Dynamic Meta-Metrics: Source-Sentence Conditioned Weighting for MT Evaluation

Luke Zhang[♥], **Justin Vasselli**[♣], **Aditya Khan**[♥], **York Hay Ng**[♥], **En-Shiun Annie Lee**[♥♦]
[♥]University of Toronto, Canada [♣]Nara Institute of Science and Technology, Japan
[♦]Ontario Tech University, Canada
lukelz.zhang@mail.utoronto.ca, vasselli_justinray.vk4@is.naist.jp
adityakhan@cs.toronto.edu, yorkng@cs.toronto.edu

Abstract

We propose Dynamic Meta-Metrics (DMM), a framework for machine translation evaluation that learns source-sentence conditioned combinations of existing metrics. Rather than relying on a single static ensemble or language-specific weighting, DMM adapts the metric combination based on properties of the source segment. We study hard conditioning, which fits an interpretable combiner per cluster, and an exploratory soft-conditioned extension whose weights vary continuously with source-cluster responsibilities. We evaluate DMM on the WMT Metrics Shared Task data across multiple language pairs using pairwise agreement measures at the system and segment levels. Across settings, MLP-based combinations outperform linear and Gaussian process-based ensembles, and introducing soft conditioning yields gains over linear models.

1 Introduction

Automatic evaluation metrics underpin machine translation (MT) research and deployment. An effective metric should align with human judgements, generalise across languages and domains, and remain reproducible. The WMT Metrics Shared Task provides a standardised setting for meta-evaluation and has shown that metric behaviour varies with language, domain, and input characteristics (Freitag et al., 2022; Kocmi et al., 2023; Freitag et al., 2024). In particular, no single metric consistently dominates across years and test conditions, even among strong neural and generative metrics (Juraska et al., 2023, 2024; Freitag et al., 2024).

Anugraha et al. (2024) addressed distribution shift by combining multiple metrics into static ensembles with fixed weights, which can be more robust than relying on a single metric, termed *meta-metrics* (see Appendix A). This is not to be confused with meta-metrics in the context of *evaluating* metrics (e.g. as in Thompson et al. (2024)).

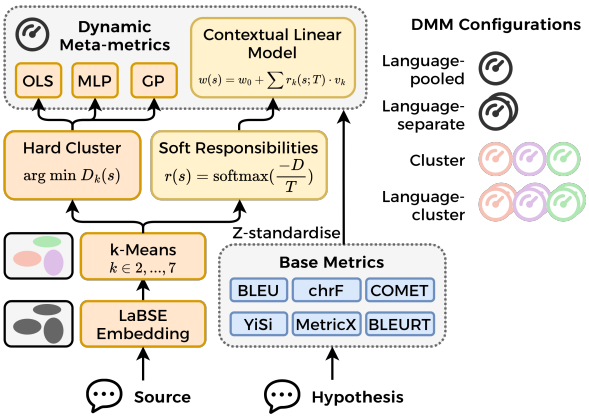

Figure 1: The DMM framework combines MT evaluation metrics conditioned on source-sentence context, with four training configurations.

However, such ensembles do not model variation within a language pair, where differences in syntax, discourse style, or domain can affect metric reliability.

To address this limitation, we introduce *Dynamic Meta-Metrics* (DMM), a framework for MT evaluation that conditions metric combination on source-sentence context rather than using a single fixed ensemble. DMM constructs a discrete context variable from sentence embeddings and clustering: each source sentence is embedded with LaBSE (Feng et al., 2022) and grouped using k-means, after which the model learns context-specific combinations of base metric scores. We call this **hard conditioning**. This preserves interpretability by yielding explicit weights for each context while allowing metric contributions to vary with the input. We instantiate DMM with three meta-metric models: (i) a linear regressor (fit with ordinary least squares – OLS) minimising MSE on segment-level human scores, (ii) a small multilayer perceptron (MLP) trained with the mean-squared error loss (MSE), and (iii) a Gaussian-process-driven Bayesian optimiser that maximises

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 903–912

Kendall correlation to human scores.

Hard conditioning assigns each source segment to a single cluster and applies the corresponding combiner. This yields an interpretable piecewise-constant rule, but it may be restrictive when a source segment exhibits properties shared by several source contexts. We therefore also define an exploratory **soft-conditioned** extension of DMM, which uses the source segment's distances to all cluster centroids to form a *soft* contextual representation, allowing the final metric weights to vary continuously with context.

Our contributions are threefold: (i) we propose DMM, a source-conditioned framework for combining MT evaluation metrics; (ii) we introduce hard and soft conditioning variants that trade off interpretability and flexibility; and (iii) we show on WMT data (Freitag et al., 2024) across four representative language pairs that MLP-based combinations outperform linear and Gaussian process ensembles, and that soft conditioning yields substantial gains over linear models.

2 Method

2.1 Problem setup and notation

Each training instance is a triple (s, t, y), where s denotes a source segment, t a system output (hypothesis) for that segment, and $y \in \mathbb{R}$ a human segment-level score. We index instances by $i \in \{1, \ldots, n\}$. For each instance i, we compute d base metric scores and form a feature vector $\mathbf{x}_i = (m_j(i))_{j=1}^{d} \in \mathbb{R}^d$, with $m_j(i)$ being the jth metric score for instance i. A meta-metric is a function $F_\theta : \mathbb{R}^d \to \mathbb{R}$ that predicts $\hat{y}_i = F_\theta(\mathbf{x}_i)$, with learnable parameters θ.

2.2 Feature standardisation

Since metric outputs differ in scale, we z-standardise each metric feature using training-set statistics. For metric j, let μ_j and σ_j be the mean and standard deviation over training instances. We standardise x_{ik}, obtaining $\tilde{x}_{ik}$ forming $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{id})$. All subsequent linear and neural models use $\tilde{\mathbf{x}}_i$. Note that standardisation does not require the metric-score distribution to be Gaussian. It is only to put the scores on a common scale.

2.3 Embedding and clustering source segments

We embed each source segment s using LaBSE (Feng et al., 2022), yielding $\mathbf{e}(s) \in \mathbb{R}^m$. The reason we use LaBSE is because DMM requires a source-only representation that is comparable across language pairs. LaBSE is designed for language-agnostic sentence embeddings and provides a fixed multilingual sentence space, making it a natural choice for clustering source segments across languages.

Following that, we fit k-means on the training-set source embeddings to obtain K centroids $\{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$. For any segment s, we define squared distances $D_k(s) = \|\mathbf{e}(s) - \mathbf{c}_k\|_2^2$ for $k = 1, ..., K$.

2.4 Soft-conditioned DMM

We now turn to the soft-conditioning extension. For a source segment s, the squared distances from its embedding to the K cluster centroids are converted into a soft responsibility vector $\mathbf{r}(s; T) = \text{softmax}(\{-D_k(s)/T\}_{k=1}^{K})$, where $T > 0$ is a temperature parameter. Smaller values of T produce sharper, near one-hot assignments, while larger values yield more diffuse mixtures over clusters. The responsibilities depend only on the source segment and are therefore shared across all candidate translations of that segment.

We then define a contextual linear meta-metric whose effective weight vector varies with source context: $\mathbf{w}_{\text{eff}}(s; T) = \mathbf{w}_0 + \sum_{k=1}^{K} r_k(s; T)\, \mathbf{v}_k$, where $\mathbf{w}_0 \in \mathbb{R}^d$ is a global baseline weight vector and $\mathbf{v}_1, \ldots, \mathbf{v}_K \in \mathbb{R}^d$ are cluster-associated deviation vectors. The prediction for instance i is $\hat{y}_i = \tilde{\mathbf{x}}_i^\top \mathbf{w}_{\text{eff}}(s_i; T) + b$ with intercept $b \in \mathbb{R}$.

This constitutes *soft conditioning* because each source segment induces a continuous weighting over clusters, so the metric-combination weights vary smoothly with that weighting. Further details such as how 1) hard conditioning appears as a special case of this model, and 2) why we do not use a neural model for soft conditioning are described in Appendix C.

2.5 Ensemble Configuration

We consider two orthogonal design choices: (i) whether training data is pooled across language pairs or separated by language, and (ii) whether clustering is used. In particular, the configurations are as follows.

- **Language-pooled:** a single model is trained on all instances across language pairs.
- **Language-separate:** one model is trained per language pair.

- **MM (meta-metrics; no conditioning):** a single model is applied within each training setup. This corresponds to a standard meta-metric.
- **DMM (dynamic meta-metrics; conditioning):** data is partitioned into K clusters based on the source segment, and a separate model is trained for each cluster.

For each configuration, we train multiple model classes: ordinary least squares (OLS), Gaussian processes (GP), and multilayer perceptrons (MLP). For DMM, we additionally include a contextual linear (CL) variant described in Section 2.4.

2.6 Evaluation measures

Following WMT24 (Freitag et al., 2024), we evaluate metrics primarily using pairwise agreement with human judgements. At the system level, we report soft pairwise accuracy (SPA) (Thompson et al., 2024), which accounts for uncertainty in both human and metric-induced rankings. At the segment level, we report group-by-item pairwise accuracy with tie calibration (acc^*_{eq}) (Deutsch et al., 2023), reflecting the metric's ability to rank alternative translations of the same source segment correctly.

3 Experiments

3.1 Experimental Setup

We use WMT Metrics Shared Task data from 2021–2024 for training and validation (Freitag et al., 2021, 2022, 2023, 2024), including English to Czech (en-cs), Chinese (en-zh), Ukrainian (en-uk), and Japanese (en-ja). We evaluate on held-out WMT25 data, citing both the WMT25 Automated Translation Evaluation shared task, which defines the metric-evaluation setting, and the WMT25 General MT shared task, which provides the underlying translation outputs and human evaluation context (Lavie et al., 2025; Kocmi et al., 2025). For each pair and year, we use source segments, system outputs, reference translations, and official segment-level human annotations, excluding segments without human scores. We use a fixed, deterministic split defined at the segment level within each year (approximately 80% train, 20% validation), created independently per year. We evaluate on held-out WMT25 shared task data.

We compute a representative set of reference-based metrics spanning overlap, embedding-based, and generative families, using default toolkit settings unless stated otherwise. We exclude the largest metric variants (e.g., MetricX-XXL) to support execution on typical hardware. We consider SacreBLEU, BLEU, and sentence-level BLEU variants (Papineni et al., 2002; Post, 2018), chrF++ (Popović, 2017), BLEURT-20 (Sellam et al., 2020; Pu et al., 2021), two variants of YiSi-1 (Lo, 2019) (using XMLRoberta (Conneau et al., 2020) and LaBSE (Feng et al., 2022)), COMET, and XCOMET-XL (Rei et al., 2020), MetricX-23/24 Large and XL (Juraska et al., 2023, 2024), and GEMBA-MQM (Kocmi and Federmann, 2023) using Qwen3-30B (Yang et al., 2025) as the underlying model.

For model fitting, we embed source segments with LaBSE and fit k-means on training sources. We evaluate $K \in \{2, 3, 4, 5, 6, 7\}$. Implementation details, hyperparameters, and model selection criteria are provided in Appendix D.

3.2 Results

Table 1 reports system- and segment-level meta-evaluation results across language pairs, comparing baseline metrics with the proposed DMM variants under different conditioning strategies.

Performance. Table 2 reports system- and segment-level results. For DMM, we report $K = 6$, which performed best on the validation set (having tried $K = 2, ..., 7$). The MLP combiner achieves the highest average performance across all configurations. All MLP variants outperform Gemini-2.5-Pro at the segment level (avg acc^*_{eq} 0.576 vs. 0.559), while Gemini retains a system-level SPA advantage (0.851 vs. 0.787). Cluster-conditioned DMM performs comparably to static MM for the MLP combiner: the best DMM configuration (language-separate, $K = 6$) achieves avg acc^*_{eq} 0.576 / SPA 0.782 vs. 0.576 / 0.784 for language-separate MM, suggesting that MLP's nonlinear hidden layers already capture source-context-dependent behaviour that clustering makes explicit for simpler models. The GP combiner shows greater sensitivity to conditioning but lower absolute performance, and OLS generalises poorly across all settings. The soft-conditioned contextual linear model (CL) underperforms the hard-conditioned variants (avg SPA 0.670 for language-pooled, $K = 6$), which may reflect overfitting in the expanded parameter space.

Cluster interpretability. We profile the $K = 6$ clusters to examine what source-level structure DMM exploits. Clusters separate primarily by sentence length and domain rather than language pair

System	EN-CS		EN-ZH		EN-JA		EN-UK		Avg	
	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA
Gemini-2.5-Pro	0.613	0.890	0.530	**0.851**	0.559	**0.825**	0.517	**0.819**	0.559	**0.851**
wmt22-comet-da	0.595	0.685	0.544	0.626	0.536	0.526	**0.572**	0.696	0.562	0.633
gemba	0.369	0.814	0.353	0.793	0.341	0.725	0.346	0.741	0.352	0.768
LANGUAGE-POOLED MM										
GP	0.607	0.883	0.548	0.666	0.570	0.757	0.543	0.718	0.567	0.756
OLS	0.486	0.380	0.473	0.276	0.494	0.465	0.472	0.372	0.481	0.373
MLP	0.612	0.898	**0.568**	0.818	0.566	0.741	0.556	0.692	**0.576**	0.787
LANGUAGE-SEPARATE MM										
GP	0.578	0.728	0.554	0.608	0.559	0.766	0.548	0.752	0.560	0.714
OLS	0.494	0.390	0.510	0.467	0.534	0.640	0.488	0.440	0.507	0.485
MLP	**0.614**	**0.902**	0.567	0.782	0.567	0.739	0.557	0.715	**0.576**	0.784
LANGUAGE-POOLED DMM (CLUSTERS) ($k = 6$)										
GP	0.578	0.781	0.553	0.633	0.566	0.728	0.555	0.664	0.563	0.702
OLS	0.470	0.345	0.463	0.190	0.479	0.441	0.464	0.364	0.469	0.335
MLP	0.610	0.887	0.567	0.815	0.562	0.749	0.554	0.679	0.573	0.782
CL	0.591	0.709	0.542	0.599	**0.573**	0.735	0.547	0.639	0.563	0.670
LANGUAGE-SEPARATE DMM (CLUSTERS) ($k = 6$)										
GP	0.570	0.702	0.531	0.673	0.551	0.770	0.551	0.748	0.551	0.723
OLS	0.464	0.334	0.449	0.175	0.494	0.454	0.467	0.315	0.469	0.320
MLP	0.613	0.896	0.567	0.807	0.568	0.743	0.555	0.682	**0.576**	0.782
CL	0.585	0.700	0.537	0.574	0.571	0.716	0.543	0.634	0.559	0.656

Table 1: Segment- and system-level meta-evaluation results. We report group-by-item pairwise accuracy with tie calibration (acc^*_{eq}) and soft pairwise accuracy (SPA), comparing our MM and DMM variants against Gemini-2.5-Pro-as-a-Judge, the top system at WMT25, as well as wmt22-comet-da and gemba, the top submetrics used by average acc^*_{eq} and SPA respectively. Higher is better.

Cluster	$\|S\|$	Med. Tok	Domains	GP Top	MLP Top
0	1,393	17	Minutes, e-comm., social	MetricX-24-L, XCOMET	spBLEU, chrF
1	871	11	Social, news, e-comm.	BLEU, MetricX-24-L,MetricX-24-XL	MetricX-24-L, MetricX-24-XL
2	998	25	News (NYT, BBC)	MetricX-24-L	MetricX-24-L
3	789	24	News, literary	MetricX-24-L,MetricX-24-XL	MetricX-24-L
4	1,482	7	E-comm., social, minutes	BLEU	BLEU
5	963	66	Literary (cross-lingual)	COMET	COMET

Table 2: $|S|$: segments; Med. Tok: median token count; If top and second metric are within 0.01 then both are listed.

(Table 1). For example, Cluster 5 contains longer literary segments (median 66 tokens), Clusters 1 and 4 contain short segments (7–11 tokens) from e-commerce and social media, and Clusters 2–3 contain medium-length news text (24–25 tokens).

We next examine per-cluster metric preferences using GP weights and MLP feature attributions. In semantically coherent clusters, both models agree on the dominant metric: BLEU for short segments and MetricX-24-Large for news, while longer literary segments favor neural metrics such as COMET. Agreement is weaker in mixed or transitional clusters, where surface and neural metrics provide comparable signal. Overall, these patterns suggest that clustering captures meaningful variation in metric reliability, with metric preferences aligning with known strengths of lexical and neural metrics.

Effect of k. Figure 2 shows the effect of the number of clusters k on performance. Across settings, performance is largely insensitive to k. The MLP model remains stable across all values of k, suggesting it captures source-dependent structure without explicit clustering. In contrast, GP and OLS models are more sensitive to k, with performance degrading under finer-grained clustering. Notably, the contextual linear (CL) model benefits the most from clustering, while remaining stable across $k \in [2, 7]$.

Cross-lingual transfer. Language-pooled configurations perform comparably to language-separate ones across all combiners (Table 2: MLP avg acc^*_{eq} 0.576 / SPA 0.787 pooled vs. 0.576 / 0.784 separate for static MM; similar patterns hold under clustering). This provides evidence that source-sentence semantics, captured by LaBSE embeddings, are more predictive of metric reliability than language

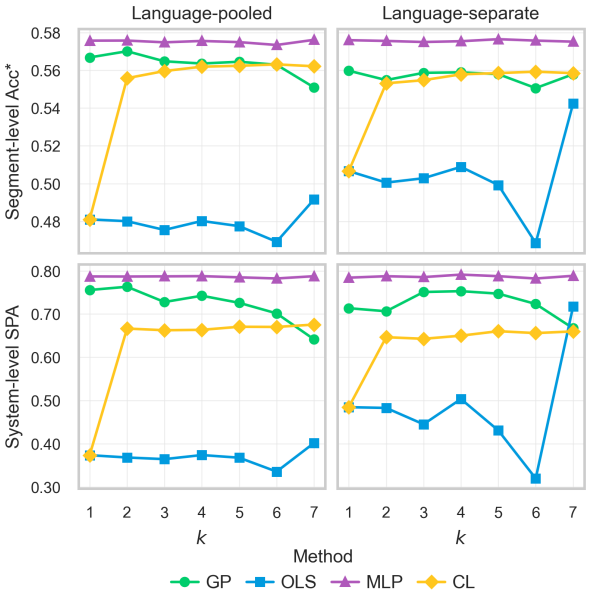

Figure 2: Effect of the number of clusters k on meta-metric performance. Top panels show Acc*, bottom panels show SPA across language pairs.

identity, and that combination strategies learned on pooled multilingual data transfer across language pairs.

Unseen language pairs. Table 3 reports results on language pairs not observed during training. Overall, the MLP-based meta-metric remains the strongest performer, showing stable generalisation across languages. In contrast, DMM with hard clustering does not provide consistent gains over standard meta-metrics in this setting. As in the seen-language experiments, Gaussian process and linear models degrade under hard clustering. Notably, the contextual linear (CL) model substantially improves over static linear regression, indicating that soft conditioning provides a more robust mechanism for adapting metric weights under distribution shift.

4 Conclusion

We introduced Dynamic Meta-Metrics (DMM), a framework for combining MT evaluation metrics based on the properties of the source segment. Across experiments, model choice and conditioning play a larger role than clustering itself. MLP-based combinations consistently outperform linear and GP-based ensembles and remain stable across seen and unseen language pairs. In contrast, hard clustering provides limited benefit and can degrade performance, particularly under distribution shift, while soft conditioning substantially improves linear models. These results suggest that source-dependent weighting is beneficial, but that flexible conditioning mechanisms are more robust than discrete clustering for meta-metric design.

Limitations

Source-only conditioning. Both hard and soft conditioning derive their context from the source segment. This is to make the routing system-invariant. However, this still does not directly capture any other phenomena, such as hypothesis-specific difficulty.

Metric set constraints. We exclude the largest metric variants to support execution on typical hardware. This may limit direct comparison with the strongest available single-metric baselines.

Metric-on-metric overfitting. When optimising for MSE (Linear Regressors) or correlation (Gaussian Processes), the combiner only sees other metrics' scores instead of human error patterns, so it can chase artefacts of those metrics (e.g., overlap bias, domain-specific scaling). For example, if a base metric has a quirk (e.g., it over-rewards literal overlap on speech transcripts), the weight search can "learn" that quirk as a shortcut to higher validation scores without actually getting closer to human judgement.

Language coverage. We train on only four language pairs, chosen to cover both high-resource and lower-resource directions. While we also evaluate on unseen language pairs, broader multilingual training would be needed to determine how well the framework scales across a wider range of languages.

Heterogeneous human evaluation protocols. Our training data pools WMT years that differ in annotation protocol and score scale. We mitigate this by normalising human targets within year–language-pair annotation cells and by reporting pairwise meta-evaluation metrics on held-out WMT25 data. Nevertheless, protocol differences may still affect the learned combiner through changes in annotation guidelines, domain composition, and rater behaviour.

Use of Generative AI

AI was only used to assist in proofreading and revising manuscripts, and to provide code completions for experiments.

References

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. 2024. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, 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, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 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*, pages 1–42, Singapore. Association for Computational Linguistics.

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

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation

evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Related Work

A.1 Single metric evaluation

Reference-based overlap metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015, 2017) are commonly used in WMT Shared Tasks due to their efficiency and historical usage. However, their emphasis on lexical overlap means that legitimate semantic or syntactic variations can be under-rewarded, leading to poorer correlation with human evaluation (Freitag et al., 2022).

Learned metrics, including BLEURT (Sellam et al., 2020; Pu et al., 2021) and COMET (Rei et al., 2020), improve correlation with human judgments by using pretrained encoders and supervised calibration to human ratings. YiSi-1 (Lo, 2019) provides an embedding-based similarity framework that can support both lower and higher-resource settings. More recently, neural generative models such as MetricX-23/24 (Kocmi et al., 2023; Freitag et al., 2022) became state-of-the-art in WMT22/23, showing stronger and more consistent results across many domains (though not all). However, no single metric dominates all domains. Freitag et al. (2024) showed that LLM-era MT outputs expose brittleness in some metrics and underscored the value of robust calibration to MQM/ESA.

A.2 Static ensembles via meta metrics

MetaMetrics-MT (MM) (Anugraha et al., 2024) learns static weights, typically per language pair, and optimises correlation with human scores, often outperforming individual metrics. However, weights are typically trained per-language and optimised against a single objective, leaving room to adapt within-language variation such as domain or

segment properties. DMM addresses this limitation by conditioning on source-derived contexts.

B Data processing and joins

Sources. We use the WMT Metrics Shared Task test sets from 2021–2024 (Freitag et al., 2024). We focus on English to Czech (en–cs) and English to Chinese (en–zh), English to Japanese (en-ja), and English to Ukrainian (en-uk) to test performance variation on languages with different resource levels and topologies. For example, Czech is lower-resource, synthetic alphabetic while Chinese is higher resource, analytic, logographic language. For each language pair, we use the source sentences, system outputs and reference translations as input, and compare against the official segment-level human annotations, using the year-specific fields: 2021 `wmt-raw:seg`, 2022 `wmt-appraise:seg`, 2023 `da-sqm:seg`, 2024 `esa:seg`. Segments without human annotation are excluded.

Units. A segment is one source sentence (WMT `seg-id`). A system is a MT run that produces one hypothesis per segment (`system-name`). Within a given (year, language pair) we identify an item by the pair (`system-name`, `seg-id`). When joining across years, we suffix the year to the identifier to keep years disjoint.

C Further details on soft-conditioned DMM

We briefly recall that the model form over the z-standardised base-metric vector for instance i is given by

$$\hat{y}_i = \tilde{\mathbf{x}}_i^\top \Big(\mathbf{w}_0 + \sum_{k=1}^{K} r_k(s_i; T) \mathbf{v}_k \Big) + b.$$

Equivalently, this may be viewed as a linear model over an expanded design matrix consisting of the original metric features together with the interaction features

$$\tilde{\mathbf{x}}_i \, r_1(s_i; T), \ldots, \tilde{\mathbf{x}}_i \, r_K(s_i; T).$$

In this subsection, we review some theoretical and implementation details.

Why a linear model? We use a *linear* contextual model, rather than a more expressive nonlinear architecture, to isolate the effect of replacing discrete source conditioning with continuous source conditioning while preserving interpretability. Under this parameterisation, one may directly inspect the global baseline weights $\mathbf{w}_0$ and the context-specific perturbations $\mathbf{v}_k$, making it possible to ask how the preferred mixture of metrics changes across source contexts.

Hard conditioning is a limiting case. Hard conditioning is a limiting case of this model. As $T \to 0^+$, the responsibility vector $\mathbf{r}(s; T)$ approaches a one-hot assignment concentrated on the nearest centroid. In that regime, $\mathbf{w}_{\text{eff}}(s; T) \to \mathbf{w}_0 + \mathbf{v}_{c(s)}$, $c(s) = \arg\min_k D_k(s)$, so prediction reduces to applying a single context-specific weight vector. Thus, the soft-conditioned model generalises the *selection mechanism* of hard conditioning, while allowing intermediate cases in which a source segment partially belongs to several source contexts.

D Implementation details and hyperparameters

This appendix specifies the implementation choices and hyperparameters required to reproduce the clustering, responsibility construction, and model training procedures described in Section 2.

D.1 Source embeddings

We embed each source segment using LaBSE sentence embeddings (Feng et al., 2022). We encode source sentences in batches of 128 and apply embedding normalisation to unit length prior to clustering and distance computation. We treat the encoder as fixed and do not fine-tune it.

D.2 Clustering via k-means on training sources

For each choice of number of clusters K, we fit a k-means model to source embeddings from the training split only. We treat a source sentence as a unique string and deduplicate before fitting. We use Euclidean distance in embedding space. Since embeddings are normalised, Euclidean distance is proportional to angular distance, which improves stability for clustering and subsequent distance-based computations. For $K \in \{2, 3, 4, 5, 6, 7\}$, we run k-means with the standard initialisation in `sklearn` (Pedregosa et al., 2011).

D.3 Hard-conditioned models

Hard-conditioned DMM realizes the four ensemble configurations described in Section 2.5 by training separate models for each condition, without parameter sharing.

- **Linear (OLS).** $M(x) = \mathbf{w}^\top \mathbf{s}(x) + b$. Inputs are z-standardized per metric and a linear regression model is fit using ordinary least squares to minimize mean squared error (MSE) on the human gold scores $h(x)$.

- **Neural (MLP).** Two hidden layers (64 and 32 units) with ReLU activations and dropout $p = 0.2$. Trained with MSE using Adam (learning rate 10^{-3}), batch size 32, for 100 epochs. Inputs are z-standardized.

- **Gaussian Process (GP).** A linear combiner $M(x) = \mathbf{w}^\top \mathbf{s}(x)$ with $\mathbf{w} \in [0, 1]^d$. We learn $\mathbf{w}$ via Bayesian optimization with a Gaussian Process to *maximize* Kendall's tau correlation to human scores. We use 5 random initial evaluations and 100 optimization iterations. Degenerate zero solutions are avoided with a small uniform initialization.

D.4 Soft-conditioned DMM

Optimisation. We fit the model by ridge regression with regularisation parameter $\alpha = 1.0$. Only the base-metric features are standardised; the centroid-distance features are left on their original scale and are used solely to construct the responsibilities. The intercept is fit jointly with the regression coefficients. Our default implementation uses `sklearn`'s ridge regression solver.

Temperature selection. The temperature T governs the sharpness of the responsibility distribution. We select T on the validation split from the grid

$$T \in \{0.1,\ 0.25,\ 0.5,\ 1.0,\ 2.0\}.$$

For each candidate value of T, we:

1. compute the corresponding responsibility vectors from the stored centroid distances,

2. fit the ridge-regression contextual model on the training split,

3. evaluate the fitted model on the validation split.

We then retain the temperature yielding the best validation performance according to a specified selection criterion. In our experiments, we use validation Pearson correlation as the default selection criterion.

E Full Table References

System	CS-DE		CS-UK		EN-AR		EN-ET		EN-IS		EN-IT	
	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA
META-METRICS												
GP	0.593	**0.884**	0.551	0.779	0.308	0.420	0.645	0.861	0.751	0.903	0.574	**0.787**
OLS	0.449	0.258	0.461	0.372	0.174	0.150	0.497	0.511	0.548	0.538	0.564	0.746
MLP	**0.602**	0.879	**0.555**	0.751	0.436	0.705	**0.659**	0.867	0.770	0.902	**0.579**	0.762
DMM $k = 6$												
GP	0.566	0.835	0.552	0.787	0.231	0.300	0.626	0.793	0.729	0.885	0.571	0.749
OLS	0.434	0.219	0.461	0.373	0.158	0.141	0.483	0.454	0.512	0.478	0.569	0.742
MLP	**0.602**	0.878	0.551	0.747	**0.452**	**0.745**	0.658	**0.874**	**0.771**	**0.904**	0.578	0.751
CL	0.577	0.812	0.547	**0.807**	0.270	0.332	0.631	0.768	0.726	0.889	0.574	0.733

System	EN-KO		EN-MAS		EN-RU		EN-SR		JA-ZH		Avg	
	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA	acc^*_{eq}	SPA
META-METRICS												
GP	0.340	0.872	0.295	0.583	**0.574**	**0.705**	**0.713**	**0.906**	0.456	0.730	0.536	0.765
OLS	0.262	0.345	0.321	**0.623**	0.496	0.456	0.462	0.416	0.323	0.241	0.422	0.430
MLP	**0.353**	**0.886**	0.307	0.510	0.568	0.682	0.712	0.900	**0.490**	**0.867**	0.558	**0.794**
DMM $k = 6$												
GP	0.332	0.785	0.293	0.531	0.563	0.680	0.648	0.815	0.456	0.733	0.513	0.713
OLS	0.252	0.265	0.309	0.600	0.491	0.450	0.375	0.251	0.322	0.239	0.402	0.381
MLP	0.351	0.883	0.311	0.502	0.564	0.663	0.711	0.898	**0.490**	0.863	**0.559**	**0.794**
CL	0.337	0.818	**0.346**	0.617	0.561	0.619	0.690	0.863	0.460	0.740	0.528	0.723

Table 3: System- and segment-level meta-evaluation results on unseen language pairs using soft pairwise accuracy (SPA) and group-by-item pairwise accuracy with tie calibration (acc^*_{eq}). All models are evaluated in a language-pooled setting. Higher values indicate better agreement with human judgements.

K	Cl.	$\lvert S \rvert$	Med. Tok	Domains	GP Top	MLP Top
—	*Global*	6,496	—	All	MetricX-24-L, -XL, YiSi-R	COMET
2	0	3,686	12	Social, e-comm., minutes	MetricX-24-L, XCOMET	XCOMET, chrF
	1	2,810	30	Literary, news	MetricX-24-L, -XL, YiSi-R	COMET
4	0	2,253	9	E-comm., social, news	BLEU, MetricX-24-L	MetricX-24-L
	1	951	23	News, literary	MetricX-24-L	MetricX-24-L
	2	1,799	17	Social, minutes, e-comm.	XCOMET, COMET	spBLEU
	3	1,493	42	Literary, news	COMET	COMET
6	0	1,393	17	Minutes, e-comm., social	MetricX-24-L, XCOMET	spBLEU
	1	871	11	Social, news, e-comm.	BLEU, MetricX-24-L, -XL	MetricX-24-L, -XL
	2	998	25	News (NYT, BBC)	MetricX-24-L	MetricX-24-L
	3	789	24	News, literary	MetricX-24-L	MetricX-24-L
	4	1,482	7	E-comm., social, minutes	BLEU	BLEU, spBLEU
	5	963	66	Literary	COMET	COMET
7	0	714	24	UK/intl. news, literary	MetricX-24-L, -XL	chrF, BLEURT
	1	1,072	7	Social, minutes, news	MetricX-24-L	MetricX-24-L
	2	847	15	News, e-comm.	MetricX-24-L, -XL, YiSi-R	MetricX-24-L
	3	1,047	19	Social, minutes	MetricX-24-XL, COMET	XCOMET, COMET
	4	773	19	E-comm., minutes	MetricX-24-L, XCOMET, COMET	XCOMET, spBLEU, MetricX-24-L
	5	792	9	Social, e-comm., minutes	MetricX-24-L	MetricX-24-XL
	6	1,251	45	Literary, news	COMET	COMET

Table 4: Cluster statistics and top-ranked metrics across $K \in \{2, 4, 6, 7\}$. $|S|$: unique source segments; Med. Tok: median token count. The Global row shows unclustered baseline weights. GP global weights are identical across all K; MLP global weights vary slightly across runs (COMET and spBLEU alternate as top-ranked within 0.04), so we report the modal ranking. Metrics are listed when within 0.01 of the top weight in each model.