release channels; this entry references the most recent published Gemma technical report.

Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Sheriff Issaka, Erick Rosas Gonzalez, Lieqi Liu, Evans Kofi Agyei, Lucas Bandarkar, Nanyun Peng, David Ifeoluwa Adelani, Francisco Guzmán, and Saadia Gabriel. 2026. Translation as a scalable proxy for multilingual evaluation. *CoRR*, abs/2601.11778.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 193–203. European Association for Machine Translation.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, and Philippe Thomas. 2022. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation*.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2026. OpenAI GPT-5 system card. *Preprint*, arXiv:2601.03267. GPT-5.5 variants accessed via the official OpenAI API.

Finn Schmidt, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2026. Who watches the watchmen? humans disagree with translation metrics on unseen domains. *CoRR*, abs/2604.17393.

Chihiro Taguchi, Seng Mai, Keita Kurabe, Yusuke Sakai, Georgina Agyei, Soudabeh Eslami, and David Chiang. 2025. Languages still left behind: Toward a better multilingual machine translation benchmark. *CoRR*, abs/2508.20511.

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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Madison Van Doren, Casey Ford, Jennifer Barajas, and Cory Holland. 2026. "be my cheese?": Cultural nuance benchmarking for machine translation in multilingual LLMs. *CoRR*, abs/2602.04729.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu Chinese: Workflow and analysis. *CoRR*, abs/2410.10278.

Kaiyan Zhao, Zheyong Xie, Zhongtao Miao, Xinze Lyu, Yao Hu, and Shaosheng Cao. 2026. Benchmarking machine translation on Chinese social media texts. *CoRR*, abs/2601.22931.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. 2026. Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild. *Preprint*, arXiv:2605.22064.