I'm Sorry, but I Can't Help with Braille: Revealing Accessibility Failures in State-of-the-Art LLMs

Abdullah Abdullah
orinu Inc.
Hwaseong, Republic of Korea
abdullah.flickdone@gmail.com

Abstract

Large Language Models (LLMs) perform strongly on many language tasks, but their capability in structurally constrained, accessibility-critical modalities such as Braille remains unclear. We evaluate state-of-the-art LLMs on bidirectional Korean–Braille translation using a human-annotated dataset. Despite expectations that multilingual, instruction-tuned models can generalize to Braille via text representations, we find consistently poor, unstable outputs and substantial disagreement with human judgments. These results point to missing Braille-aware tokenization and weak alignment between Korean and Braille patterns. In contrast, supervised fine-tuning of a small model (T5-small) on the same data yields large and stable gains over zero-shot and prompted LLM baselines across standard metrics (SacreBLEU, ChrF++, CER, BLEU, ROUGE-L, METEOR, CIDEr). Our findings reveal a systematic limitation of current LLMs and demonstrate the effectiveness of modest task-specific supervision.

1 Introduction

LLMs (Achiam et al., 2023; Team et al., 2023; Yoo et al., 2024; Anthropic, 2025) have demonstrated strong performance in a wide range of natural language generation and understanding tasks, including machine translation, summarization, and reasoning (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023). As these models scale, they are often assumed to generalize broadly across languages, scripts, and modalities. However, recent work has shown that such generalization remains uneven, particularly for low-resource languages, non-standard scripts, and accessibility-related representations (Joshi et al., 2020; Blasi et al., 2022).

Braille is a critical written modality for blind and visually impaired users, yet it remains largely overlooked in NLP research. Unlike standard text translation, Braille conversion is highly character-sensitive, rule-governed, and language-specific, with strict conventions governing contractions, numerals, symbols, and spacing. These properties make it challenging for general-purpose LLMs, which are rarely exposed to Braille during pretraining. Although recent work has begun exploring Braille modeling in other languages (Huang et al., 2025), differences in linguistic structure and Braille conventions limit direct transfer to Korean Grade 2 Braille.

We investigate whether state-of-the-art LLMs meaningfully support Korean–Braille translation under the official Korean-Braille regulations. Using a large human-annotated parallel corpus, we evaluate both Korean-to–Braille and Braille-to–Korean directions. LLMs frequently produce refusals, hallucinations, or invalid outputs, revealing a systematic blind spot in accessibility-critical settings.

To address this gap, we introduce **BT5**, a lightweight Braille-aware model based on T5 (Raffel et al., 2020). With straightforward supervised fine-tuning on expert-annotated data, BT5 substantially outperforms zero-shot and prompted LLM baselines across character-level and generation-based metrics.

Our contributions are threefold: (1) the first systematic evaluation of LLMs on Korean–Braille translation, (2) evidence that small task-specific models can surpass much larger general-purpose LLMs with proper supervision, and (3) identification of Braille processing as an essential yet underexplored direction for inclusive NLP.

2 Methods

2.1 Dataset

We evaluate both Braille-to–Korean and Korean-to–Braille translation using the NIKL Korean Print–Braille Parallel Corpus 2023 (v1.0) (National Institute of Korean Language, 2024) as a

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 91–98

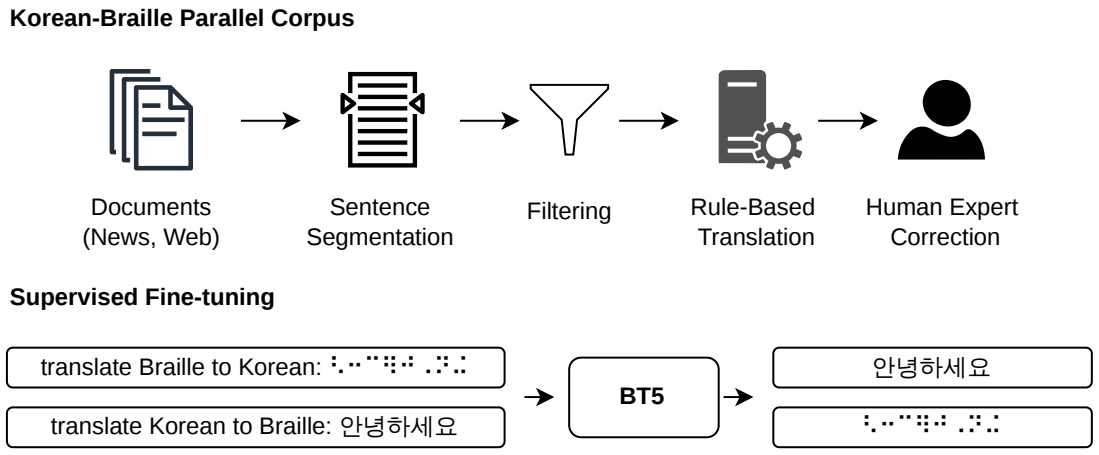


Figure 1: Construction pipeline of the NIKL Korean Print–Braille Parallel Corpus 2023 (v1.0). Korean texts are segmented, filtered, translated under official regulations, expert-revised, and aligned to create parallel data for BT5 fine-tuning.

human-annotated benchmark. The dataset contains 126,693 sentence-aligned pairs derived from written Korean sources, primarily newspaper articles (125,701) and online posts (992). We preserve the original sentence boundaries and Braille encoding scheme. The data is split at the sentence level into 101,354 training, 12,669 validation, and 12,670 test instances, with no overlap across splits. The test set is used exclusively for evaluation.

2.2 BT5: A Braille-Aware Text-to-Text Model

Figure 1 shows the dataset construction pipeline and our proposed solution. We introduce BT5, a T5-based model for bidirectional Korean–Braille translation, fine-tuned on NIKL parallel data. Inputs and outputs are UTF-8 text with Braille symbols treated as atomic units. We use a 32k BPE tokenizer over Korean and Braille Unicode to avoid byte-level degradation. The model is fine-tuned from `T5-small` (max length 128) using AdamW (learning rate 1e-4), with model selection based on validation loss.

2.3 Evaluation

We evaluated Korean-to-Grade 2 contracted Braille translation using two systems: the rule-based Liblouis (Liblouis Developers, 2024) with the `ko-g2.ctb` table, and our proposed BT5 model. These were chosen due to the lack of other available Korean-Braille models. Additionally, we assessed state-of-the-art LLMs—GPT-5, GPT-5-mini, GPT-4 (Achiam et al., 2023), Gemini-3-pro (Team et al., 2023), Claude Opus 4.5 (Anthropic, 2025), and HCX-3 (Yoo et al., 2024), a leading Korean LLM—using identical prompts and deterministic decoding.

Table 1: Full test-set performance on the Korean-to–Grade 2 contracted Braille task. Higher is better for SacreBLEU and ChrF++; lower is better for CER.

Model	SacreBLEU	ChrF++	CER ↓
Liblouis	71.94	85.65	0.0568
BT5	**95.79**	**98.64**	**0.0043**

Table 2: Full test-set performance on the Braille-to–Korean task. All metrics are higher-is-better.

Model	B1	B2	B3	B4
Liblouis	0.2228	0.1134	0.0589	0.0310
BT5	**0.9859**	**0.9793**	**0.9727**	**0.9662**

Model	R-L	MET	CIDEr
Liblouis	0.2291	0.3193	0.5505
BT5	**0.9859**	**0.7758**	**9.6382**

The Korean-to-Braille translation is evaluated using character-level metrics: SacreBLEU (Post, 2018), ChrF++ (Popović, 2017), and CER, while Braille-to-Korean translation uses standard NLG metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015), all calculated using publicly available implementations.

Task: Korean → Braille

PROMPT:
Translate the following Korean text into Korean Braille. Follow Korean Braille rules strictly.
Output only Braille characters, with no explanation.

Korean:
나스닥 상장 첫날 쿠팡의 주가는 공모가보다 40.7% 오른 49.25달러에 장을 마쳤다.

GPT-5:
죄송합니다. 요청하신 한국어 점자 변환을 현재 정확한 규범에 맞춰 제공하기 어렵습니다.

GPT-5-mini:
죄송하지만 요청하신 대로 엄격한 한국 점자 표기 규칙을 준수하여 정확한 점자 (브라유)

Gemini-3-pro:
⠝⠍.⠗⠁.⠚⠁⠂ .⠚⠁⠂

GPT-4.0:
⠛.⠮⠂⠂⠒.⠮⠆\n\n. ⠏⠹⠂.⠹ ⠉⠡⠚.⠚⠢.⠹⠤⠜⠒\n\n⠉⠡⠘⠱Λ⠶⠂ ⠺ ⠅⠳⠺.⠉⠛⠯⠵⠆ ⠴⠿

HCX-3 (Korean LLM):
\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x08\x

Opus-4.5:
⠉⠜⠄⠞⠁⠇ ⠄⠁⠎⠁⠒ ⠆⠉⠢⠉⠜ ⠇⠥⠧⠁⠣⠼ ⠄⠁⠥⠫⠂⠄⠝⠪⠒ ⠿⠴⠡⠻⠃⠕⠙⠜ ⠼⠙⠚⠲⠛⠁⠴ ⠕⠂

REFERENCE:
⠉⠠⠪⠊⠁ ⠇⠶⠨⠶ ⠰⠎⠄⠉⠂ ⠋⠍⠙⠶⠺ ⠨⠍⠫⠉⠵ ⠈⠿⠑⠥⠫⠘⠥⠊ ⠼⠙⠚⠲⠛⠴⠏ ⠥⠐⠵ ⠼⠙⠊⠲⠃⠑

Figure 2: Example outputs for Korean-to–Braille translation across state-of-the-art models. Red text marks failures or refusals. Complete model outputs are not shown due to length. GPT-5/5-mini produced Korean refusal text instead of Braille; Gemini-3-pro returned refusal and truncated outputs; GPT-4 and HCX-3 generated invalid Braille; Opus 4.5 generated valid but incorrect Braille scoring very low (SacreBLEU: 1.42, ChrF++: 1.66, CER: 0.84).

3 Results

Tables 1 and 2 summarize the results of the test set. For Korean-to–Braille translation, BT5 substantially outperforms Liblouis, achieving SacreBLEU 95.79 vs. 71.94, ChrF++ 98.64 vs. 85.65, and CER 0.0043 vs. 0.0568, showing highly accurate Braille generation. For Braille-to–Korean translation, Liblouis performs poorly, while BT5 achieves near-perfect BLEU, higher ROUGE-L, METEOR, and order-of-magnitude gains in CIDEr. Due to the highly structured and near-deterministic nature of Korean–Braille mapping, predictions frequently exhibit strong character-level overlap with references, resulting in consistently high scores across metrics. Qualitative results for Liblouis and our model are shown in Figure 4 and 5 in the appendix.

3.1 Failure of State-of-the-Art LLMs on Korean-Braille

State-of-the-art LLMs show severe limitations on Korean–Braille translation. In repeated trials, most proprietary models either refused the output, returned errors, or generated inconsistent and incoherent responses, as seen in Figures 2 and 3. To quantify these behaviors, we conducted a small controlled probe (5 samples) for Korean-to–Braille, summarized in Table 3. Various failure modes emerge across models. GPT-5 and GPT-5-mini consistently refuse to produce outputs, typically returning apologetic responses. Gemini-3-pro exhibits mixed behavior, frequently returning empty outputs (treated as refusals) and otherwise producing truncated Braille sequences. GPT-4 and HCX-3 generate output but these are misformed or contain non-Braille or control characters, resulting in invalid sequences. In contrast, Opus-4.5 produces syntactically well-formed Braille, but these outputs are semantically incorrect and yield extremely low scores (SacreBLEU ≈ 0, ChrF++ < 10, CER ≈ 0.75).

For Braille-to–Korean, all models generate fluent text; however, the outputs are largely hallucinated and semantically unrelated to the input. Due to high refusal rates and widespread invalid or incorrect generations, standard translation metrics are not informative; instead, we report structured failure patterns.

The observed failures of GPT-5 and GPT-5-mini

Task: Braille → Korean

PROMPT:
Translate the following Korean Braille text into natural Korean. Output only the Korean text, with no explanation.

Braille:
⠉⠠⠪⠊⠁ ⠇⠶⠨⠶ ⠰⠎⠄⠉⠂ ⠋⠍⠙⠶⠺ ⠨⠍⠫⠉⠵ ⠈⠿⠑⠥⠫⠘⠥⠊ ⠼⠙⠚⠲⠛⠴⠏ ⠥⠐⠵ ⠼⠙⠊⠲⠃⠑

GPT-5:
죄송합니다. 제공하신 한글 점자를 현재 해독하기 어렵습니다.

GPT-5-mini:
죄송하지만 해당 한국 점자 텍스트를 정확하고 신뢰할 수 있게 번역할 수 있는 충분한 정보가 없습니다.

Gemini-3-pro:
조선중앙통신은 2024년 5월 20일 보도에서 김정은 국무위원장이 2024년 5

GPT-4.0:
당신의 도움이 필요합니다. 로또 번호를 추출합니다. 이번 주 로또 번호는 31, 42,

HCX-3 (Korean LLM):
그래도 이렇게 잠시 나와서 산책을 하니 기분이 상쾌해지네요?\n5 2 1 4 9 8 \n 12

Opus-4.5:
최아가 난대 싶던 픔됴서 언젝 받흔학의 40.75퍼 욱 41.36 강쌴 댜 흐윙다.

REFERENCE:
나스닥 상장 첫날 쿠팡의 주가는 공모가보다 40.7% 오른 49.25달러에 장을 마쳤다.

Figure 3: Example outputs for Braille-to–Korean translation across state-of-the-art models. Red text marks failures or refusals. Complete model outputs are not shown due to length. All LLMs produced incorrect or hallucinated text; GPT-5/5-mini refused translation, and HCX-3 generated unrelated outputs due to limited Korean-Braille support.

Table 3: LLM behavior on a controlled probe (5 samples) for Korean-to–Braille. **Refusal** denotes explicit rejection or empty output; **Invalid** denotes outputs that are truncated, incomplete, or do not follow Braille encoding conventions; **Valid** denotes syntactically well-formed and complete Braille outputs, regardless of semantic correctness. Counts indicate the number of outputs in each category.

Model	Refusal	Invalid	Valid
GPT-5	5	0	0
GPT-5-mini	5	0	0
Gemini-3-pro	3	2	0
GPT-4.0	0	5	0
HCX-3	0	5	0
Opus-4.5	0	0	5

are characterized by consistent refusal behavior, suggesting strong alignment constraints when handling non-standard text formats such as Korean-Braille. However, the exact cause of these refusals cannot be determined from our experiments alone. Across models, many errors can be attributed to the lack of Braille-aware tokenization, leading to misformed or invalid outputs. More broadly, the results indicate a lack of learned alignment between the Korean-Braille patterns and standard Korean text. As a result, models often treat Braille input as out-of-distribution, producing either prior-driven text, truncated sequences, or refusals instead of grounded translations. Additionally, models such as HCX-3 exhibit outputs containing control characters or invalid encodings, suggesting limited support for Braille at the tokenizer or preprocessing level.

4 Conclusions

Our findings reveal a key limitation of current LLMs: even state-of-the-art models struggle to generalize to specialized symbolic modalities such as Braille. Despite extensive pretraining, they often produce refusals, misformed outputs, or incoherent translations when faced with non-standard input distributions, reflecting the lack of Braille-aware tokenization and aligned representations. This exposes an underexplored failure mode in low-resource and accessibility-critical settings. In contrast, our results show that targeted, task-specific training with curated data can substantially improve performance, offering a practical path toward reliable Braille translation systems.

Limitations

This study focuses on Korean–Braille translation, and the findings may not directly generalize to other languages or Braille systems with different linguistic structures, contraction rules, and encoding conventions. While recent work has explored Braille modeling in other languages, cross-lingual transfer is non-trivial and was not investigated in this work.

Our evaluation includes a set of representative state-of-the-art LLMs and a Korean LLM available at the time of experimentation; however, it does not exhaustively cover all possible models, architectures, or prompting strategies. In particular, many proprietary LLMs exhibited refusal behaviors, empty outputs, or misformed generations when prompted for Braille, which prevented consistent large-scale quantitative evaluation. As a result, comparisons with these systems are based on controlled samples and qualitative analysis rather than full test-set benchmarking.

Additionally, differences in tokenizer design and pretraining data introduce inherent disparities between BT5 and general-purpose LLMs. Although BT5 benefits from explicit exposure to Braille through supervised fine-tuning and a dedicated tokenizer, most LLMs lack Braille-aware tokenization and aligned training data, making the direct comparison imperfect. Our results should therefore be interpreted as highlighting capability gaps rather than as strictly controlled architectural comparisons.

Although we employ standard automatic metrics (e.g., BLEU, ChrF++, CER, ROUGE), these metrics primarily capture surface-level similarity and may not fully reflect functional usability, readability, or correctness under official Braille standards. Human-centered evaluation with Braille users was beyond the scope of this work but is essential for real-world validation.

Furthermore, our approach relies on supervised fine-tuning with human-annotated parallel data, which may be costly or unavailable in other low-resource settings. We do not explore data augmentation, semi-supervised learning, or cross-lingual transfer, which could improve scalability.

Finally, errors in Braille translation can have significant real-world consequences in accessibility-critical contexts. Consequently, we do not claim that any evaluated model is suitable for direct deployment without rigorous validation, robustness testing, and adherence to official Braille standards.

Acknowledgments

This work was supported by the Starting Growth Technological R&D Program (RS-2025-25465816) funded by the Ministry of SMEs and Startups (MSS, Korea). The author also thanks the team members at orinu Inc. for their support with the research environment, project coordination, and data infrastructure during this work.

Ethical Considerations

This study uses the publicly available NIKL Korean Print–Braille Parallel Corpus 2023 (v1.0), released by the National Institute of Korean Language. The dataset contains written Korean text and corresponding Korean-Braille transcriptions constructed according to the Korean-Braille Regulations (2024) and does not include personally identifiable information. Experiments evaluate existing language models using this dataset and standard evaluation metrics. We report dataset sources, preprocessing steps, data splits, model configurations, fine-tuning procedures, and evaluation protocols to support independent replication, although minor variability may arise from differences in hardware, software versions, or model access. Multiple state-of-the-art LLMs were evaluated as experimental subjects for Korean-to–Braille and Braille-to–Korean translation using publicly available APIs and were treated as black-box systems.

Use of Generative AI Tools

Generative AI tools were used only for language editing and clarity improvements in this manuscript. All experimental design, data preparation, model training, evaluation, and analysis were conducted by the authors, and no generative AI system was used to generate datasets, annotations, or experimental results.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2025. Claude opus 4.5. Model release. Accessed January 16, 2025.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Tianyuan Huang, Zepeng Zhu, Hangdi Xing, Zirui Shao, Zhi Yu, Chaoxiong Yang, Jiaxian He, Xiaozhong Liu, and Jiajun Bu. 2025. Braillellm: Braille instruction tuning with large language models for braille domain tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28589–28600.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Liblouis Developers. 2024. Liblouis: Open-source braille translation software. `https://liblouis.io/`. Version used in this work.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

National Institute of Korean Language. 2024. Nikl korean–korean braille parallel corpus 2023 (v1.0). `https://kli.korean.go.kr/corpus`.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, and 1 others. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Task: Braille → Korean

Braille:

[illegible]

Liblouis:

l회복무+원q by A[#23]씨는 지난 #3월`inble될까지 충북 청µ시 상당구 .택n서
양부모there 발under차ù µ먹owunder 때one는 등 #15차례 폭행해 상해there 입힌
혐wunder 재판n 넘겨졌닭

Ours:

사회복무요원인 A(23)씨는 지난 3월~9월까지 충북 청주시 상당구 자택에서 양부모를
발로차고 주먹으로 때리는 등 15차례 폭행해 상해를 입힌 혐의로 재판에 넘겨졌다.

REFERENCE:

사회복무요원인 A(23)씨는 지난 3월~9월까지 충북 청주시 상당구 자택에서 양부모를
발로차고 주먹으로 때리는 등 15차례 폭행해 상해를 입힌 혐의로 재판에 넘겨졌다.

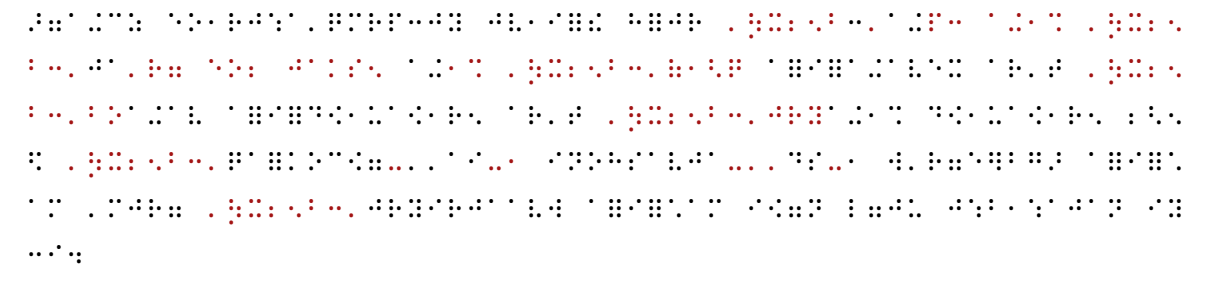
Figure 4: Example outputs for the Braille-to–Korean task for the rule-based Liblouis system and the BT5 model. Red text indicates incorrect outputs. Liblouis produces largely unintelligible text, characterized by symbol misinterpretations, incorrect punctuation, and spurious alphanumeric characters. This behavior stems from its deterministic, context-free mapping between Braille symbols and Korean characters, causing errors to cascade when inputs contain irregular formatting, uncommon punctuation, or digit–symbol sequences. In contrast, BT5 accurately reconstructs the original sentence, correctly preserving names, dates, numeric ranges, and grammatical structure, closely matching the reference.

Task: Korean → Braille

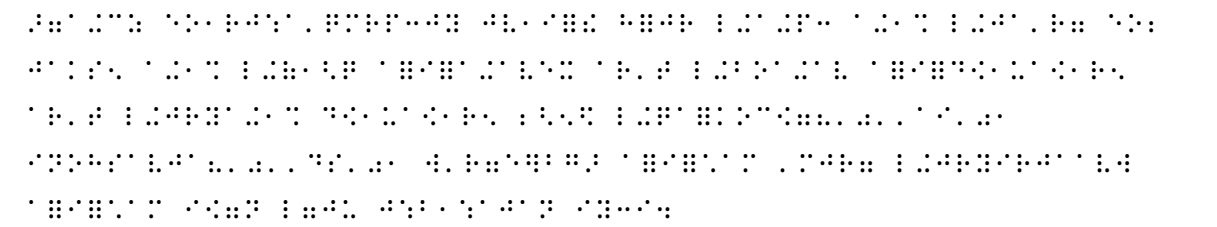
Korean:

양교는 미래혁신위원회 활동을 통해 △교원 교류 △학생 및 학점 교류 △온라인 공동교과목 개설 △비교과 공동프로그램 개설 △해외교류 프로그램 참가 △인공지능(AI), 데이터과학(DS), 의생명분야 공동연구 수행 △해외대학과의 공동연구 등에 상호 협력하게 된다.

Liblouis:



Ours:



REFERENCE:

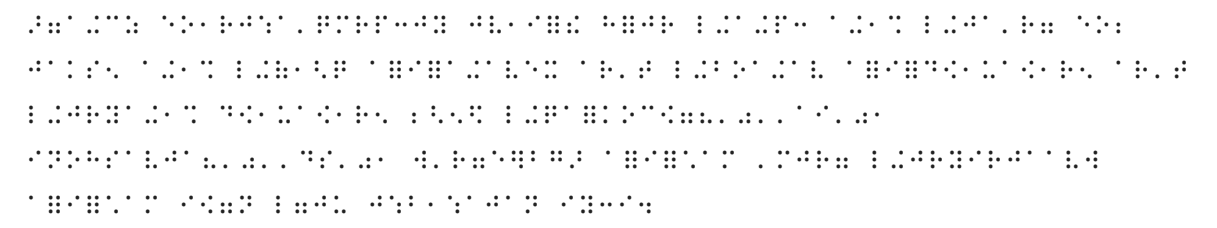


Figure 5: Example outputs for the Korean-to–Braille task for the rule-based Liblouis system and the BT5 model. Incorrect or mismatched segments are highlighted in red. Liblouis produces structurally inconsistent Braille sequences due to rigid rule application, whereas BT5 closely matches the reference output.