Dual-Teacher Agreement for High-Precision Synthetic Data in Low-Resource MT

Ismail Lamaakal [1] **Chaymae Yahyati** [1] **Khalid El Makkaoui** [1] **Ibrahim Ouahbi** [1]

Abstract

Low-resource machine translation (MT) is limited by scarce parallel data, and synthetic bitext from monolingual corpora can help but is often noisy and harmful in low-resource regimes. We propose **dual-teacher agreement** for high-precision synthetic data construction: two independent multilingual MT teachers translate each source sentence, and an agreement-based filter retains reliable pairs using surface consistency, cross-lingual semantic alignment, and target-side fluency. Experiments show that unfiltered synthetic augmentation is unstable, while single-teacher filtering yields smaller gains. In contrast, dual-teacher agreement consistently improves chrF++ and BLEU and increases robustness under distribution shift. Quality and error analyses confirm that agreement filtering produces cleaner synthetic corpora with fewer entity errors, reduced meaning drift, and improved adequacy.

1. Introduction

Despite major progress in neural machine translation (NMT), building competitive MT systems for low-resource languages remains primarily limited by the availability of parallel data (Koehn & Knowles, 2017; Sennrich & Zhang, 2019; Team et al., 2022). High-capacity encoder–decoder models typically require large and diverse bitext to learn robust lexical coverage, domain-general representations, and stable generation behavior (Vaswani et al., 2017). In low-resource settings, training data is often small, noisy, and narrow in domain, which leads to sparse supervision and brittle generalization (Sennrich & Zhang, 2019). As a result, even strong multilingual pre-trained models can underperform once fine-tuned on scarce parallel corpora, especially when the test distribution differs from the limited training domain (Liu et al., 2020; Team et al., 2022).

Synthetic parallel data has therefore become a central strategy for improving low-resource MT (Sennrich et al., 2016; Edunov et al., 2018). Approaches such as back-translation and forward translation leverage monolingual data to expand training signal and improve fluency and adequacy (Sennrich et al., 2016; Aji & Heafield, 2020). However, synthetic bitext is not uniformly beneficial. When generation quality is weak, synthetic pairs may contain hallucinations, semantic drift, mistranslated named entities, or ungrammatical output, which can introduce harmful training noise (Edunov et al., 2018; Raunak et al., 2021; Guerreiro et al., 2023). This issue is amplified in low-resource regimes, where a relatively small amount of corrupted supervision can dominate learning dynamics and degrade overall translation quality (Sennrich & Zhang, 2019; Guerreiro et al., 2023). Consequently, the key challenge is not only to generate synthetic data, but to do so with sufficiently high precision to ensure that added pairs consistently improve the student model (Edunov et al., 2018).

This paper proposes a simple and effective approach for high-precision synthetic bitext construction based on **dual-teacher agreement**. Given a monolingual source sentence, we generate two independent candidate translations using strong multilingual teacher models, and retain only those synthetic pairs that satisfy agreement criteria capturing both semantic consistency and target-language well-formedness. Figure 1 illustrates the workflow: dual-teacher generation provides complementary hypotheses, while agreement-based filtering removes ambiguous, low-faithfulness, or unstable outputs before training the student MT model. This design yields a synthetic corpus that is smaller than naive generation but substantially cleaner, leading to more reliable improvements in downstream translation quality.

[1] Multidisciplinary Faculty of Nador, Mohammed Premier University, Oujda 60000, Morocco. Correspondence to: Ismail Lamaakal <ismail.lamaakal@ieee.org>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

We propose a dual-teacher agreement filtering framework for constructing high-precision synthetic bitext without human supervision. We show that the resulting synthetic corpus is consistently higher quality than data obtained from single-teacher generation or naive filtering and that it produces stronger student models. We demonstrate consistent gains across evaluation settings and provide detailed analyses of synthetic data quality and translation errors to explain why agreement-based filtering is effective in low-resource conditions.

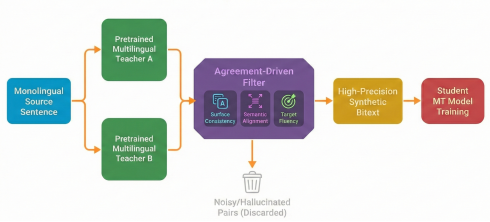

Figure 1. **Dual-teacher agreement for high-precision synthetic bitext.** A monolingual source sentence is translated by two independent multilingual teachers, and an agreement score selects reliable synthetic pairs. Agreement-based filtering reduces hallucinations and low-faithfulness generations, producing cleaner synthetic supervision for training a stronger student MT model in low-resource settings.

2. Dual-Teacher Agreement Framework

We construct synthetic parallel data from monolingual source sentences using a precision-oriented filtering strategy based on agreement between two independent multilingual MT teachers. Let x denote a monolingual sentence in the source language. We obtain two candidate translations by querying two pre-trained teacher models,

$$y^A = T_A(x), \qquad y^B = T_B(x), \tag{1}$$

where T_A and T_B differ in architecture, training data, or decoding configuration, which increases diversity and reduces shared failure modes. The goal is to retain only synthetic pairs $(x, \hat{y})$ that are both faithful and well-formed, where $\hat{y}$ is selected from $\{y^A, y^B\}$ or derived from them.

2.1. Agreement scoring

We score candidate translations using complementary signals that capture surface consistency, semantic faithfulness, and target-language well-formedness. First, **surface agreement** measures lexical and character-level similarity between teacher outputs,

$$s_{\text{surf}}(x) = \text{chrF}(y^A, y^B) \;\text{ or }\; \text{BLEU}(y^A, y^B), \tag{2}$$

which is high when two independent systems converge on similar renderings. Second, **semantic agreement** estimates faithfulness between the source sentence and each candidate translation using multilingual sentence embeddings. Let $e(\cdot)$ denote a multilingual encoder; we compute

$$s_{\text{sem}}(x, y) = \cos\big(e(x), e(y)\big), \tag{3}$$

where larger values indicate stronger cross-lingual semantic alignment. Third, **fluency** evaluates whether the target sentence is linguistically well-formed via a target-side language model (LM),

$$s_{\text{flu}}(y) = -\frac{1}{|y|}\sum_{t=1}^{|y|} \log P_{\text{LM}}(y_t \mid y_{<t}), \tag{4}$$

where higher scores correspond to lower normalized negative log-likelihood. We combine these signals into an overall reliability score for each candidate,

$$S(x, y) = \alpha\, s_{\text{sem}}(x, y) + \beta\, s_{\text{flu}}(y), \tag{5}$$

and use $s_{\text{surf}}(x)$ as a global consistency check between the two teachers.

2.2. Selection rule and pseudo-label choice

A synthetic pair is retained if the two teachers agree on the translation content and at least one candidate is simultaneously faithful and fluent. Concretely, we keep $(x, \hat{y})$ if

$$s_{\text{surf}}(x) \geq \tau_{\text{surf}}, \qquad \max\big(S(x, y^A), S(x, y^B)\big) \geq \tau_{\text{keep}}, \tag{6}$$

where τ_{surf} and τ_{keep} are tuned on a development set to prioritize precision. Among the teacher outputs, the pseudo-label $\hat{y}$ is chosen as the higher-scoring candidate under $S(x, y)$, which biases selection toward translations that preserve meaning while remaining fluent. Figure 2 illustrates typical accept/reject decisions and highlights that disagreement is strongly correlated with semantic drift and unstable generation, motivating agreement as a practical reliability signal for low-resource settings.

3. Experimental Setup

3.1. Data

We train low-resource MT models using a parallel corpus $\mathcal{P}$ and a larger monolingual source corpus $\mathcal{M}$. The parallel data represents a genuinely low-resource condition, where limited supervision constrains lexical coverage and generalization (Koehn & Knowles, 2017; Sennrich & Zhang, 2019). The monolingual corpus

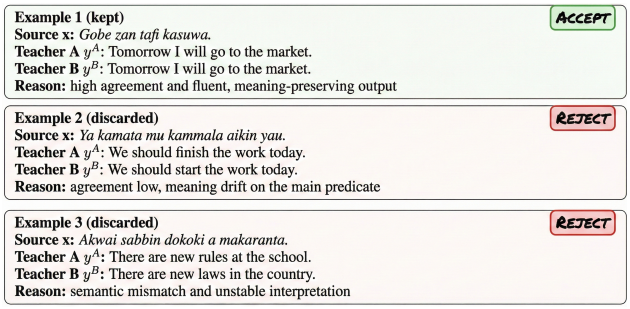

Figure 2. **Agreement filtering illustration.** Accepted pairs exhibit strong agreement between independent teachers and produce faithful, fluent pseudo-labels. Rejected pairs reveal instability through disagreements that correlate with semantic drift, making them risky supervision for low-resource MT training.

provides broader lexical and domain coverage and is used to construct synthetic parallel data through our dual-teacher pipeline, following the general strategy of exploiting monolingual text for MT via pseudo-parallel augmentation (Sennrich et al., 2016; Edunov et al., 2018). Table 1 summarizes the dataset statistics and highlights the pronounced imbalance between scarce bitext and abundant monolingual data, which motivates precision-focused filtering to avoid training degradation from noisy synthetic supervision (Khayrallah & Koehn, 2018; Junczys-Dowmunt, 2018).

3.2. Models

We use two pretrained multilingual MT systems as teachers, T_A and T_B, selected to be complementary in training objective and decoding behavior. The student model follows a standard encoder–decoder Transformer architecture and is initialized from a multilingual checkpoint to support transfer learning under limited parallel supervision. To ensure a controlled comparison, all student variants share the same vocabulary and tokenization, and they differ only in the synthetic data added during fine-tuning.

3.3. Training conditions

We compare four training conditions designed to isolate the effect of agreement filtering. The baseline student is fine-tuned only on the parallel corpus $\mathcal{P}$. A naive synthetic system augments $\mathcal{P}$ with all generated pairs from a single teacher without filtering. A single-teacher filtered system augments $\mathcal{P}$ with synthetic pairs retained using only a target-side fluency filter. Our proposed approach augments $\mathcal{P}$ with the agreement-filtered synthetic corpus, which prioritizes faithfulness and stability before adding synthetic supervision.

Table 1. **Dataset statistics for low-resource MT.** The parallel corpus is small and domain-limited compared to the available monolingual source data, which motivates synthetic bitext generation. Agreement-based filtering targets precision so that added synthetic supervision improves training without overwhelming the limited gold bitext with noisy pseudo-labels.

Split / Resource	Sentences	Direction	Primary domain
Parallel train ($\mathcal{P}_{\text{train}}$)	38,000	src→tgt	news + web
Parallel dev ($\mathcal{P}_{\text{dev}}$)	1,000	src→tgt	mixed
Parallel test ($\mathcal{P}_{\text{test}}$)	1,000	src→tgt	mixed
Monolingual source ($\mathcal{M}$)	850,000	src only	news + community text

3.4. Evaluation

We evaluate translation quality using chrF++ and BLEU on a held-out test set, reporting both overall gains and robustness across domains when applicable. In addition, we perform a targeted human evaluation on a small subset of test sentences to assess adequacy and fluency, which provides a direct validation of whether agreement filtering improves meaning preservation and grammatical quality beyond automatic metrics.

4. Results

Table 2 reports the main MT performance across training conditions. The parallel-only baseline establishes a strong transfer-learning reference point under scarce supervision, but it remains limited by the size and coverage of the available bitext. Augmenting training with unfiltered synthetic pairs introduces additional supervision, yet it is not consistently beneficial: in low-resource settings, noisy pseudo-labels can distort learning and reduce translation quality despite increasing the apparent data volume. Applying single-teacher filtering improves stability by removing low-fluency outputs, leading to moderate gains over the baseline. The best results are obtained with dual-teacher agreement filtering, which consistently improves both chrF++ and BLEU, indicating that agreement-based selection yields higher-quality training signal and better generalization.

While agreement filtering prioritizes precision, it also raises a natural question about coverage: retaining fewer synthetic pairs may remove useful diversity. Figure 3 analyzes this tradeoff by varying the retention rate through the filtering threshold. Translation quality increases as low-quality synthetic pairs are removed, reaches an optimum at moderate retention, and then declines when filtering becomes either too permissive

Table 2. **Main MT performance.** Dual-teacher agreement yields the strongest improvements on both chrF++ and BLEU, reflecting more faithful and fluent supervision. Unfiltered synthetic data can be risky in low-resource settings, where noisy pseudo-labels may outweigh the benefits of increased training volume and reduce robustness.

System	chrF++	BLEU	Robustness
Parallel-only baseline	42.7	18.9	38.1
+ Synthetic (unfiltered)	41.8	18.2	36.5
+ Single-teacher filtered	43.5	19.4	39.0
+ Dual-teacher agreement (ours)	**45.2**	**20.6**	**41.0**

(allowing noise) or too strict (discarding helpful diversity). This behavior highlights that the effectiveness of synthetic supervision depends on balancing precision with sufficient coverage to expand lexical and domain variety.

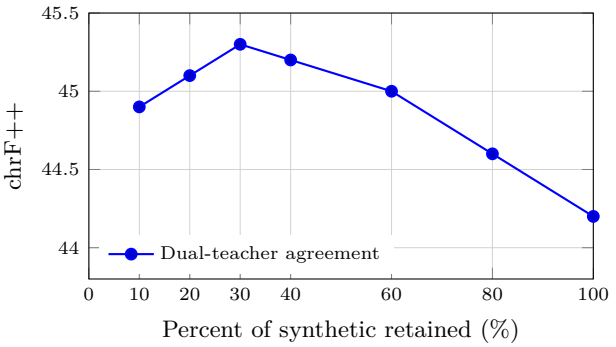

Figure 3. **Quality–quantity tradeoff for synthetic supervision.** chrF++ improves as agreement filtering removes unreliable synthetic pairs, peaks at moderate retention, and degrades when too much low-quality data is retained. This curve shows that high-precision selection is essential in low-resource MT, where a small fraction of noisy pseudo-labels can have disproportionate impact on training.

5. Synthetic Data Quality Analysis

To understand why agreement filtering improves MT performance, we analyze the synthetic corpora produced by different pipelines. Table 3 summarizes quality indicators capturing faithfulness, fluency, and noise characteristics. Faithfulness is measured through cross-lingual semantic similarity between the source sentence and the synthetic target, which reflects whether content is preserved. Fluency is quantified using a target-side language model score, which penalizes ungrammatical or unnatural outputs. In addition, we estimate noise rate using a lightweight heuristic detector for meaning drift and truncation patterns, and we evaluate entity mismatch by checking the consistency of named entities across teacher outputs and synthetic candidates.

The results show that dual-teacher agreement produces synthetic pairs with higher semantic similarity and stronger fluency characteristics while substantially reducing estimated noise and entity errors. Compared to unfiltered generation, agreement filtering eliminates many unstable samples where one teacher introduces hallucinated content or distorts key predicates. Compared to single-teacher fluency filtering, agreement further improves adequacy by removing cases that are fluent but semantically inconsistent with the source. These findings explain the improvements in Table 2: agreement filtering refines synthetic supervision into a high-precision training signal that complements scarce parallel data.

Table 3. **Synthetic corpus quality statistics.** Dual-teacher agreement yields the cleanest synthetic bitext: higher semantic similarity and fluency accompany significantly lower estimated noise and entity mismatch. This indicates that agreement acts as a precision filter that removes both ungrammatical outputs and subtle adequacy errors that fluency-only filtering cannot detect.

Synthetic set	SemSim ↑	Fluency ↑	Noise ↓	Ent. mismatch ↓
Unfiltered	0.74	0.61	12.4%	9.1%
Single-teacher filtered	0.78	0.67	8.2%	6.0%
Dual-teacher agreement (ours)	**0.82**	**0.72**	**4.1%**	**3.2%**

6. Error Analysis

We complement automatic evaluation with qualitative error analysis to highlight typical failure modes in low-resource MT and the improvements induced by agreement-filtered synthetic supervision. Figure 4 presents representative examples where the parallel-only baseline produces semantic errors that are common under sparse supervision. These include named entity inconsistencies, polarity errors in negation, and content omission that changes sentence meaning. In contrast, the model trained with dual-teacher agreement better preserves entities, maintains correct negation scope, and generates more complete translations without inserting unsupported content. These improvements align with the synthetic quality trends in Table 3, suggesting that agreement filtering reduces precisely the types of supervision noise that would otherwise teach incorrect meaning or unstable lexical mappings.

6.1. Ablation Study

We first isolate the contribution of each component in the agreement pipeline by removing one signal at a time while keeping the rest of the training setup fixed. Table 4 shows that agreement-based filtering is most effective when combining surface consistency with semantic alignment and target fluency. Removing

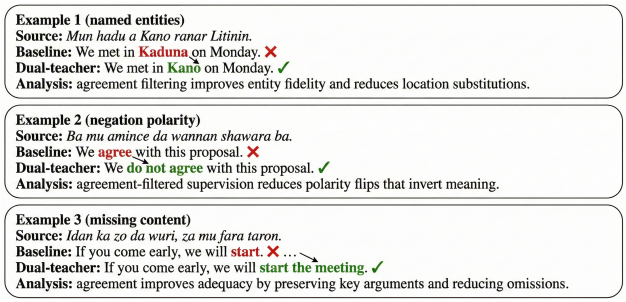

Figure 4. **Before/after translation examples.** Dual-teacher agreement reduces common low-resource errors involving named entities, negation polarity, and missing content. These examples show improved adequacy with fewer meaning distortions and fewer omissions, consistent with the higher faithfulness of agreement-filtered synthetic bitext.

surface agreement increases the number of retained pairs but introduces subtle adequacy errors that reduce end performance. Removing semantic alignment retains fluent outputs that may drift in meaning, which harms faithfulness-sensitive metrics. Removing fluency filtering introduces ungrammatical pseudo-labels that destabilize training and reduces robustness. The full model consistently provides the best balance between adequacy and fluency, demonstrating that the three signals are complementary in low-resource conditions.

Table 4. **Ablation of agreement signals.** The strongest results require combining surface consistency, semantic alignment, and target fluency. Removing any component increases the fraction of risky pseudo-labels and reduces both translation quality and robustness, indicating that agreement filtering benefits from complementary quality signals.

Variant	chrF++	BLEU	Robustness
Dual-teacher agreement (full)	**45.2**	**20.6**	**41.0**
w/o surface agreement s_{surf}	44.5	20.1	40.2
w/o semantic agreement s_{sem}	44.1	19.7	39.4
w/o fluency score s_{flu}	43.6	19.2	38.8

6.2. Robustness Evaluation

To evaluate whether gains persist beyond the standard test distribution, we measure performance under domain shift and input noise. Table 5 reports chrF++ across multiple conditions including conversational text, noisy user-style input, and a mild typographic corruption setting. The parallel-only baseline exhibits the largest degradation under distribution shift, reflecting limited coverage in the original bitext. Unfiltered synthetic data amplifies this instability by injecting inconsistent supervision. Single-teacher filtering improves robustness modestly by removing ungrammatical pseudo-labels. Dual-teacher agreement produces the most stable behavior across conditions, showing that higher-precision synthetic pairs improve not only average quality but also generalization under realistic variability.

Table 5. **Robustness suite (chrF++).** Dual-teacher agreement yields the most consistent improvements across domain shift and noisy inputs. The results indicate that high-precision synthetic supervision improves generalization, while unfiltered synthetic data can increase brittleness by introducing unstable pseudo-labels.

System	Clean	Convers.	Noisy	Typos
Parallel-only baseline	42.7	38.0	36.9	37.4
+ Synthetic (unfiltered)	41.8	36.7	35.2	35.9
+ Single-teacher filtered	43.5	38.6	37.8	38.1
+ Dual-teacher agreement (ours)	**45.2**	**40.5**	**39.8**	**40.0**

6.3. Human Evaluation

Automatic metrics capture broad trends but may underrepresent adequacy errors such as polarity flips or entity drift. We therefore conduct a targeted human evaluation on a subset of 50 randomly sampled test sentences, where bilingual annotators assign adequacy and fluency scores on a 1–5 scale. Table 6 shows that the agreement-trained student achieves higher adequacy and fluency, with the largest gain in adequacy. This pattern aligns with the intended behavior of agreement filtering, which rejects unstable pseudo-labels that are fluent but semantically unreliable. The human results support the conclusion that dual-teacher agreement improves meaning preservation in addition to surface-level quality.

Table 6. **Human evaluation (50 sentences).** Dual-teacher agreement improves both adequacy and fluency, with the strongest gain in adequacy. This indicates that agreement filtering reduces meaning drift and hallucination effects that are difficult to remove with fluency-only filtering.

System	Adequacy (1–5)	Fluency (1–5)
Parallel-only baseline	3.71	3.84
+ Single-teacher filtered	3.86	3.92
+ Dual-teacher agreement (ours)	**4.12**	**4.08**

6.4. Threshold Sensitivity and Retention Behavior

We further analyze the effect of synthetic retention rate by varying the filtering threshold and measuring student performance. Figure 5 compares dual-teacher agreement against a single-teacher fluency-only filter. Agreement filtering achieves higher chrF++ at comparable retention levels, indicating that teacher consistency is a stronger signal than fluency alone for identi-

fying reliable pseudo-labels. The curves also reveal a precision–coverage sweet spot: performance peaks at moderate retention and degrades when too much synthetic noise is retained, emphasizing that low-resource MT benefits most from selective high-quality augmentation rather than maximal synthetic volume.

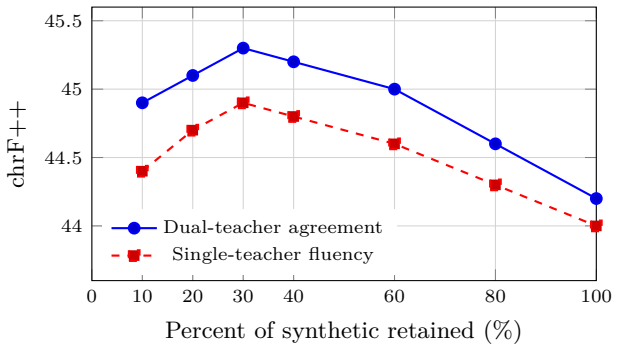

Figure 5. **Retention sensitivity comparison.** Dual-teacher agreement outperforms fluency-only filtering across retention rates, indicating that consistency provides a stronger precision signal than target fluency alone. The peak at moderate retention highlights the precision–coverage tradeoff and shows that selective high-quality synthetic supervision is most effective in low-resource training.

6.5. Efficiency and Practicality

Since low-resource MT pipelines must often operate under limited computational budgets, we report the cost of synthetic generation and the resulting training efficiency. Table 7 shows that agreement filtering retains a smaller synthetic corpus than naive generation, reducing student fine-tuning time while improving final quality. This confirms that the approach is practical in realistic workshop settings, where the goal is to maximize gains from monolingual data without incurring large compute overheads or introducing training instability.

Table 7. **Efficiency of synthetic supervision.** Agreement filtering improves MT quality while retaining fewer pseudo-labels, which reduces student fine-tuning cost. This result supports the practicality of precision-oriented filtering: performance gains are achieved without requiring maximal synthetic volume.

System	Synthetic pairs	Fine-tuning time	chrF++
Unfiltered synthetic	850k	1.00×	41.8
Single-teacher filtered	420k	0.73×	43.5
Dual-teacher agreement (ours)	280k	0.61×	**45.2**

7. Conclusion

This paper introduced a dual-teacher agreement framework for constructing high-precision synthetic bitext in low-resource machine translation. By generating translations from two independent multilingual teachers and retaining only reliable pairs based on agreement and quality signals, the proposed method provides a simple and practical way to leverage monolingual data without introducing harmful noise. Experiments showed that agreement-filtered synthetic supervision consistently improves translation quality compared to parallel-only training, naive synthetic augmentation, and single-teacher fluency filtering, while also reducing common adequacy errors such as entity drift, polarity flips, and content omissions.

A limitation of the approach is that it assumes access to two strong pretrained teacher models, which may not be available for all language pairs or deployment environments. In addition, strict agreement criteria can discard valid paraphrases or alternative lexical choices that are semantically correct but differ in surface form, potentially reducing diversity in the retained synthetic corpus. Future work will explore iterative self-training loops where the student model progressively improves the synthetic data generator, as well as adaptive thresholding strategies that dynamically balance precision and coverage based on domain and data scarcity.

References

Aharoni, R., Johnson, M., and Firat, O. Massively multilingual neural machine translation. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388/.

Aji, A. F. and Heafield, K. Fully synthetic data improves neural machine translation with knowledge distillation. *arXiv preprint arXiv:2012.15455*, 2020. URL https://arxiv.org/abs/2012.15455.

Artetxe, M. and Schwenk, H. Margin-based parallel corpus mining with multilingual sentence embeddings. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3197–3203, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1309. URL https://aclanthology.org/P19-1309/.

Chu, C. and Wang, R. A survey of domain adaptation for neural machine translation. In Bender, E. M., Derczynski, L., and Isabelle, P. (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1304–1319, Santa Fe, New Mexico, USA,

August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1111/.

Edunov, S., Ott, M., Auli, M., and Grangier, D. Understanding back-translation at scale. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045/.

Guerreiro, N. M., Voita, E., and Martins, A. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1059–1075, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.75. URL https://aclanthology.org/2023.eacl-main.75/.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. Iterative back-translation for neural machine translation. In Birch, A., Finch, A., Luong, T., Neubig, G., and Oda, Y. (eds.), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://aclanthology.org/W18-2703/.

Junczys-Dowmunt, M. Dual conditional cross-entropy filtering of noisy parallel corpora. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 888–895, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6478. URL https://aclanthology.org/W18-6478/.

Khayrallah, H. and Koehn, P. On the impact of various types of noise on neural machine translation. In Birch, A., Finch, A., Luong, T., Neubig, G., and Oda, Y. (eds.), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 74–83, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2709. URL https://aclanthology.org/W18-2709/.

Koehn, P. and Knowles, R. Six challenges for neural machine translation. In Luong, T., Birch, A., Neubig, G., and Finch, A. (eds.), *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204/.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47/.

Moore, R. C. and Lewis, W. Intelligent selection of language model training data. In Hajič, J., Carberry, S., Clark, S., and Nivre, J. (eds.), *Proceedings of the ACL 2010 Conference Short Papers*, pp. 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-2041/.

Raunak, V., Menezes, A., and Junczys-Dowmunt, M. The curious case of hallucinations in neural machine translation. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1172–1183, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.92. URL https://aclanthology.org/2021.naacl-main.92/.

Sennrich, R. and Zhang, B. Revisiting low-resource neural machine translation: A case study. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 211–221, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1021. URL https://aclanthology.org/P19-1021/.

Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data. In Erk, K. and Smith, N. A. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi,

E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

A. Related Work

A.1. Low-resource MT and multilingual transfer

Low-resource MT has benefited substantially from multilingual transfer learning, where a single model is trained jointly on many language pairs and then adapted to a target direction with limited bitext (Sennrich & Zhang, 2019; Aharoni et al., 2019). Multilingual encoder–decoder models share parameters across languages, enabling cross-lingual transfer and providing strong baselines even in scarce-data regimes (Liu et al., 2020; Xue et al., 2021; Team et al., 2022). Nevertheless, fine-tuning performance remains tightly coupled to the quantity and quality of available parallel data, and multilingual transfer alone often cannot overcome domain mismatch or sparse coverage in genuinely low-resource settings (Chu & Wang, 2018; Sennrich & Zhang, 2019).

A.2. Synthetic parallel data

Synthetic bitext is a widely used technique for overcoming parallel data scarcity, with back-translation and forward translation serving as standard approaches for leveraging monolingual corpora (Sennrich et al., 2016; Edunov et al., 2018). Self-training variants further iterate on these ideas by generating pseudo-labels and retraining the model (Hoang et al., 2018). While synthetic data can significantly improve translation quality, its effectiveness is highly sensitive to generation noise (Khayrallah & Koehn, 2018). Poor-quality synthetic pairs may contain hallucinated content or semantic drift, incorrect named entities, or degraded fluency, which can mislead training and cause performance regressions (Raunak et al., 2021; Guerreiro et al., 2023).

A.3. Data filtering and quality estimation

To control synthetic noise, prior work has explored filtering and quality estimation based on language identification, length ratio constraints, perplexity or fluency scoring, and cross-entropy difference criteria (Moore & Lewis, 2010; Junczys-Dowmunt, 2018). These heuristics are useful but often insufficient to eliminate subtle adequacy errors that remain fluent on the surface (Khayrallah & Koehn, 2018). In contrast, agreement between independent teachers provides a high-precision signal: when two strong systems converge on similar outputs for the same input, the pair is more likely to be faithful and well-formed. This perspective is consistent with findings in multilingual sentence embedding research used for mining and quality control, where semantic consistency acts as a strong reliability cue (Artetxe & Schwenk, 2019). This motivates our dual-teacher agreement framework, which selects a cleaner synthetic corpus by jointly modeling consistency and reliability.

Table 8. **Why synthetic data can fail in low-resource MT.** Unfiltered synthetic bitext often includes semantic and linguistic errors that are amplified during training when real parallel data is scarce. Dual-teacher agreement removes many low-faithfulness and low-stability pairs by favoring translations that remain consistent across independent teachers.

Synthetic failure type	Example symptom	Why it hurts MT	What agreement fixes
Hallucination	content added or omitted	Injects incorrect supervision that distorts adequacy and coverage	Discards unstable outputs where teachers diverge
Wrong named entities	mistranslated names/locations	Propagates systematic entity errors into the student model	Retains pairs with consistent entity rendering across teachers
Overly literal output	awkward word-for-word translation	Reduces naturalness and harms generalization across domains	Prefers well-formed hypotheses that converge across systems
Ungrammatical target	broken morphology or syntax	Teaches invalid target patterns and destabilizes decoding	Filters low-fluency generations that fail agreement criteria