Appendix

The following appendix sections propose additional experimental results and specifications for LAMA-UT, which can be instrumental in deeper understanding or validating the strengths of the proposed pipeline. The first section of the appendix presents the specifications of the languages we leveraged in the experiments, given the corresponding ISO-639 language code. Consequently, the second section demonstrates the transcription performance of LAMA-UT for each language used in the experiments. Then, the third section suggests the end-to-end transcription performance comparison between IPA-based LAMA-UT and Romanization-based LAMA-UT, further to validate the effectiveness of Romanization in orthography unification. Finally, in the last section, we would like to discuss possible advancements of LAMA-UT and our further research direction.

Languages and Language Codes

Table 8 indicates the languages that were leveraged in the experiments of our manuscript. All of the seen languages comprise 102 languages in the FLEURS dataset, and they are used in both training and evaluation of LAMA-UT. Evaluation samples for unseen languages were chosen from the official test split of the CommonVoice 17.0 dataset, which ensured that they possessed an adequate volume of data.

Language-Specific Performance of LAMA-UT

In this section, we present a detailed performance analysis for each language, comparing transcription performance from the universal transcription generation phase to that after passing through the universal converter. Additionally, we analyze these results further to explore the correlation between orthographic characteristics and transcription performance. First of all, Table 9 suggests the language-specific transcription performance of the Romanization-based universal transcription generator, and Table 10 presents the language-specific end-to-end (E2E) language-specific transcription performance of LAMA-UT. In the case of the universal transcription generator, CERs were consistently similar across most languages with minimal variation. However, in the end-to-end pipeline with language-specific transliteration, notably higher error rates were observed for certain languages, and the languages highlighted in bold represent the top 10 languages with the most significant increases in error rate observed in the end-to-end pipeline compared to the CER values during the universal transcription phase. Interestingly, the 10 languages that exhibited the most significant performance degradation after passing through the universal converter were all non-Latin script languages, with the majority being languages that do not employ spacing in their orthography. This suggests that when a pre-trained LLM is utilized as a universal converter, languages with non-Latin orthography which inherently exhibit different structural characteristics compared to Latin-based languages, are more prone to error propagation.

E2E Comparison on Unification Methods

In this section, we would like to further clarify the effectiveness of utilizing Romanization as an intermediate representation. Since the experimental results from Table 2 were limited to the universal transcription generation phase, we supplement the previous results with the full performance comparison between IPA-based and Romanization-based end-to-end architecture of LAMA-UT, both in upper bound assessment and transcription performance. Table 11 proposes the end-to-end performance comparison between IPA-based LAMA-UT and Romanization-based LAMA-UT. The only difference between the two models is the orthography unification method utilized in the universal transcription generator. (e.g., the IPA-based model utilizes IPA as an intermediate representation, while the Romanization-based one leverages Romanized transcription as an intermediate representation. In the results, Romanization-based LAMA-UT consistently outperforms IPA-based LAMA-UT as a substantial difference in both CER and WER. These results strongly demonstrate the superiority of Romanization over IPA when leveraging LLMs as a universal converter, since LLMs are primarily trained with Latin scripts and optimized tokenization strategy for them. Furthermore, there are some inconsistencies in the results of the IPA-based LAMA-UT, where passing predicted IPA transcriptions along with a few examples to the universal converter yielded better performance than passing ground truth IPA transcriptions without examples. This is presumably because the LLMs did not frequently encounter IPA-driven tokens during its training process.

Discussion and Future Work

LAMA-UT showed comparable or better performance compared to previous works even without language-specific modules (e.g., adapters, lexicons, n-gram LMs), while achieving efficient training with a significantly reduced dataset size. However, there is still room for further improvement in both the universal transcription generator and the universal converter. First, the transcription performance of the universal transcription generator can be improved. For instance, the universal transcription generator of LAMA-UT can leverage additional linguistic information (e.g., embedding from a pre-trained language classifier) to further enhance the transcription quality of the first phase. Secondly, our pipeline shows relatively lower performance for languages with distinct linguistic structures, like Korean, and those with additional features (e.g., tones), such as Chinese, in the language-specific transliteration phase. Since our universal converter is replaceable, this issue will naturally be resolved in line with the development of LLMs. Finally, the utilization of prompt learning techniques (Li and Liang 2021; Liu et al. 2022; Gu et al. 2022) might improve transliteration performance. We plan to address these aspects in future research.

Seen	Afrikaans (af), Amharic (am), Arabic (ar), Assamese (as), Asturian (ast), Azerbaijani (az), Belarusian (be), Bulgarian (bg), Bengali (bn), Bosnian (bs), Catalan (ca), Cebuano (ceb), Sorani-Kurdish (ku), Mandarin Chinese (cmn), Czech (cs), Welsh (cy), Danish (da), German (de), Greek (el), English (en), Spanish (es), Estonian (et), Persian (fa), Fula (ff), Finnish (fi), Filipino (fil), French (fr), Irish (ga), Galician (gl), Gujarati (gu), Hausa (ha), Hebrew (he), Hindi (hi), Croatian (hr),Hungarian (hu), Armenian (hy), Indonesian (id), Igbo (ig), Icelandic (is), Italian (it), Japanese (ja), Javanese (jv), Georgian (ka), Kamba (kam),Kabuverdianu (kea), Kazakh (kk), Khmer (km), Kannada (kn), Korean (ko), Kyrgyz (ky), Luxembourgish (lb), Ganda (lg), Lingala (ln), Lao (lo), Lithuanian (lt), Luo (luw), Latvian (lv), Maori (mi), Macedonian (mk), Malayalam (ml), Mongolian (mn), Marathi (mr), Malay (ms), Maltese (mt), Burmese (my), Norwegian (no), Nepali (ne), Dutch (nl), Northern-Sotho (nso), Nyanja (ny), Occitan (oc), Oromo (om), Oriya (or), Punjabi (pa), Polish (pl), Pashto (ps), Portuguese (pt), Romanian (ro), Russian (ru), Sindhi (sd), Slovak (sk), Slovenian (sl), Shona (sn), Somali (so), Serbian (sr), Swedish (sv), Swahili (sw), Tamil (ta), Telugu (te), Tajik (tg), Thai (th), Turkish (tr), Ukrainian (uk), Umbundu (umb), Urdu (ur), Uzbek (uz), Vietnamese (vi), Wolof (wo), Xhosa (xh), Yoruba (yo), Cantonese Chinese (yue), Zulu (zu)
Unseen	Abkhazian (ab), Albanian (sq), Basaa (bas), Bashkir (ba), Basque (eu), Breton (br), Chuvash (cv), Eastern Mari (mhr), Erzya (myv), Esperanto (eo), Guarani (gn), Hakha Chin (cnh), Interlingua (ia), Kinyarwanda (rw), Latgalian (ltg), Norwegian Nynorsk (nn), Romansh (rm), Tatar (tt), Toki Pona (tok), Turkmen (tk), Uighur (ug), Upper Sorbian (hsb), Western Frisian (fy), Western Mari (mrj), Yakut (sah)

Table 8: Specifications of 102 seen languages and 25 unseen languages for experiments. We primarily reported ISO-639-1 codes and provided ISO-639-3 codes when the former were unavailable. Sursilvan and Vallader dialects are both considered Romansh languages.

	Language	CER ↓	Language	CER ↓	Language	CER ↓	Language	CER ↓
Seen	Afrikaans	19.3	Ganda	10.6	Lithuanian	8.7	Shona	9.0
	Amharic	8.7	Georgian	9.4	Luo	7.4	Sindhi	29.7
	Arabic	8.8	German	8.4	Luxembourgish	13.6	Slovak	5.4
	Armenian	5.6	Greek	11.5	Macedonian	5.3	Slovenian	9.1
	Assamese	12.9	Gujarati	8.9	Malay	8.3	Somali	16.8
	Asturian	9.3	Hausa	9.7	Malayalam	7.7	Sorani-Kurdish	11.8
	Azerbaijani	10.7	Hebrew	20.5	Maltese	8.4	Spanish	4.7
	Belarusian	8.0	Hindi	11.6	Mandarin Chinese	6.5	Swahili	6.4
	Bengali	9.8	Hungarian	14.5	Maori	8.7	Swedish	11.8
	Bosnian	6.9	Icelandic	18.7	Marathi	11.9	Tajik	5.3
	Bulgarian	7.6	Igbo	14.0	Mongolian	14.7	Tamil	11.9
	Burmese	19.4	Indonesian	6.0	Nepali	11.2	Telugu	10.8
	Cantonese Chinese	16.6	Irish	28.4	Northern-Sotho	13.8	Thai	15.9
	Catalan	7.5	Italian	3.7	Norwegian	9.2	Turkish	6.4
	Cebuano	7.5	Japanese	33.4	Nyanja	12.4	Ukrainian	11.2
	Croatian	6.3	Javanese	7.0	Occitan	17.7	Umbundu	9.5
	Czech	11.0	Kabuverdianu	6.8	Oriya	14.3	Urdu	44.0
	Danish	15.6	Kamba	10.9	Oromo	20.2	Uzbek	16.3
	Dutch	11.3	Kannada	7.4	Pashto	20.0	Vietnamese	13.0
	English	13.1	Kazakh	5.1	Persian	6.9	Welsh	12.8
	Estonian	4.4	Khmer	23.6	Polish	10.9	Wolof	14.9
	Filipino	4.8	Korean	9.6	Portuguese	9.0	Xhosa	10.4
	Finnish	4.9	Kyrgyz	5.2	Punjabi	20.3	Yoruba	10.9
	French	12.3	Lao	17.8	Romanian	11.8	Zulu	9.5
	Fula	14.7	Latvian	7.0	Russian	9.0		
	Galician	6.7	Lingala	7.7	Serbian	8.9		
Unseen	Abkhazian	45.8	Eastern Mari	30.9	Latgalian	28.8	Upper Sorbian	35.3
	Albanian	34.8	Erzya	29.9	Norwegian Nynorsk	27.1	Western Frisian	33.8
	Basa	33.4	Esperanto	22.1	Romansh	31.9	Western Mari	36.7
	Bashkir	33.5	Guarani	33.5	Tatar	31.8	Yakut	36.3
	Basque	33.0	Hakha Chin	42.9	Toki Pona	28.7		
	Breton	49.1	Interlingua	17.3	Turkmen	41.3		
	Chuvash	37.4	Kinyarwanda	37.2	Uighur	27.1		

Table 9: Performance of Romanization-based universal transcription generator in 102 seen and 25 unseen languages. Results are reported in CER, which calculates the character-level difference between ground truth Romanized transcription and predicted Romanized transcription.

	Language	CER ↓	Language	CER ↓	Language	CER ↓	Language	CER ↓
Seen	Afrikaans	18.3	Ganda	10.4	Luo	7.3	Sindhi	29.1
	Amharic	**47.4**	Georgian	23.6	Luxembourgish	16.5	Slovak	3.8
	Arabic	13.9	German	5.5	Macedonian	5.5	Slovenian	9.4
	Armenian	**22.4**	Greek	15.6	Malay	6.9	Somali	15.4
	Assamese	18.9	Gujarati	11.9	**Malayalam**	**24.1**	Sorani-Kurdish	26.4
	Asturian	9.8	Hausa	10.2	Maltese	9.8	Spanish	2.8
	Azerbaijani	11.9	**Hebrew**	**35.2**	**Mandarin Chinese**	**36.0**	Swahili	5.3
	Belarusian	11.3	Hindi	8.3	Maori	12.0	Swedish	9.5
	Bengali	13.8	Hungarian	16.7	Marathi	14.0	Tajik	8.9
	Bosnian	5.9	Icelandic	18.1	Mongolian	21.4	Tamil	19.6
	Bulgarian	8.2	Igbo	17.9	Nepali	13.2	Telugu	15.5
	Burmese	**52.6**	Indonesian	4.3	Northern-Sotho	15.1	**Thai**	**45.3**
	Cantonese Chinese	**61.6**	Irish	34.8	Norwegian	7.0	Turkish	5.1
	Catalan	6.3	Italian	2.0	Nyanja	12.1	Ukrainian	9.7
	Cebuano	7.4	Javanese	6.6	Occitan	19.9	Umbundu	10.5
	Croatian	5.2	Kabuverdianu	8.3	Oriya	18.3	Urdu	14.9
	Czech	9.9	Kamba	14.6	Oromo	19.8	Uzbek	14.5
	Danish	12.9	Kannada	13.9	Pashto	32.7	Vietnamese	19.0
	Dutch	8.7	Kazakh	6.7	Persian	11.9	Welsh	12.3
	English	8.2	**Khmer**	**66.3**	Polish	8.6	Wolof	18.0
	Estonian	5.7	Korean	18.8	Portuguese	7.9	Xhosa	10.5
	Filipino	4.4	Kyrgyz	8.8	Punjabi	19.5	Yoruba	25.5
	Finnish	4.7	**Lao**	**55.0**	Romanian	11.0	Zulu	9.6
	French	8.2	Latvian	7.9	Russian	6.0		
	Fula	16.6	Lingala	7.8	Serbian	5.7		
	Galician	6.0	Lithuanian	8.6	Shona	8.6		
Unseen	Abkhazian	55.9	Eastern Mari	35.3	Latgalian	34.7	Upper Sorbian	35.2
	Albanian	35.4	Erzya	35.7	Norwegian Nynorsk	23.9	Western Frisian	29.6
	Basa	42.1	Esperanto	19.7	Romansh	30.0	Western Mari	41.4
	Bashkir	29.7	Guarani	40.0	Tatar	24.2	Yakut	38.2
	Basque	32.8	Hakha Chin	42.8	Toki Pona	29.0		
	Breton	49.9	Interlingua	15.2	Turkmen	43.1		
	Chuvash	44.3	Kinyarwanda	36.7	Uighur	24.5		

Table 10: Performance of LAMA-UT in 101 seen and 25 unseen languages. We leveraged a Romanization-based universal transcription generator and GPT-4o-mini model with a few-shot prompting as a universal converter. All of the results are reported in CER, which calculates the character-level difference between ground truth language-specific transcription and predicted language-specific transcription. Japanese was excluded from the results due to the ambiguity in evaluation criteria arising from its mixed use of Hiragana, Katakana, and Kanji.

	Universal Converter	IPA				Roman			
		Upper Bound		Few Shot		Upper Bound		Few Shot	
		CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓	WER ↓
Seen	LLaMA-8B	51.4	82.4	56.5	80.8	33.0	44.4	38.1	64.7
	LLaMA-70B	40.0	64.6	33.8	67.8	15.6	34.6	23.1	56.2
	GPT-4o-mini	27.3	49.2	33.0	70.6	10.0	28.6	19.9	49.4
Unseen	LLaMA-8B	52.0	93.0	43.4	93.3	32.7	60.2	41.6	94.4
	LLaMA-70B	42.5	89.0	38.1	94.6	17.0	57.5	34.7	93.9
	GPT-4o-mini	40.4	78.2	35.2	84.1	12.3	42.1	29.7	83.0

Table 11: An end-to-end (E2E) comparison between IPA and Romanization. Since *Phonemizer* (IPA transliterator) supports a smaller amount of language compared to *Uroman* (Romanization transliterator), all of the metrics were calculated and averaged within the common languages which are both available in the IPA-based and the Romanization-based version of the universal transcription generator.