Cross-Language Evaluation of Prompt Inversion: Similarity Metrics, Decoding Strategies, and Prefix Sensitivity in Japanese and English

Yusei Kitamura [1] **Ahmad Akmal Aminuddin Mohd Kamal** [2] **Masaya Fujisawa** [3]

Abstract

Prompt inversion, a process of inferring a prompt from an observed output and regenerating text from that inferred prompt, has emerged as a method for analyzing machine-generated text, yet its empirical behavior across diverse languages, similarity metrics, decoding strategies, and model families remains insufficiently understood. We present a systematic evaluation of prompt inversion using English and Japanese text generated by large language models (LLMs) based on Alpaca-style instruction-response datasets. Our analysis spans five similarity metrics, multiple decoding configurations, varying prefix augmentation settings, and two generator architectures (T5-style and GPT-style). In our settings, English yields higher BLEU-4, ROUGE-Lsum, METEOR, and BERTScore F1 scores than Japanese, whereas Japanese demonstrates greater Sentence-BERT cosine similarity. Furthermore, the strongest observed decoding strategy varies by language in our T5 experiments: beam search performs best for English T5 reconstruction, while hybrid decoding performs best for Japanese. We observe that prefix augmentation substantially benefits English T5 reconstruction but hinders performance on Japanese. Finally, we demonstrate that reconstruction quality differs significantly between T5 and GPT architectures depending on the applied metric, suggesting that no single score can fully characterize performance. These findings highlight that prompt inversion is highly sensitive to language, metric choice, decoding strategy, and model configuration, rather than acting as a universally stable evaluation framework.

[1]Graduate School of Engineering, Tokyo University of Science, Japan [2]Department of Electronic and Computer Engineering, Ritsumeikan University, Japan [3]Department of Information and Computer Technology, Tokyo University of Science, Japan. Correspondence to: Yusei Kitamura <4625515@ed.tus.ac.jp>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

1. Introduction

Large language models (LLMs) have been widely adopted for natural language generation across diverse domains, including education, communication, and content creation. As the quality of model-generated text continues to advance, it becomes increasingly important to understand how such outputs are produced and how reliably their generation behavior can be analyzed. In particular, methods capable of characterizing the relationship between an output text and its underlying generation process are essential for evaluating the behavior of modern text generation systems. Such analysis is especially critical in authorship-sensitive contexts, such as academic essays or educational writing, where understanding textual provenance is more important than surface fluency alone.

A recent line of research approaches this problem through prompt inversion (also known as prompt reconstruction), which we examine here as a reconstruction-based analysis framework (1; 2). In prompt-inversion frameworks such as DPIC (1) and IPAD (2), a prompt is inferred from an observed output text and then used to regenerate a new text. The similarity between the original and regenerated texts is measured to assess how faithfully the pipeline recovers the underlying generation behavior. This paradigm is appealing because it does not rely solely on surface stylistic cues; rather, it investigates whether an observed text can be consistently explained via a reconstructed generation path. Despite its promise, the empirical behavior of prompt-reconstruction-based analysis remains insufficiently understood. Prior prompt-inversion studies, including DPIC (1) and IPAD (2), have predominantly focused on English data, and several crucial design factors have yet to be systematically disentangled. Specifically, it remains unclear how reconstruction quality varies across languages, how sensitive the process is to distinct similarity metrics, how decoding strategies and their hyperparameters influence regenerated outputs, whether prefix augmentation alters reconstruction behavior, and how results diverge between T5-style and GPT-style models.

In this study, we present a systematic evaluation of prompt inversion for analyzing LLM-generated text in English and Japanese. Utilizing Alpaca-style instruction-response datasets, we assess reconstruction quality across multiple similarity metrics, decoding configurations, prefix augmentation settings, and generator families. Our objective is not to propose a novel reconstruction algorithm, but rather to clarify the conditions under which prompt inversion yields stable results and to identify the design choices that most significantly impact reconstruction quality. Crucially, we do not claim that these findings generalize universally across all languages or model families. Rather, we explicitly scope this work as a controlled empirical analysis to demonstrate that prompt inversion is inherently sensitive to its setting. Our contribution is a systematic benchmark-oriented evaluation of prompt inversion across languages, metrics, decoding strategies, prefix settings, and generator families.

- We evaluate prompt inversion in both Japanese and English.
- We compare five similarity metrics and show that the observed ranking is metric-dependent.
- We analyze sensitivity to decoding strategies, including beam search, sampling, and hybrid approaches.
- We examine the influence of prefix augmentation and the choice of generator family.

2. Related Work and Limitations

2.1. Prompt Inversion and Inversion-Based Analysis

Prompt inversion aims to recover an input prompt from an observed output and to use the reconstructed prompt to analyze the generation process. Representative approaches such as DPIC (1) and IPAD (2) reconstruct candidate prompts from observed outputs and evaluate the consistency between the original and regenerated texts. Prior work has shown that this framework is useful for analyzing model-generated text and for studying the relationship between prompts and outputs. These approaches are appealing because they move beyond purely surface-level analysis and instead exploit the consistency between an output text and a plausible reconstructed generation path. However, most prior studies have focused on English datasets, and the empirical behavior of prompt inversion under different experimental conditions remains fragmented. In particular, the effects of language, decoding configuration, and generation model family have not been systematically studied within a unified evaluation setting.

2.2. Evaluation Sensitivity in Reconstruction-Based Generation Analysis

An important issue in reconstruction-based analysis is evaluation sensitivity. Reported reconstruction quality may depend not only on the reconstruction model itself, but also on the evaluation protocol used to measure similarity between the original and regenerated outputs. If different similarity metrics emphasize different aspects of textual correspondence, then the apparent quality of prompt inversion may vary substantially even under the same generation pipeline. From a benchmarking perspective, this sensitivity matters because it affects how results should be interpreted. A method that appears strong under one metric may be much less convincing under another. Therefore, understanding metric sensitivity is essential for making reconstruction-based analysis more reliable and for designing evaluation protocols that remain informative across settings.

2.3. Metrics, Decoding, Multilinguality, and Generator Effects

The quality of regenerated text is strongly influenced by decoding choices such as beam search (3), top-k sampling (4), and top-p sampling (5). Prior work on text generation (6) has shown that these strategies trade off determinism, diversity, and semantic stability. However, their role in prompt-reconstruction-based analysis remains insufficiently characterized, especially when decoding hyperparameters are varied more extensively. A related issue concerns similarity metrics. Surface-overlap metrics such as BLEU and ROUGE capture lexical or structural similarity, whereas metrics such as METEOR, BERTScore, and sentence-level cosine similarity capture increasingly semantic aspects of correspondence. Because reconstruction quality may look different depending on whether lexical or semantic similarity is emphasized, metric choice is itself an important object of study. Finally, multilingual evaluation remains challenging. Many methods and metrics have been developed primarily for English and do not always transfer cleanly to Japanese, where tokenization, flexible word order, and paraphrastic variation can affect both reconstruction difficulty and similarity measurement. In addition, the behavior of prompt inversion may also depend on the family of the original generator, such as T5-style (13) versus GPT-style models (14), and on simple formatting interventions such as prefix augmentation. Our study integrates these dimensions within a single evaluation framework.

3. Prompt Inversion Framework

3.1. Problem Formulation

We study prompt inversion as a reconstruction-quality evaluation problem. Let x denote an observed output text generated by a language model. The goal is to infer a candidate prompt $\hat{p}$ from x, regenerate a text $\hat{x}$ from $\hat{p}$, and evaluate how closely $\hat{x}$ matches x under multiple similarity measures.

Formally, let f_{inv} be a prompt inversion model and let f_{gen} be a text generation model. Given an observed output x, we first estimate a prompt

$$\hat{p} = f_{\text{inv}}(x).$$

We then produce a regenerated output

$$\hat{x} = f_{\text{gen}}(\hat{p}).$$

The central quantity of interest is not a binary detection label, but the degree to which the original output can be reconstructed through this two-stage pipeline.

3.2. Prompt Inversion and Regeneration Pipeline

Figure 1 illustrates the overall prompt inversion pipeline used in this study. Our pipeline follows the general two-stage prompt inversion formulation used in prior work such as IPAD (2), while extending the evaluation to multiple languages, metrics, decoding strategies, prefix settings, and generator families.

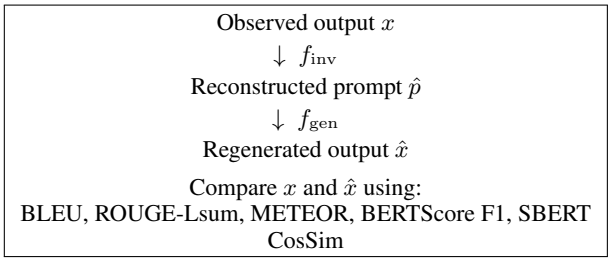

Figure 1. Overview of the prompt inversion framework. An observed output x is mapped to a reconstructed prompt $\hat{p}$, which is then used to generate a regenerated output $\hat{x}$. Reconstruction quality is evaluated by comparing x and $\hat{x}$ with multiple similarity metrics.

Prompt inversion. In the first stage, the observed output text x is given to a prompt inversion model f_{inv}, which predicts a candidate prompt $\hat{p}$. This stage attempts to recover an input condition that could plausibly have produced x.

Output regeneration. In the second stage, the reconstructed prompt $\hat{p}$ is given to a generation model f_{gen} to produce a regenerated output $\hat{x}$. This stage approximates the forward generation process implied by the reconstructed prompt.

The overall procedure can be summarized as

$$x \rightarrow \hat{p} \rightarrow \hat{x}.$$

Rather than assigning a class label to x, this framework evaluates whether the observed output can be reproduced consistently through a reconstructed generation path. This perspective enables a direct analysis of reconstruction quality under different design choices.

3.3. Similarity-Based Evaluation

After obtaining the regenerated text $\hat{x}$, we compare it with the original text x using multiple similarity metrics. Let $\mathcal{M} = \{m_1, \ldots, m_5\}$ denote the set of metrics used in the experiments. For each metric $m_i \in \mathcal{M}$, where $i = 1, 2, \ldots, 5$, we compute

$$s_{m_i}(x, \hat{x}),$$

where larger values indicate stronger similarity between the original and regenerated texts.

In this work, similarity scores are used as evaluation signals for reconstruction quality. We do not use a threshold-based decision rule or a binary classification objective. Instead, for each experimental condition, we compute similarity scores over the full dataset and compare their aggregate statistics, primarily the mean values. This design reflects our goal of characterizing how the reconstruction pipeline behaves under different settings rather than evaluating a classifier.

Because different metrics emphasize different aspects of textual correspondence, we adopt a multi-metric view of reconstruction quality. Surface-oriented metrics capture lexical and structural overlap, while embedding-based metrics better reflect semantic consistency. Comparing these metrics makes it possible to identify whether a given experimental setting improves lexical similarity, semantic similarity, or both.

3.4. Factors Analyzed in This Study

Our study examines how reconstruction quality varies across several design factors.

Similarity metrics. We compare five similarity metrics to analyze how different notions of textual correspondence affect reconstruction quality: BLEU, ROUGE-Lsum, METEOR, BERTScore F1, and Sentence-BERT cosine similarity.

Decoding configurations. Both prompt inversion and output regeneration depend on decoding decisions. We therefore examine multiple decoding configurations, including beam-search-based and sampling-based settings, and analyze how these choices affect the similarity between x and $\hat{x}$.

Prefix augmentation. We also study whether adding a prefix to the reconstruction or generation input changes the resulting reconstruction quality. This allows us to test whether simple prompt-formatting interventions improve consistency.

Generator family. Finally, we compare outputs from different model families, including T5-style and GPT-style generators, to examine whether the behavior of the reconstruction pipeline depends on the type of generator that produced the original text.

Taken together, these factors allow us to move beyond reporting a single reconstruction score and instead provide a more systematic characterization of when prompt inversion appears stable, when it becomes sensitive to design choices, and which factors most strongly affect the resulting reconstruction quality.

4. Experimental Setup

4.1. Datasets

We use Alpaca-style instruction–response datasets in both Japanese and English. For Japanese, we use `FreedomIntelligence/alpaca-gpt4-japanese`. For English, we use `vicgalle/alpaca-gpt4`. In both datasets, the instruction field is treated as the original prompt p, and the response field is treated as the observed output text x.

Our evaluation is conducted on held-out response texts that are not used for training. Specifically, we evaluate reconstruction quality on 9,994 Japanese samples and 10,401 English samples. To ensure separation between training and evaluation, we first split each processed dataset into training and test portions with a test ratio of 0.2 and a random seed of 36. The held-out test portion is used exclusively for similarity-based evaluation. The remaining training portion is further split into training and validation subsets with a validation ratio of 0.2 and a random seed of 35.

It is important to note that this study does not use human-written text. All observed outputs are model-generated responses from the Alpaca-style datasets, and our goal is to evaluate reconstruction quality under different experimental conditions rather than to perform binary human-versus-AI classification.

4.2. Models and Training Setup

We compare two generator families, namely T5-style and GPT-style models, in both Japanese and English. For the T5-style family (13), we use `sonoisa/t5-base-japanese` for Japanese and `mrm8488/t5-base-finetuned-question-generation-ap` for English. For the GPT-style family (14), we use `rinna/japanese-gpt2-medium` for Japanese and `gpt2-medium` for English.

It is important to note a discrepancy in the pretraining and prior fine-tuning conditions of the selected T5 models. In particular, the English T5 model was fine-tuned for question generation, whereas the Japanese T5 model is a standard base model. Although both models are subsequently trained under an identical prompt-inversion and regeneration setup in our experiments, this difference in their initialization may affect the English–Japanese comparison. Therefore, these cross-lingual findings should be interpreted as a comparison of the available, language-specific model pipelines rather than an isolated measurement of language effects.

To mitigate this mismatch during task adaptation, we standardize the instruction–response training procedure, optimization hyperparameters, and number of training epochs across both language settings. Nevertheless, the discrepancy in the underlying T5 checkpoints remains an inherent limitation of this experimental setup.

The same model is used for both stages of the reconstruction pipeline: prompt inversion ($x \rightarrow \hat{p}$) and output regeneration ($\hat{p} \rightarrow \hat{x}$). This allows us to compare reconstruction behavior across model families without introducing an additional model mismatch between the two stages.

All models are trained under stage-specific but otherwise shared optimization settings across languages and model families. For prompt inversion ($x \rightarrow \hat{p}$), we train for 20 epochs with a learning rate of 1×10^{-4}, a per-device training batch size of 16, a per-device evaluation batch size of 16, weight decay of 0.01, and a save limit of 3 checkpoints. For output regeneration ($\hat{p} \rightarrow \hat{x}$), we train for 15 epochs with the same learning rate, batch sizes, weight decay, and checkpoint limit. Model selection is performed using epoch-level evaluation on the validation split. For sequence length control, GPT-based models use `max_length=900`. No explicit maximum input or output length is imposed for the T5-based models in our implementation.

4.3. Decoding Configurations and Prefix Augmentation

We evaluate reconstruction quality under three groups of decoding configurations: beam search, sampling-based decoding, and hybrid decoding. The same decoding groups are applied in both stages of the pipeline ($x \rightarrow \hat{p}$ and $\hat{p} \rightarrow \hat{x}$). Detailed settings are summarized in Table 1.

We additionally examine prefix augmentation in both stages. For prompt inversion ($x \rightarrow \hat{p}$), we prepend a Japanese prefix meaning "Recover the original instruction" and the English prefix "Recover the original instruction: ". For output regeneration ($\hat{p} \rightarrow \hat{x}$), we prepend a Japanese

prefix meaning "Generate the answer" and the English prefix "Generate the answer: ".

4.4. Similarity Metrics and Evaluation Protocol

We evaluate reconstruction quality using five similarity metrics: BLEU (7), ROUGE (8), METEOR (9), BERTScore (10), and Sentence-BERT cosine similarity (11) (12). These metrics jointly cover lexical overlap, structural correspondence, and semantic similarity between the original output x and the regenerated output $\hat{x}$. For each sample, we execute the reconstruction pipeline

$$x \rightarrow \hat{p} \rightarrow \hat{x},$$

and compute a similarity score $s_{m_i}(x, \hat{x})$ for each metric m_i. We then aggregate the results at the dataset level and compare the mean similarity values across experimental conditions.

Our evaluation is similarity-based rather than threshold-based. We do not use a binary decision rule, and we do not report classification metrics such as accuracy or ROC-AUC. Instead, the objective of the experiments is to characterize how reconstruction quality varies as a function of language, similarity metric, decoding configuration, prefix augmentation, and generator family.

5. Results

Table 2 summarizes the representative T5 results with and without prefix augmentation, and Table 3 reports representative GPT results. Representative T5 configurations were selected as the highest-BERTScore setting within each decoding strategy, and the detailed decoding configurations are listed in Table 1. Overall, reconstruction quality depends strongly on language, metric, decoding strategy, prefix usage, and generator family.

5.1. Japanese vs. English in Similarity Metrics

Across the T5 results, the English setting achieves much higher BLEU-4, ROUGE-Lsum, METEOR, and BERTScore F1 than the Japanese models. However, as previously noted, the English and Japanese T5 backbones are not fully aligned; thus, this performance gap cannot be attributed solely to language. Instead, it likely reflects a complex combination of language-specific characteristics, metric sensitivities, dataset properties, and differences in the underlying model checkpoints. Under the representative beam setting with prefix, the English model reaches BLEU-4 of 0.07638 and BERTScore F1 of 0.87138, whereas the Japanese model reaches BLEU-4 of 0.00374 and BERTScore F1 of 0.66980. However, Sentence-BERT cosine similarity shows the opposite tendency, with Japanese remaining higher than English across the representative settings. This contrast indicates that lexical-overlap metrics and sentence-level embedding similarity capture different aspects of reconstruction quality.

The same metric-dependent contrast is also visible in the GPT results. English GPT remains higher in BLEU-4, ROUGE-Lsum, METEOR, and BERTScore F1, while Japanese GPT shows higher Sentence-BERT cosine similarity. Therefore, the performance discrepancy between Japanese and English is not uniform across evaluation criteria. It should neither be reduced to a single aggregate score nor interpreted as an isolated language-driven effect.

5.2. Decoding Strategy Sensitivity

For T5-based reconstruction, the preferred decoding strategy differs by language. In English, beam search gives the strongest representative results, hybrid decoding is second, and sampling is weakest. In Japanese, the strongest representative results are obtained by hybrid decoding, followed closely by sampling, while beam search is weaker. Thus, the ranking of decoding strategies is language-dependent.

The full T5 results also show finer-grained variation inside each strategy. In English beam search, increasing the beam width from 1 to 5 improves the reported scores, while the additional gain from 5 to 10 is small. In Japanese, the differences inside the beam group are also small once the beam width becomes larger than 5. Within the sampling group, English slightly favors top-p-based settings, whereas Japanese slightly favors top-k-based settings.

In contrast, the GPT-based results exhibit minimal variation across the available representative configurations. The evaluations of the Japanese GPT, presented in Table 3, require particular attention, as the performance metrics remain almost identical across beam search, sampling, and hybrid decoding strategies. However, this uniformity should not be misconstrued as evidence that the Japanese GPT model is highly robust to decoding choices. Upon further examination, the regenerated outputs were highly similar and, in some cases, effectively unchanged, regardless of the decoding configuration used. Therefore, these near-identical scores are more likely to reflect the limited output diversity within this specific setting rather than true robustness to decoding variations.

5.3. Prefix and Generator Family

Prefix augmentation produces opposite effects in English and Japanese T5 reconstruction. In English, adding prefixes substantially improves all representative metrics. For example, under the representative beam setting, BLEU-4 increases from 0.01011 to 0.07638, and BERTScore F1 increases from 0.83398 to 0.87138 when prefixes are added. In Japanese, the pattern is reversed: removing prefixes gen-

Table 1. Decoding configurations used in both stages of the prompt inversion pipeline.

Group	Variable settings	Shared settings
Beam	num_beams = {1, 5, 10}	early_stopping=True
Sampling	top-p = {0.70, 0.95} top-k = {4, 10} temperature = {0.5, 1.0}	do_sample=True, no_repeat_ngram_size=3, repetition_penalty=1.0, max_new_tokens=128, length_penalty=1.0, early_stopping=True
Hybrid	num_beams = {3, 6} top-p = 0.95 or top-k = 7 temperature = 0.7	do_sample=True, no_repeat_ngram_size=3, repetition_penalty=1.0, max_new_tokens=128, length_penalty=1.0, early_stopping=True

Table 2. Representative T5 reconstruction results with and without prefix augmentation. For each language–strategy pair, one representative configuration is shown. Best values within each language block are shown in bold.

Language	Strategy	Config	Prefix	BLEU-4	ROUGE-Lsum	METEOR	BERTScore F1	SBERT CosSim
English	Beam	beam=5	with	**0.07638**	**0.28360**	**0.25940**	**0.87138**	**0.52920**
English	Beam	beam=5	without	0.01011	0.11760	0.06133	0.83398	0.41935
English	Sampling	top-p=0.95, temp=0.5	with	0.02889	0.15008	0.18355	0.85098	0.51451
English	Sampling	top-p=0.95, temp=0.5	without	0.00807	0.10877	0.05633	0.83065	0.34953
English	Hybrid	beam=6, top-k=7	with	0.03484	0.16286	0.20090	0.85421	0.52090
English	Hybrid	beam=6, top-k=7	without	0.01035	0.11886	0.06181	0.83578	0.42529
Japanese	Beam	beam=5	with	0.00374	0.09595	0.01131	0.66980	0.63171
Japanese	Beam	beam=5	without	**0.01106**	0.10746	**0.03004**	0.69358	0.67801
Japanese	Sampling	top-k=4, temp=0.5	with	0.00370	0.10516	0.01076	0.69238	0.65351
Japanese	Sampling	top-k=4, temp=0.5	without	0.01074	0.10642	0.02922	0.69439	0.67771
Japanese	Hybrid	beam=6, top-k=7	with	0.00385	**0.10815**	0.01075	0.69385	0.65563
Japanese	Hybrid	beam=6, top-k=7	without	0.01072	0.10664	0.02905	**0.69484**	**0.67982**

erally improves the reported scores, especially for BLEU-4, METEOR, and Sentence-BERT cosine similarity. Under the representative beam setting, BLEU-4 increases from 0.00374 to 0.01106, and Sentence-BERT cosine increases from 0.63171 to 0.67801 when prefixes are removed.

Under the evaluated checkpoints and training configurations, the English T5 models achieve higher BLEU-4, ROUGE-Lsum, METEOR, and BERTScore F1 scores than the English GPT models, whereas the GPT architecture demonstrates slightly stronger Sentence-BERT cosine similarity. However, this comparison must be approached with caution, as the evaluated checkpoints differ in their pretraining and fine-tuning histories. Therefore, these results suggest that reconstruction behavior is highly sensitive to the specific generator configuration, rather than serving as a definitive performance ranking between the T5 and GPT model families. In the Japanese context, GPT models outperform the prefix-enabled T5 models on several metrics, although T5 achieves the highest BERTScore F1 in the optimal hybrid decoding setting. Ultimately, comparative assessments between generator families remain heavily dependent on the specific similarity metric prioritized.

6. Discussion

6.1. Language Effects and Model-Backbone Confounds

A central limitation in interpreting the comparison between English and Japanese T5 models is that the evaluated checkpoints are not fully comparable. The English T5 model has undergone prior fine-tuning specifically for question generation, while the Japanese T5 model is a base version without this fine-tuning. Since prompt inversion involves generating prompt-like text from observed outputs, this discrepancy may give an advantage to the English model, regardless of language. Therefore, the observed performance gap between the T5 models should be understood as a result of the combined effects of language, dataset properties, metric behavior, and differences in model architecture.

The results from GPT provide a valuable secondary reference, as they demonstrate a similar overall trend for BLEU-4, ROUGE-Lsum, METEOR, and BERTScore F1. However, these results do not completely eliminate the confounding factors, since the comparison with GPT is less comprehensive, and the outputs from the Japanese GPT show limited variation across different decoding settings. Therefore, while the cross-family results reinforce the broader conclusion that prompt inversion is sensitive to configuration, they should not be considered as definitive evidence of a purely linguistic effect.

Table 3. Representative GPT reconstruction results. One available configuration is shown for each strategy and language. Best values within each language block are shown in bold.

Language	Strategy	Config	BLEU-4	ROUGE-Lsum	METEOR	BERTScore F1	SBERT CosSim
English	Beam	beam=1	**0.02837**	**0.13610**	**0.18708**	**0.84517**	**0.55287**
English	Sampling	top-p=0.70, temp=0.5	0.02750	0.13404	0.18649	0.84495	0.55268
English	Hybrid	beam=3, top-p=0.95	0.02663	0.13122	0.18423	0.84471	0.55061
Japanese	Beam	beam=1	0.00633	0.11000	0.01841	0.68142	0.66877
Japanese	Sampling	top-p=0.70, temp=0.5	0.00633	0.11000	0.01841	0.68142	**0.66878**
Japanese	Hybrid	beam=3, top-p=0.95	0.00633	0.11000	0.01841	0.68142	0.66877

6.2. Metric-Dependent Interpretation of Reconstruction Quality

The results demonstrate that reconstruction quality should not be interpreted through a single metric. BLEU-4, ROUGE-Lsum, and METEOR are highly sensitive to surface realization and therefore emphasize lexical overlap between the original output and the regenerated output. BERTScore F1 remains more semantically oriented, but still favors English strongly in the current experiments. By contrast, Sentence-BERT cosine similarity often yields a different ranking, especially for Japanese and GPT-based outputs. This difference matters because a system may appear weak under overlap-based metrics while remaining relatively strong under sentence-level semantic similarity. Accordingly, prompt-reconstruction-based analysis should be evaluated through multiple complementary metrics rather than reduced to a single reconstruction score.

6.3. Ambiguity in The Prompt-Output Relationship

A further source of low reconstruction quality is the inherent ambiguity of the prompt–output relationship. Prompt inversion is not a one-to-one recovery problem: multiple prompts may plausibly correspond to the same output, and a reconstructed prompt may in turn generate a different but still valid output. As a result, a low similarity score does not necessarily imply that the reconstructed prompt is entirely unreasonable.

For example, the original text "Jim went to the store and bought eggs." may yield a reconstructed prompt such as "Edit the following sentence: Jim went to the store and bought eggs.", which then leads to a regenerated output such as "Rewrite the following concisely using complex sentences:". Similarly, the original text "Sure, x and y have now been assigned values of 2 and 3 respectively." may yield a reconstructed prompt such as "Assign the following values to the following pairs: x and y.", which then leads to "Assign two values to the following variables: $x = 2$, $y = 3$". In such cases, the pipeline drifts toward a different but still plausible generation path, and the final similarity score decreases even though the intermediate reconstruction is not entirely meaningless.

This observation suggests that reconstruction quality is influenced not only by model limitations, but also by the non-uniqueness of the mapping between prompts and outputs. In other words, part of the measured error may come from structural ambiguity in the task itself.

6.4. Implications for Decoding and Prefix Design

The experiments indicate that preferred decoding strategies can vary between languages and model configurations. For English T5 reconstruction, beam search emerges as the most effective strategy. In contrast, for Japanese T5 reconstruction, the strongest strategy is hybrid decoding. However, the results from the Japanese GPT demonstrate nearly identical scores across different decoding settings. This suggests that the sensitivity to decoding methods may depend on both the model used and the diversity of the generated outputs. Similarly, prefix augmentation is beneficial in English but generally harmful in Japanese. These results indicate that decoding and prefix design should not be transferred across languages without adjustment.

More broadly, the findings suggest that prompt-reconstruction pipelines are highly configuration-sensitive. A strategy that improves lexical recovery in one language may not improve semantic consistency in another, and a prefix that stabilizes one language may destabilize another. Therefore, practical use of reconstruction-based analysis requires language- and model-specific validation rather than assuming a single universal decoding or prefix setting.

6.5. Limitations and Future Work

This study has several limitations. First, our experiments are conducted exclusively on Alpaca-style, model-generated outputs and do not incorporate human-written text. Consequently, these findings illuminate prompt reconstruction behavior but should not be misconstrued as a direct measurement of human-versus-AI text detection performance. Second, our evaluation methodology relies on aggregate similarity scores rather than binary, threshold-based classification. Third, the scope of evaluated languages and model families is restricted. Our analysis encompasses only Japanese and English and focuses specifically on the T5 and GPT architectures. As a result, the observed trends may not readily generalize

to other languages, massively multilingual models, larger instruction-tuned models, or alternative prompt-inversion frameworks. Fourth, the English and Japanese T5 backbones are not identically initialized. Specifically, the English T5 model was fine-tuned for question generation, whereas its Japanese model is a standard base model. Although we standardized the task-specific training procedure and optimization hyperparameters across both languages, this baseline discrepancy remains a persistent confounding factor. Therefore, the observed performance gaps between the English and Japanese T5 models cannot be attributed solely to language-driven effects. Lastly, the GPT evaluation is inherently less comprehensive than the T5 comparison, as valid outputs could not be reliably obtained across all decoding configurations. Furthermore, the Japanese GPT evaluations show near-identical scores across the representative decoding strategies. As previously discussed, this uniformity is more indicative of limited output diversity in these specific settings than of true strong robustness to decoding choices.

Taken together, these results should be viewed as an initial empirical characterization of prompt inversion behavior across languages, metrics, and decoding settings, rather than as a definitive account of reconstruction-based detection performance. Future work should extend this framework to encompass mixed datasets containing both human-written and model-generated text, thereby enabling rigorous threshold-based detection evaluations. Future work should also evaluate more closely matched multilingual or architecture-equivalent checkpoints in order to separate language effects from checkpoint-specific effects more clearly. Additionally, it would be highly valuable to develop evaluation methodologies that explicitly account for the intrinsic many-to-one and one-to-many ambiguities in the prompt–output relationship, rather than assuming that successful reconstruction requires strict surface-level similarity.

7. Conclusion

In this paper, we presented a systematic evaluation of prompt inversion for LLM-generated text analysis in Japanese and English. Rather than proposing a new reconstruction algorithm, we focused on clarifying how reconstruction quality changes across similarity metrics, decoding strategies, prefix settings, and generator families. The experiments revealed four main findings. First, under the evaluated model and dataset settings, English obtains higher BLEU-4, ROUGE-Lsum, METEOR, and BERTScore F1 scores than Japanese, while Japanese achieves higher Sentence-BERT cosine similarity. Second, the preferred decoding strategy depends on the target language: beam search is strongest in English T5 reconstruction, whereas hybrid decoding is strongest in Japanese T5 reconstruction. Third, prefix augmentation substantially improves English T5 reconstruction but generally degrades Japanese T5 reconstruction. Fourth, the comparison between T5-style and GPT-style models shows that reconstruction quality is strongly metric-dependent and that different model families preserve different aspects of similarity. Furthermore, we also argue that measured reconstruction errors stem not solely from model limitations, but also from the inherent ambiguity of the prompt–output mapping.

Overall, our findings indicate that prompt inversion must be treated as a highly language-, metric-, and configuration-sensitive analysis framework. While the restricted scope of our evaluated models means that these results serve as a foundational empirical characterization rather than universally generalizable conclusions, they provide a critical analytical stepping stone. As detailed in Section 6.5, future work includes transitioning toward mixed human-AI datasets to facilitate threshold-based detection and developing ambiguity-aware evaluation methods.

References

[1] X. Yu. et al.: DPIC: Decoupling prompt and intrinsic characteristics for LLM generated text detection, Proceedings of the 38th Annual Conference on Neural Information Processing Systems, pp. 16194–16212 (2024).

[2] Z. Chen, Y. Feng, C. He, Y. Deng, H. Pu, Bo Li: IPAD: Inverse prompt for AI detection - a reliable and explainable LLM-generated essay detector, arXiv Preprint arXiv:2502.15902 (2025).

[3] M. Freitag, Y. Al-Onaizan: Beam search strategies for neural machine translation, arXiv Preprint arXiv:1702.01806 (2017).

[4] A. Fan, M. Lewis, Y. Dauphin: Hierarchical neural story generation, arXiv Preprint arXiv:1805.04833 (2018).

[5] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi: The curious case of neural text degeneration, arXiv Preprint arXiv:1904.09751 (2019).

[6] S. Zarrieß, H. Voigt, S. Schüz: Decoding methods in neural language generation: a survey, Information, vol. 12, no. 9, article 355 (2021).

[7] K. Papineni, S. Roukos, T. Ward, W. Zhu: Bleu: a method for automatic evaluation of machine translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002).

[8] C. Lin: Rouge: A package for automatic evaluation of summaries, Text Summarization Branches Out, pp. 74–81 (2004).

[9] S. Banerjee, A. Lavie: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 (2005).

[10] T. Zhang, V. Kishore, F. Wu, K. Weinberger, Y. Artzi: Bertscore: Evaluating text generation with bert, arXiv Preprint arXiv:1904.09675 (2019).

[11] N. Reimers, I. Gurevych: Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv Preprint arXiv:1908.10084 (2019).

[12] N. Shibayama, H. Shinnou: Construction and evaluation of Japanese sentence-BERT models, Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 731–738 (2021).

[13] C. Raffel et al.: Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67 (2020).

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever: Language models are unsupervised multitask learners, OpenAI Technical Report (2019).