Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39098

When Language Rewrites History: Quantifying and Mitigating Cross-Lingual Inconsistency in LLM-Based Historical Question Answering via Quantitative Analysis and LLM Fine-Tuning

Katelyn Chen

Carmel High School, Carmel, USA
kyc.10934@gmail.com

Abstract. Open-ended historical questions can elicit different responses from a large language model when only the input language changes. We tested this with 200 questions about the Great Leap Forward and related modern Chinese history, written in English, Chinese, and Japanese. ChatGPT and Gemini generated 1,200 answers. We compared each answer with a reference in the same language using Emotion Risk (ER) and Embedding Distance (ED), combined these measures in a Final Score, and measured score variation across languages. Both models had their highest mean Final Scores for Japanese questions: 39.04 for ChatGPT and 39.68 for Gemini. We then split the questions by ID and fine-tuned Qwen2.5-1.5B-Instruct with supervised fine-tuning (SFT) and Direct Preference Optimization (DPO). On 90 held-out language-level samples, mean Final Score fell from 34.55 in the original model to 34.36 after SFT and 31.88 after DPO. SFT gave the lowest mean Cross-Language Disparity (5.12, compared with 6.90 originally and 5.91 after DPO). These results show measurable language-conditioned differences in this historical question set and a reduction in the proposed response-risk score after training. The automatic metrics do not establish factual accuracy or historical bias on their own.

Keywords: large language models, multilingual historical question answering, cross-language consistency, Emotion Risk, Embedding Distance, supervised fine-tuning, Direct Preference Optimization

1. Introduction

Ask the same historical question in English, Chinese, and Japanese, and a model may change more than its wording. Differences in multilingual answers have already been documented in health-related question answering [1]. Historical questions add a complication: the model must write an explanation, not choose from a short list. It may name the same event in all three languages while changing what it emphasizes, omits, or treats as a cause. This study compares matched questions so that those shifts can be examined directly.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39098

Historical interpretation depends on evidence, sequence, and causal explanation. Research comparing human and model interpretations has identified cultural biases in historical reading [2]. For our questions, finding the right person or date is therefore only part of the task. We also look at whether an answer changes the account of an event. A multiple-choice response would conceal much of that variation because it does not require the model to explain its reasoning in prose.

Some variation may arise before a model produces its final wording. Prior work has examined whether multilingual models rely on English-centered internal processing [3]. Our experiment does not test internal representations. It tests the answers a reader would actually receive: each Question ID has three semantically matched prompts, and we compare the responses within that ID.

Differences in historical narratives matter because AI-generated accounts can influence readers' opinions [4]. A change in causal attribution, an official's role, or the description of a famine could affect a reader even when the prose is fluent. We use questions on the Great Leap Forward and related modern Chinese history to keep the historical topic fixed while the question language varies.

Work on cross-lingual factual knowledge asks whether equivalent prompts elicit consistent information [5]. At the same time, normal translation choices should not be counted as errors. We therefore prepare corresponding reference answers in the three languages, score each generated answer against a reference in its own language, and then compare scores across the matched questions.

Translation-based evaluation offers one way to test multilingual consistency [6]. Here, each answer is generated directly from its English, Chinese, or Japanese question. ChatGPT and Gemini each answer all 200 Question IDs, giving 600 responses per model. Those responses establish the initial cross-language comparison; the same question bank later supports a separate Qwen training and test experiment.

Language-specific latent processes have also been linked to weaker cross-lingual performance [7]. We make no claim about the internal cause of any difference observed here. Instead, we record two properties of the output: the strength of emotional expression and its semantic distance from a language-matched reference. Reporting them separately helps show whether a higher composite score comes from tone, content, or both.

Frameworks for evaluating historical text also consider cultural misconceptions [8]. We therefore treat our automatic scores as diagnostic measures, not judgments of historical truth. For each Question ID, we compare the three language versions and calculate their score dispersion. We then ask whether supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) change Qwen's scores on 30 Question IDs held out from training. The study contributes a focused trilingual dataset, paired response comparisons, and an evaluation of these training stages.

2. Methodological framework

2.1. Construction of a multilingual historical question-answering dataset

The dataset contains 200 Question IDs about the Great Leap Forward and related topics in modern Chinese history. Each ID has an English, Chinese, and Japanese question with a corresponding reference answer, giving 600 language-level question-answer samples. The ID links the three versions throughout response collection, paired comparisons, and data partitioning. Keeping them together is essential: splitting individual language rows could put one version of a test question into training.

Reference Answers were compiled manually, mainly from Wikipedia and related historical resource pages. Each answer records core details such as people, dates, background, and causal

DOI: 10.54254/2755-2721/2026.39098

relationships, while the three language versions aim to preserve the same central information. These answers provide a common comparator for training and scoring; they are not presented as the only defensible interpretation of the historical events.

Multilingual benchmarks compare model behavior across languages and tasks [9]. We submitted the 600 questions to ChatGPT and Gemini separately. Each system returned 600 answers, for 1,200 commercial-model responses in total. A response record links the Question ID, language, question, reference answer, generated text, and later risk scores. These responses were not used for Qwen supervised fine-tuning; selected commercial-model responses were subsequently considered when building DPO preference pairs.

Multilingual evaluation must account for linguistic and cultural variation [10]. We split the data by Question ID rather than by individual response: 140 IDs for training, 30 for validation, and 30 for testing. These correspond to 420, 90, and 90 language-level samples. All three versions of a test question remain in the test set, so no translation of that question is included in training or validation.

Before the split, we ranked Question IDs by their mean risk across the two commercial models and three languages. We then distributed the IDs across low-, medium-, and high-risk groups. The validation and test sets each contain 10 IDs from each group, allowing stage comparisons on a test set that spans the observed risk range.

2.2. Multidimensional response-risk framework

We score two distinct properties of each answer. Emotion Risk (ER) reflects the strength of emotional tone; Embedding Distance (ED) reflects how far the answer is from a reference in the same language. A single metric cannot capture every aspect of cultural alignment or bias [11]. For example, a response may be emotionally restrained but omit a central fact, or use emphatic language while remaining close to the reference. We report ER and ED separately before combining them in the Final Score.

2.2.1. Emotion Risk

The multilingual sentiment classifier cardiffnlp/twitter-xlm-roberta-base-sentiment returns probabilities for negative, neutral, and positive sentiment [12]. We first take the positive probability minus the negative probability to obtain a signed sentiment tendency. The next step measures its magnitude.

Our risk measure concerns the strength of that tendency, regardless of whether the response is positive or negative. We therefore use its absolute value as Emotion Risk:

$$ER = |S_{\text{emotion}}| = |P_{\text{positive}} - P_{\text{negative}}| \tag{1}$$

Emotion Risk is a property of the classifier output. A high value signals stronger measured sentiment; it does not identify a political position, a factual error, or bias by itself. We read it together with the semantic measure rather than treating emotional language as proof of historical inaccuracy.

2.2.2. Embedding Distance

BGE-M3 maps a generated answer and its reference answer in the same language to embedding vectors [13]. We compare those vectors with cosine similarity. This gives a way to compare overall meaning when the two answers use different wording:

DOI: 10.54254/2755-2721/2026.39098

$$Sim = \cos\left(e_{AI}, e_{Ref}\right) = \frac{e_{AI} \cdot e_{Ref}}{|e_{AI}||e_{Ref}|} \tag{2}$$

To express the result as deviation, we subtract cosine similarity from one. The resulting Embedding Distance increases as the two representations move further apart:

$$ED = 1 - Sim \tag{3}$$

Embedding Distance is a broad measure of semantic proximity, not a fact checker. An answer can stay close to its reference in embedding space while omitting one important date or adding a mistaken detail. We use it to compare many responses consistently and inspect representative answers at the text level when interpreting the scores.

2.2.3. Composite risk metric: Final Score

For ranking responses and comparing models or training stages, we combine the two measurements in a Final Score. The equation assigns 40% of the weight to Emotion Risk and 60% to Embedding Distance:

$$Final\ Score = 40 \times Emotion\ Risk + 60 \times Embedding\ Distance \tag{4}$$

Lower Final Scores mean lower measured emotional intensity and semantic deviation under this weighting. The score is a relative comparison within this experiment; it is not a direct probability of factual error.

The 40:60 weighting is an empirical design choice, not the result of a parameter search. We give semantic deviation the larger share because the task asks for historically informative answers, but the chosen weights are not uniquely correct. For that reason, ER and ED are also shown separately wherever the combined score is interpreted.

2.2.4. Cross-language consistency and sample-level changes

A survey of multilingual language models identifies alignment and bias as continuing concerns [14]. Pairwise comparisons show how two language versions differ, but this study also needs one value for the spread across all three. For each Question ID, we first calculate the mean of its English, Chinese, and Japanese Final Scores:

$$\overline{FS}_i^{(m)} = \frac{FS_{i,EN}^{(m)} + FS_{i,CN}^{(m)} + FS_{i,JP}^{(m)}}{3} \tag{5}$$

Cross-Language Disparity is the standard deviation of the three Final Scores for one Question ID:

$$D_i^{(m)} = \sqrt{\frac{1}{3} \sum_{l \in \{\mathrm{EN,CN,JP}\}} \left(FS_{i,l}^{(m)} - \overline{FS}_i^{(m)} \right)^2} \tag{6}$$

A lower Cross-Language Disparity means the three language versions have more similar risk scores. A higher value means the score depends more strongly on the language of the question. $D_i^{(m)}$

DOI: 10.54254/2755-2721/2026.39098

Similar scores do not guarantee good answers: three equally high Final Scores can still produce a low disparity. We therefore report Final Score and Cross-Language Disparity together. One describes measured response risk; the other describes how evenly that risk is distributed across languages.

2.3. Optimization of multilingual historical responses with Qwen

For supervised fine-tuning, each training example pairs a historical question with its Reference Answer and an instruction to respond in the question's language. Low-Rank Adaptation (LoRA) updates a small set of trainable matrices rather than every parameter of Qwen2.5-1.5B-Instruct [15]. The 420 language-level training samples cover all three languages, with the same task instruction applied throughout.

Direct Preference Optimization (DPO) follows SFT and uses chosen-rejected answer pairs to teach a relative preference [16]. Candidates for the same question are compared under ER, ED, and Final Score, then screened before training. Clear non-answers, formatting failures, and preferred candidates with substantial factual problems are removed. This screening matters because the automatic metrics used to choose pairs are also used later for evaluation; an apparent gain may partly reflect alignment to those metrics.

Original Qwen, SFT Qwen, and DPO Qwen are evaluated on the same 30 held-out Question IDs, giving 90 language-level test responses at each stage. The question text, system instruction, generation method, and scoring procedure are held fixed. We compare overall scores, language-specific changes, and score dispersion across languages; the Question-ID split prevents direct overlap between the training and test questions.

3. Experimental design and results

3.1. Experimental setup and evaluation metrics

All experiments use a common data partition, generation setup, and evaluation procedure to maintain comparability across languages, models, and training stages. The experiments have two parts. First, ChatGPT and Gemini generate responses to the complete trilingual historical question set to analyze cross-language narrative differences in distinct LLMs. Second, Qwen2.5-1.5B-Instruct undergoes progressive optimization through LoRA-based supervised fine-tuning (SFT) and Direct Preference Optimization (DPO), and the Original, SFT, and DPO stages are compared on a fixed test set [17]. Table 1 lists the training settings used for the two optimization stages.

Table 1. Training configurations for SFT and DPO

Parameter	LoRA-SFT	DPO
Base model	Qwen2.5-1.5B-Instruct	SFT-Qwen
Training samples / pairs	420 samples	60 pairs
Validation samples / pairs	90 samples	32 pairs
Epochs	4	4
Learning rate	1×10^{-4}	3×10^{-6}
Batch size	1	1

DOI: 10.54254/2755-2721/2026.39098

Table 1. (continued)

Gradient accumulation	8	4
Maximum sequence length	512	512
LoRA rank r	16	inherited
LoRA alpha (α)	32	inherited
LoRA dropout	0.05	inherited
DPO beta (β)	-	0.1
Precision	BF16	BF16
Checkpoint criterion	Minimum validation loss	Minimum validation loss

3.2. Cross-language risk analysis of ChatGPT and Gemini

The OpenAI API (GPT-5.5) and Google API (3.6 Flash), referred to below as ChatGPT and Gemini respectively, are tested in English, Chinese, and Japanese using all 200 Question IDs. Each model generates 600 responses, which are evaluated using Emotion Risk, Embedding Distance, and Final Score as defined in Section 2.2. The purpose is not to rank the absolute performance of ChatGPT and Gemini, but to test whether the same historical question exhibits reproducible language-conditioned variation across different linguistic expressions.

3.2.1. Differences in composite risk across languages

The Final Scores of both models are first compared across the three language conditions. ChatGPT has mean Final Scores of 33.65, 33.73, and 39.04 in English, Chinese, and Japanese, respectively; the corresponding Gemini means are 35.30, 34.34, and 39.68. Both models exhibit a similar overall pattern: the English and Chinese means are relatively close, while the Japanese mean is noticeably higher.

Specifically, ChatGPT's Japanese mean Final Score exceeds its English and Chinese means by approximately 5.39 and 5.31, respectively. The corresponding differences for Gemini are approximately 4.38 and 5.34. Thus, the higher composite risk under the Japanese condition is observable in two independent LLMs rather than only one model. Figure 1 shows the mean composite risk under each language condition.

For descriptive reporting, Final Score > 40 is treated as an empirical higher-risk threshold, not a universal boundary. For ChatGPT, the proportions of higher-risk responses are 42.0%, 37.5%, and 50.5% in English, Chinese, and Japanese, respectively; the corresponding Gemini proportions are 44.0%, 43.5%, and 55.5%. Japanese has the highest proportion in both models, consistent with the mean Final Score pattern.

The higher mean risk under the Japanese condition is therefore not attributable entirely to a small number of outliers; a greater share of its responses is also classified as higher risk.

DOI: 10.54254/2755-2721/2026.39098

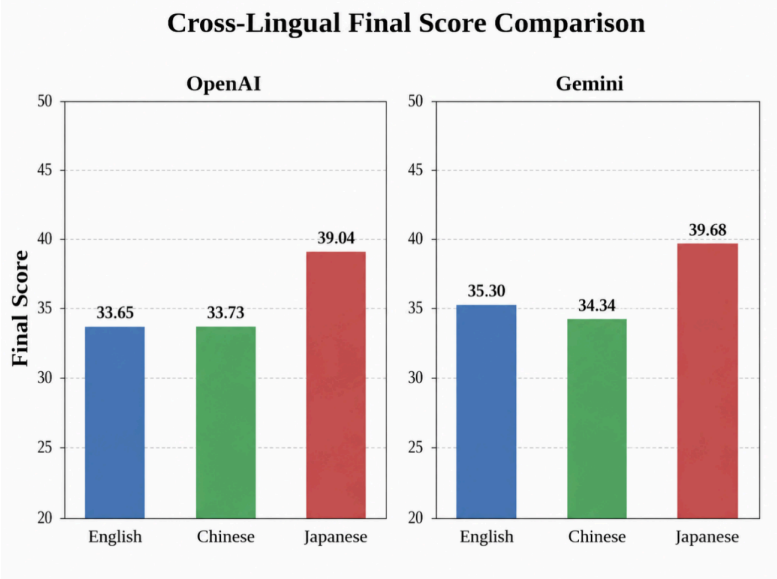


Figure 1. Mean Final Scores for ChatGPT and Gemini under different language conditions

3.2.2. Decomposition into Emotional Risk and semantic deviation

Because Final Score combines Emotion Risk and Embedding Distance, the composite score alone cannot identify which risk dimension primarily accounts for cross-language differences. Emotional risk and semantic deviation are therefore analyzed separately across the three languages.

Both models again exhibit similar trends. Mean Emotion Risk is relatively close in English and Chinese and highest in Japanese. Relative to English, Japanese Emotion Risk increases by approximately 0.065 in ChatGPT and 0.067 in Gemini; relative to Chinese, the increases are approximately 0.081 and 0.094, respectively.

The higher Final Score under the Japanese condition is therefore partly attributable to stronger emotional tendencies.

For semantic consistency, ChatGPT's mean Embedding Distances in English, Chinese, and Japanese are, respectively:

0.231,0.243,0.278

The corresponding Gemini means are:

0.252,0.254,0.280

Greater semantic deviation is again observed under the Japanese condition. To control further for differences between historical questions, paired Embedding Distance differences between languages are calculated within each Question ID.

Most notably, both Japanese–English and Japanese–Chinese show positive shifts in both models, whereas the mean semantic difference between Chinese and English is substantially smaller. In other words, expressing the same historical question in Japanese generally tends to increase the semantic distance between the generated response and its corresponding reference answer, while differences between English and Chinese are comparatively limited.

The proportions of sample pairs showing language shifts in the same direction are 66.5% for Japanese-English, 64.0% for Japanese-Chinese, and 58.0% for Chinese-English.

The Pearson correlations are weak but positive ($r = 0.245$ for Japanese-English, 0.217 for Japanese-Chinese, and 0.263 for Chinese-English). Together with the directional agreement above,

DOI: 10.54254/2755-2721/2026.39098

these results suggest limited cross-model agreement in language sensitivity, rather than strong consistency.

Overall, ChatGPT and Gemini exhibit nonidentical historical-response risk distributions across the three languages. Japanese has the highest mean Final Score in both models, together with higher Emotion Risk, Embedding Distance, and proportions of higher-risk responses. Within-question paired analysis further reveals positive Japanese–English and Japanese–Chinese semantic-distance shifts in both models, while cross-model analysis indicates some, albeit limited, question-level reproducibility of language sensitivity. These findings motivate the subsequent investigation of whether targeted supervised fine-tuning and preference optimization can mitigate multilingual historical-response risks.

3.3. Comparison of original Qwen, SFT, and DPO

To evaluate whether supervised fine-tuning and preference optimization reduce composite risk in multilingual historical question answering, the outputs of Original Qwen, SFT Qwen, and DPO Qwen are compared on a fixed test set. All three stages use the same 30 Question IDs, comprising 90 language-level samples in English, Chinese, and Japanese, with identical generation parameters and evaluation procedures. The observed metric changes therefore primarily reflect differences between training stages rather than changes in questions or scoring conditions.

3.3.1. Changes in overall composite risk

Final Score is first compared across the three training stages using all 90 test samples. The mean is 34.55 for Original Qwen, decreases to 34.36 after SFT, and decreases further to 31.88 after DPO.

Relative to Original Qwen, SFT reduces the mean Final Score by only 0.19, or approximately 0.55%. DPO reduces it by 2.67, or approximately 7.7%, relative to Original Qwen, and by a further 2.48 relative to SFT. Thus, supervised fine-tuning alone provides only limited improvement in the composite score, while subsequent preference optimization yields a more pronounced reduction in overall risk.

3.3.2. Fine-tuning effects under different language conditions

Aggregate means may conceal differences in optimization effects between languages. Performance is therefore analyzed separately for English, Chinese, and Japanese.

SFT reduces the score by 1.26 relative to Original. The DPO mean is 34.63, a slight increase of 0.11 relative to SFT, but still 1.15 below Original. Most of the improvement in English therefore occurs during SFT, with no substantial additional benefit from DPO.

SFT slightly increases the mean relative to Original, whereas DPO reduces it to 28.43, a decrease of 2.67 relative to Original and 3.25 relative to SFT.

SFT remains close to Original, whereas DPO reduces the mean Final Score to 32.58, a decrease of 4.18 relative to Original and 4.30 relative to SFT.

DPO benefits therefore vary substantially by language. Relative to Original Qwen, DPO reduces the mean Final Score by approximately 11.4% in Japanese, 8.6% in Chinese, and 3.2% in English. Japanese shows the largest improvement, followed by Chinese, while the English change is comparatively limited. Figure 2 compares the stage-wise results for each language.

DOI: 10.54254/2755-2721/2026.39098

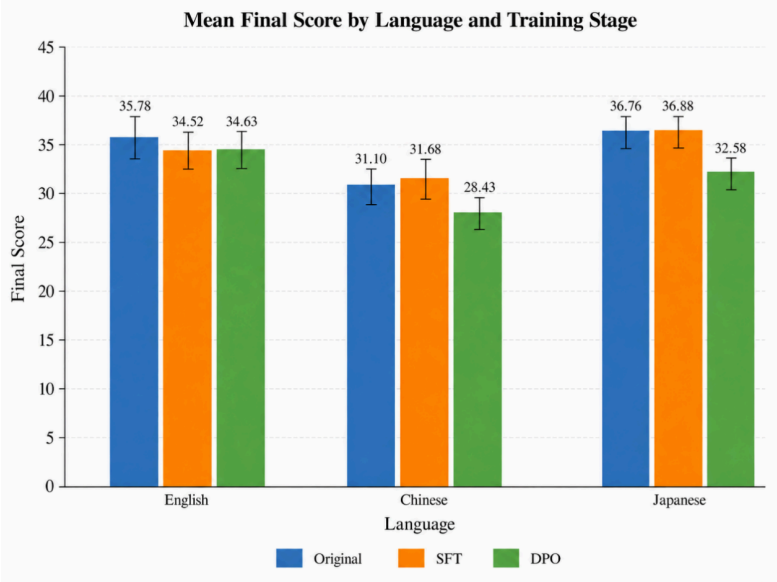


Figure 2. Mean Final Scores of Original, SFT, and DPO under different language conditions. DPO improvements are concentrated in Chinese and Japanese, with relatively limited changes in English

3.3.3. Decomposition into Emotion Risk and Embedding Distance

Because Final Score combines Emotion Risk and Embedding Distance, the composite score alone cannot explain the source of DPO's improvement. The two components are therefore analyzed separately.

SFT does not reduce mean Emotion Risk; instead, it increases slightly from 0.464 to 0.470. DPO subsequently reduces it to 0.433, the lowest value among the three stages.

SFT reduces mean Embedding Distance from 0.267 to 0.259, and DPO reduces it further to 0.243. Both training stages therefore bring model outputs closer to the corresponding Reference Answers in semantic space, with DPO achieving the lowest mean semantic deviation.

DPO's reduction in composite risk is not driven by a single metric, but by simultaneous improvements in Emotion Risk and Embedding Distance.

SFT and DPO have different effects on the risk structure. SFT primarily reduces Embedding Distance without a corresponding improvement in Emotion Risk, resulting in a limited overall Final Score change. DPO reduces both components, lowering the overall mean Final Score from 34.55 to 31.88. Language-specific analysis shows that these improvements are not uniform: DPO's principal benefits occur in Japanese and Chinese, while English improvements are relatively limited.

3.4. Sample-level changes and cross-language consistency

The preceding analysis shows lower mean risk after DPO in Final Score, Emotion Risk, and Embedding Distance. Aggregate means, however, do not indicate how broadly these improvements extend across test samples. A few large reductions can substantially affect the overall mean while other responses remain unchanged or worsen. This study therefore analyzes the generality of optimization benefits and cross-language stability through Question-ID-level trajectories, sample-level improvement proportions, and cross-language disparity.

Among the 30 Test Question IDs, 18 show a lower mean Final Score from Original to SFT, and 22 show a further decrease from SFT to DPO. Comparing Original directly with DPO, 20 questions

DOI: 10.54254/2755-2721/2026.39098

ultimately obtain a lower mean Final Score.

3.4.1. Proportions of improved, worsened, and essentially unchanged samples

Question-level mean trajectories may still conceal different changes among the three languages within a question. The 90 language-level test samples are therefore analyzed individually, and each sample change in Final Score across training is classified. Let deltaFS = FS_after - FS_before and epsilon = 0.5: deltaFS < -epsilon is improved, -epsilon <= deltaFS <= epsilon is essentially unchanged, and deltaFS > epsilon is worsened.

For the direct Original→DPO comparison:

63.3% of samples ultimately obtain lower Final Scores;

1.1% remain essentially unchanged;

35.6% have higher scores than Original.

Nearly two-thirds of language-level test samples therefore exhibit lower risk from Original to DPO, providing sample-level support for the aggregate improvement.

Nevertheless, approximately 35.6% of samples do not improve, indicating that preference optimization does not benefit every question and language condition consistently. Figure 3 shows how often individual test responses improve or worsen.

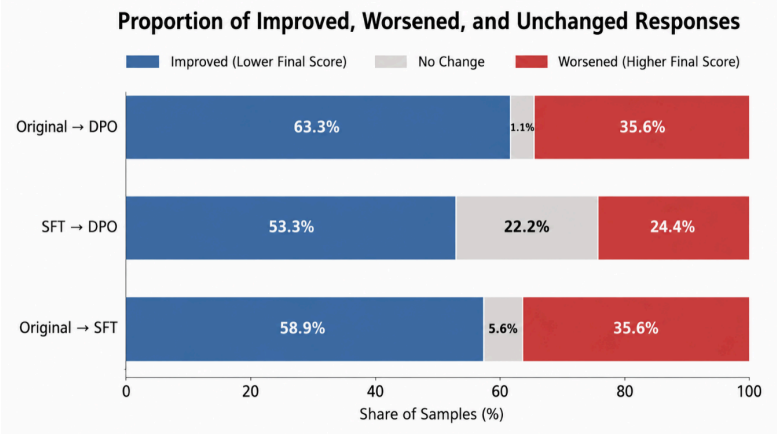


Figure 3. Sample-level Final Score changes across Original, SFT, and DPO. Lower scores indicate improvement

3.4.2. Changes in cross-language consistency

Lower overall risk does not necessarily imply more consistent performance across English, Chinese, and Japanese. This study therefore uses Cross-Language Disparity, defined in the methods section, to measure dispersion in Final Score among the three versions of each Question ID.

A key finding is therefore that SFT achieves the best cross-language consistency. Although DPO still improves stability relative to Original, it does not retain all of the consistency gains achieved by SFT.

Both question-level trajectories and sample-level proportions indicate that DPO's overall improvement is not driven solely by a few extreme samples. Of the 30 Test Question IDs, 20 have lower mean Final Scores after DPO; among the 90 language-level samples, 63.3% improve relative to Original. Nevertheless, approximately one-third of test samples show increased risk, so the benefits are not uniform. Cross-language dispersion further reveals differences between training objectives. SFT reduces mean trilingual disparity from 6.90 to 5.12, the lowest among the three

DOI: 10.54254/2755-2721/2026.39098

stages. Although DPO further reduces overall Final Score to its minimum, Cross-Language Disparity rises to 5.91. Lower composite risk and greater cross-language consistency are therefore not fully equivalent optimization objectives, and training may involve trade-offs across evaluation dimensions.

3.5. Qualitative observations

Representative Question IDs from the fixed test set were manually compared across Original Qwen, SFT Qwen, and DPO Qwen. The review included cases with lower, similar, and higher Final Scores after fine-tuning, and considered directness, agreement with the Reference Answer, omissions or added errors, and neutrality of expression.

In the reviewed examples, some reductions in Final Score corresponded to more direct responses, less semantic deviation, and more restrained expression. However, automatic-score improvements did not always align with historical answer quality. This limited qualitative observation is not a blinded human evaluation; automatic metrics are used for large-scale comparison, not as the sole basis for judging historical correctness or objectivity.

4. Discussion

Prompt framing can alter LLM stances in related settings [18]. This study examines language sensitivity and risk control in multilingual historical question answering. Observable cross-language differences are found in two independent LLMs, ChatGPT and Gemini. Overall Final Scores are relatively close in English and Chinese, while Japanese exhibits higher mean risk together with higher Emotion Risk and Embedding Distance. This trend appears not only in aggregate means but also in question-level paired comparisons and the proportions of higher-risk samples, indicating that it is not entirely attributable to a few anomalous responses. Cross-model analysis further shows some agreement in the direction of language shifts for certain questions. Although the correlations are weak, they suggest that language sensitivity in multilingual historical question answering is not confined to a single model.

Bias evaluation has multiple competing definitions and methods [19]. The evaluation framework also offers interpretive advantages. Unlike a single accuracy or sentiment measure, Emotion Risk and Embedding Distance describe emotional tendencies and semantic deviation separately,. Nevertheless, the approach has limitations. Emotion Risk reflects the strength of emotional expression and cannot be equated directly with historical bias. Embedding Distance measures semantic proximity to reference answers and cannot fully replace judgments of factual accuracy. The 40:60 weighting in Final Score is also a fixed empirical choice rather than a uniquely optimal theoretical parameter. Accordingly, all three metrics are reported separately, and representative responses are checked at the text level. DPO preference pairs are constructed using the same ER, ED, and Final Score framework later used for evaluation, so improvements may partly reflect metric alignment; splitting by Question ID prevents direct overlap between training and held-out test questions. The representative-response review is a limited qualitative check, and no blinded human evaluation was conducted. XLM-R sentiment classification and BGE-M3 embeddings may also have language-specific calibration differences, so Japanese contrasts should be interpreted relative to these evaluators rather than as absolute bias. Reference Answers were compiled primarily from Wikipedia and related historical pages; without documented expert adjudication or systematic multi-source corroboration, they serve as study-specific comparators rather than definitive historical ground truth.

DOI: 10.54254/2755-2721/2026.39098

5. Conclusion

Using 200 parallel English, Chinese, and Japanese questions on the Great Leap Forward and related modern Chinese history, this study finds higher mean response risk in Japanese for both ChatGPT and Gemini within the dataset examined. On the fixed Qwen test set, mean Final Score decreases from 34.55 (Original) to 34.36 (SFT) and 31.88 (DPO). DPO reduces both Emotion Risk and Embedding Distance, although some individual responses worsen.

Cross-Language Disparity falls from 6.90 in Original Qwen to 5.12 after SFT, then rises to 5.91 after DPO. Thus, SFT gives the greatest cross-language consistency, while DPO gives the lowest composite risk. Future optimization should address both outcomes and include direct factual checks.

References

[1] Schlicht, I. B., Zhao, Z., Sayin, B., et al. (2025). Do LLMs provide consistent answers to health-related questions across languages? arXiv:2501.14719.

[2] Celli, F., & Spathulas, G. (2025). Cultural biases of large language models and humans in historical interpretation. arXiv:2504.02572.

[3] Schut, L., Gal, Y., & Farquhar, S. (2025). Do multilingual LLMs think in English? arXiv:2502.15603.

[4] Shu, M., Karell, D., Okura, K., & Davidson, T. R. (2026). How latent and prompting biases in AI-generated historical narratives influence opinions. *PNAS Nexus, 5*(3), pgag022. https://doi.org/10.1093/pnasnexus/pgag022

[5] Qi, J., Fernández, R., & Bisazza, A. (2023). Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of EMNLP 2023* (pp. 10650–10666). Singapore: ACL. https://doi.org/10.18653/v1/2023.emnlp-main.658

[6] Gupta, A., Mehta, M., Xu, Z., & Srikumar, V. (2025). Found in translation: Measuring multilingual LLM consistency as simple as translate then evaluate. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th AACL* (pp. 3477–3496). Mumbai: AFNLP/ACL. DOI: 10.18653/v1/2025.ijcnlp-long.185.

[7] Lim, Z. W., Aji, A. F., & Cohn, T. (2025). Language-specific latent process hinders cross-lingual performance. arXiv:2505.13141.

[8] Mak, M. K., & Luo, T. (2025). A framework for evaluating cultural bias and historical misconceptions in LLMs outputs. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 5*(3), 100235. DOI: 10.1016/j.tbench.2025.100235.

[9] Ahuja, S., Aggarwal, D., Gumma, V., et al. (2024). MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of NAACL 2024* (pp. 2598–2637). Mexico City: ACL. DOI: 10.18653/v1/2024.naacl-long.143.

[10] Singh, S., Romanou, A., Fourrier, C., et al. (2025). Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 18761–18799). Vienna: ACL. DOI: 10.18653/v1/2025.acl-long.919.

[11] Alkhamissi, B., Elnokrashy, M., Alkhamissi, M., & Diab, M. (2024). Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 12404–12422). Bangkok: ACL. DOI: 10.18653/v1/2024.acl-long.671.

[12] Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 258–266). Marseille: ELRA.

[13] Chen, J., Xiao, S., Zhang, P., et al. (2024). BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv:2402.03216.

[14] Xu, Y., Hu, L., Zhao, J., et al. (2025). A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science, 19*, 1911362. DOI: 10.1007/s11704-024-40579-4.

[15] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

[16] Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems, 36*, 53728–53741. DOI: 10.52202/075280-2338.

[17] Yang, A., Yang, B., Zhang, B., et al. (2024). Qwen2.5 technical report. arXiv:2412.15115.

DOI: 10.54254/2755-2721/2026.39098

[18] Ratnakar, S., & Raghavendra, S. (2025). The chameleon nature of LLMs: Quantifying multi-turn stance instability in search-enabled language models. arXiv:2510.16712.
[19] Gao, Z., Tong, L., & Zhang, Z. (2026). Detecting and evaluating bias in large language models: Concepts, methods, and challenges. *Journal of Behavioral Data Science, 6*(1), 68–135. DOI: 10.35566/jbds/gao.