

Lost Between Scripts: Recovering Arabic Words from Noisy Learner Orthographies

Mohamed Adel Nisa Shahid Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi
{mohamed.ade1,ns5376,nizar.habash}@nyu.edu

Abstract

Recovering an intended word from a noisy spelling is a general word-retrieval problem relevant to language learning, spelling correction, and cross-script search. We study this problem in Arabic, where learners produce variable transcriptions in Arabic or Roman script that may preserve only part of the target form. Using ArabScribe, we evaluate four large language models (LLMs) on ranked intended-word recovery under three cue conditions: the noisy form, an English meaning gloss, and both. Combining form and meaning consistently outperforms either cue alone across models and scripts, with particularly strong benefits for distorted and Roman-script input. An expert-human evaluation shows the same benefit from glosses and places the strongest models near individual expert performance on the sampled items. Analyses of edit distance, learner proficiency, and target frequency show that semantic information improves recovery across difficulty conditions but does not eliminate sensitivity to form noise and lexical rarity. These findings establish gloss-grounded intended-word recovery as a useful framework for studying how LLMs integrate form and meaning under noisy, cross-script input.

1 Introduction

People often search for a word with only partial knowledge of its form and meaning. A learner may remember approximately how it sounded and what it meant but not its conventional spelling, especially in an unfamiliar writing system. Spelling correction and phonetic search emphasize observed form, whereas reverse dictionaries begin from a definition (Zhang et al., 2017; Hill et al., 2016; Zhang et al., 2019; Qi et al., 2020). In realistic lookup, both cues may be available and uncertain. We ask whether large language models (LLMs) can combine them to recover an intended word.

Arabic provides an informative setting for this

Target: قَرِيب qariyb ‘near’			
قريب qryb	كريب kryb	گریب gryb	قریب qryyb
qareeb	kareb	gareeb	qreeb
8rib	kareeb	8arib	8areb

Table 1: Example intended-word recovery task in Arabic and Roman script, showing learner transcription variability across scripts and conventions. Arabic transliterations are in the HSB scheme (Habash et al., 2007).

question. Learners may transcribe a heard or remembered word in Arabic or Roman script, preserving only part of its form. Arabic writing commonly omits short vowels and contains morpho-orthographic ambiguity (Habash, 2010), while Roman-script spellings follow no single convention (Darwish, 2014; Eryani and Habash, 2021). One target can therefore produce many plausible transcriptions, as Table 1 illustrates.

We augment ArabScribe’s learner transcriptions of spoken Modern Standard Arabic (MSA) words (Zhang et al., 2017) with curated English glosses and formulate *gloss-grounded intended-word recovery*.¹ Given a noisy transcription, optionally paired with a meaning cue, a model returns ranked Arabic candidates. We evaluate prompted LLMs using form alone, meaning alone, and both across Arabic- and Roman-script input. Expert human responses and analyses of cue interaction, edit distance, learner proficiency, target frequency, and unresolved cases complement the model evaluation. Across models and scripts, combining form and meaning produces the strongest performance, especially for Roman-script and heavily distorted input, and substantially reduces the disparity between scripts. Human experts show the same benefit from glosses, while the strongest model approaches human expert performance. Semantic information

¹<https://github.com/CAMeL-Lab/ArabScribe>

also narrows proficiency-related differences and helps across the frequency range, although rare targets, misleading transcriptions, and exact-form confusions remain difficult. Cases recovered only with both cues indicate complementary evidence rather than duplication.

This work contributes a formulation of the underexplored task of gloss-grounded intended-word recovery, joining noisy-form recovery with semantic retrieval; a controlled extension of ArabScribe (Zhang et al., 2017) for comparing cues across scripts; and human-grounded analyses of when semantic information helps and why recovery fails. The framework connects spelling correction, phonetic search, reverse dictionaries, and assistive language tools, and can extend beyond Arabic.

2 Background and Related Work

Recovering intended Arabic words from learner transcriptions is difficult because Arabic writing is primarily consonantal, often omitting short vowels, and learners may confuse unfamiliar uvular, pharyngeal, and emphatic consonants (Habash, 2010). Cross-script input adds ambiguity: Arabic–Latin transliteration lacks one-to-one mappings (Habash et al., 2007), and Arabizi uses variable conventions and numeric substitutions such as 3 for ع and 7 for ح (Darwish, 2014). This variation is documented in Arabizi corpora and modeled by neural transliteration systems (Bies et al., 2014; Shazal et al., 2020). Unlike transliteration, our task recovers the intended Arabic word from noisy learner input rather than converting between scripts.

ArabScribe studies this problem through Arabic- and Roman-script transcriptions of spoken MSA words from learners with diverse backgrounds (Zhang et al., 2017). Zhang et al. (2017) introduce a MultiScript Phonetic Search algorithm that maps Arabic- and Roman-script inputs into a shared phonetic representation and applies specialized edit-distance matching. Direct comparison is limited because their method performs dictionary lookup over a fixed lexicon, while our LLM-based setup generates ranked candidates and can incorporate gloss information.

Our gloss-augmented setting also connects to reverse dictionary work, where systems retrieve words from definitions or descriptions (Hill et al., 2016; Zhang et al., 2019; Qi et al., 2020). Unlike standard reverse dictionaries, our task combines a meaning cue with a noisy Arabic or Roman-script

learner transcription. We address this gap by evaluating whether LLMs can recover intended Arabic words from noisy learner spellings, especially when spelling alone is ambiguous.

3 Experimental Setup

Problem formulation. We formulate the task as ranked intended-word recovery. Given a noisy learner transcription in Arabic or Roman script, the model returns ten ranked Arabic word candidates. A prediction is correct if the intended gold word appears in the ranked list.

Prompting. All models receive the same instruction and must return exactly ten unique Arabic word candidates ranked from most to least likely. To reduce formatting errors, outputs are requested as JSON through model API structured-output modules. We evaluate three cue conditions: *Word*, where the model receives only the noisy transcription; *Gloss*, where it receives only the English gloss as an ablation and semantic-control condition; and *Word+Gloss*, where the model receives both cues, simulating a learner who knows the meaning but not the standard Arabic form. Prompt templates are in Appendix C. Our primary evaluation is 0-shot. We additionally evaluate 5-shot and 20-shot prompting, with demonstrations randomly drawn from the training split and matched to the test condition format.

Dataset. We use ArabScribe (Zhang et al., 2017), which contains Arabic- and Roman-script learner transcriptions of spoken MSA words from users with different Arabic experience levels and native-language backgrounds. We enrich ArabScribe with English glosses from the Standard Arabic Morphological Analyzer (SAMA) (Graff et al., 2009), extending the original form-only task to cases where learners may also know or describe a word’s meaning. We query the SAMA Arabic morphological database using the diacritized gold word form. When SAMA returns multiple analyses or glosses for a word, we retain all corresponding glosses rather than manually selecting a single one. Following Zhang et al. (2017), we use the standard train/test split: training items are used only for random few-shot selection, and evaluation is conducted on the test set. The test set contains 654 Arabic-script and 1,342 Roman-script transcriptions. Additional statistics, licenses, and access terms appear in Appendices A and D.

Models. We evaluate Gemini-3.1-Pro (Team et al., 2025), GPT-5.4 (Singh et al., 2026), Llama-4-Scout (Meta, 2025), and Fanar-2 (FANAR TEAM et al., 2026) using prompting only, without fine-tuning. These models cover closed and open-weight systems, multilingual models, and an Arabic-centric model. Model licenses or service terms are listed in Appendix A, and model sizes and computational budget are reported in Appendix B.

Metrics. We report Acc@1, Acc@10, and mean reciprocal rank (MRR). Acc@1 measures top-ranked accuracy, Acc@10 measures whether the gold word appears anywhere in the list, and MRR rewards higher-ranked correct predictions. We evaluate exact, undiacritized, and full normalization, where full normalization includes diacritization and normalization of Alif, Ya, and Ta-Marbuta variants. We use full normalization for the main results because the task concerns lexical recovery rather than strict surface-form matching: predictions that differ from the gold only in diacritics or common orthographic variants still recover the intended word.

Human baseline. Two expert Arabic annotators (Egyptian and Lebanese native speakers) independently evaluated 200 test items: for each script, we randomly selected one learner transcription for each of the 100 target words. The annotators were told that targets were MSA words and could be rare. Each annotator first completed all *Word* items without access to the glosses, then completed the same items under *Word+Gloss*; they were instructed not to revise their first responses after seeing the glosses. Because each annotator supplied one candidate per item, we evaluate human responses as Acc@1 under the same full normalization as the models. We compare them with the top-ranked 0-shot predictions of the four models on exactly the same items. Annotator backgrounds, procedure, and detailed agreement appear in Appendix E.

4 Results

Table 2 presents the primary 0-shot results. Full-normalization, exact-match, and undiacritized results for 0, 5, and 20 shots appear in Appendix F.

Benefit of combined cues. The strongest pattern is that *Word+Gloss* consistently outperforms either cue alone across both scripts and all models. Averaged across the four models, Arabic-script Acc@1 rises to 58.2 with *Word+Gloss*, compared with 40.0

for *Word* and 28.6 for *Gloss*. For Roman-script input, the corresponding averages are 57.1, 24.6, and 26.4. Acc@10 and MRR show the same combined-cue advantage, indicating that the gloss complements rather than replaces the noisy transcription.

For reference, Zhang et al. (2017) report Acc@1 of 46.0% and 55.0% for Roman-script and Arabic-script input, respectively, in the original form-only retrieval setting. These numbers are closest to our *Word* condition, while *Word+Gloss* evaluates the additional benefit of semantic information.

Script effects. Roman-script transcriptions are harder than Arabic-script transcriptions when used alone: averaged across models, *Word* Acc@1 is 24.6 for Roman-script input and 40.0 for Arabic-script input, a 15.4-point gap. With *Word+Gloss*, the averages converge to 57.1 and 58.2, reducing the gap to 1.1 points. Semantic grounding therefore largely offsets the additional ambiguity introduced by Roman-script learner transcriptions.

Model effects. Gemini-3.1-Pro is strongest overall in the 0-shot setting, followed by GPT-5.4. Llama-4-Scout and Fanar-2 also benefit from combining the cues but remain lower in absolute performance, suggesting substantial model differences in the ability to integrate form and meaning. The secondary 5-shot and 20-shot experiments show smaller and less consistent changes than cue choice; complete results appear in Appendix F.

Human baseline. The expert results in Table 3 corroborate the value of semantic cues. On Arabic-script items, the two annotators obtain 43 Acc@1 with *Word* and 67–77 with *Word+Gloss*; on Roman-script items, they improve from 38–40 to 70–71. Exact inter-annotator response agreement similarly rises from 57% to 75% for Arabic-script input and from 53% to 74% for Roman-script input. On this subset, 0-shot Gemini-3.1-Pro reaches 76 Acc@1 with *Word+Gloss* on Arabic-script input, close to the stronger expert’s 77, and 78 on Roman-script input, above both experts’ 70–71. GPT-5.4 reaches 70 on both scripts. These results show that the value of glosses is not specific to LLM prompting: expert annotators also benefit substantially from semantic information. Their relatively low *Word*-only accuracy further indicates that recovering intended words from noisy learner transcriptions is intrinsically difficult, even for annotators with extensive Arabic NLP experience.

Model	Cue	Arabic script			Roman script		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	46.3	59.9	0.51	38.2	62.0	0.46
Gemini-3.1-Pro	Gloss	43.1	69.4	0.52	42.8	73.5	0.52
Gemini-3.1-Pro	Word+Gloss	76.0	93.1	0.83	82.3	93.7	0.87
GPT-5.4	Word	41.7	58.0	0.47	25.3	49.9	0.33
GPT-5.4	Gloss	36.9	57.6	0.43	33.8	57.9	0.43
GPT-5.4	Word+Gloss	67.1	80.4	0.72	70.4	84.4	0.75
Llama-4-Scout	Word	37.3	45.9	0.40	19.1	33.0	0.24
Llama-4-Scout	Gloss	22.3	36.5	0.27	16.8	34.6	0.22
Llama-4-Scout	Word+Gloss	49.2	66.2	0.55	42.0	59.1	0.48
Fanar-2	Word	34.7	43.6	0.38	15.9	27.4	0.20
Fanar-2	Gloss	12.2	21.4	0.16	12.2	26.4	0.18
Fanar-2	Word+Gloss	40.5	54.4	0.45	33.8	47.5	0.39

Table 2: 0-shot retrieval results under full normalization. Bold indicates the best result in each script and metric.

System	Arabic		Roman	
	Word	Word+Gloss	Word	Word+Gloss
Expert 1	43	77	40	71
Expert 2	43	67	38	70
Gemini-3.1-Pro	51	76	43	78
GPT-5.4	47	70	29	70
Llama-4-Scout	40	55	19	38
Fanar-2	38	40	16	30

Table 3: Acc@1 under full normalization on the human-baseline subset (100 items per script). W and W+G denote *Word* and *Word+Gloss*. Models use 0-shot prompting; Bold marks the best score in each column.

5 Analysis

We analyze five factors behind the aggregate results: cue interaction, normalized edit distance, learner proficiency, target frequency, and remaining hard cases. Full quantitative tables are provided in Appendix G.

Word and gloss interaction. In the primary 0-shot setting, *Word+Gloss* has the largest exclusive contribution at Acc@1. Averaged across the four models, 13.6% of Arabic-script items and 21.9% of Roman-script items are recovered only in the combined condition. These shares are substantially larger than the corresponding *Word*-only successes (6.0% and 3.1%) or *Gloss*-only successes (4.3% and 2.7%). The larger combined-only contribution for Roman-script input indicates that semantic information is particularly valuable when the form cue is more ambiguous. The smaller single-cue-only sets show that combining cues can occasionally suppress an otherwise correct prediction, but cue complementarity is much more common than

such interference. Table 15 reports the per-model results.

Edit distance. Performance decreases as normalized edit distance increases. For Arabic-script input, distance is computed against the true word; for Roman-script input, against the CAPHI representation of the true word (Habash et al., 2018). Averaged across the four models (Table 4), *Word* accuracy falls from 88.3 to 0.6 between the lowest and highest Arabic-script distance bins and from 59.6 to 7.7 for Roman-script input. *Word+Gloss* raises accuracy in the highest-distance bin to 24.7 and 44.2, respectively, showing that semantic information is especially useful when the observed form is far from the target.

Learner proficiency. Proficiency strongly affects transcription-only retrieval. Averaged across models, advanced learners exceed beginners by 34.7 Acc@1 points for Arabic-script *Word* input and by 13.1 points for Roman-script input. Under *Word+Gloss*, these gaps narrow to 12.5 and 7.1 points, respectively (Table 4). Glosses therefore partially compensate for the less recoverable transcriptions produced by beginners. Per-model Acc@1, Acc@10, and MRR results for both analyses appear in Appendix G.

Target frequency. To test whether common words are easier to recover, we use the CAMELBERT-MIX token counts (Inoue et al., 2021; Khalifa et al., 2021) as a frequency proxy. After applying the same full normalization as in the main evaluation, we place target words into four bins: <1k, [1k,10k), [10k,100k), and $\geq 100k$. For each 0-shot run, we calculate Acc@1 over all

Factor	Arabic		Roman	
	Word	Word+Gloss	Word	Word+Gloss
<i>Edit distance</i>				
[0, 0]	88.3	79.5	59.6	71.7
(0, 0.25]	25.2	54.9	34.2	63.4
(0.25, 0.5]	4.9	48.1	22.0	57.9
(0.5, 1]	0.6	24.7	7.7	44.2
<i>Proficiency</i>				
Beginner	32.9	55.6	22.7	56.1
Advanced	67.6	68.1	35.8	63.2

Table 4: Average 0-shot Acc@1 (%) across the four models by normalized edit distance and learner proficiency under full normalization. Beginner combines the no-experience, 0–1-year, and 1–2-year categories; Advanced combines the 2+-year, heritage-speaker, and native-speaker categories. Tables 16–19 give the per-model results.

Script	Frequency	Gloss	Word	Word+Gloss
Arabic	<1k	10.0	25.2	34.8
	[1k,10k)	21.3	37.8	57.1
	[10k,100k)	22.7	44.3	62.3
	≥100k	53.6	48.8	73.0
Roman	<1k	15.5	15.0	37.2
	[1k,10k)	16.8	23.5	53.7
	[10k,100k)	31.2	25.3	61.0
	≥100k	38.3	32.4	70.8

Table 5: Average 0-shot Acc@1 (%) across the four models by target frequency in the CAMELBERT-MIX list under full normalization. Table 20 gives the per-model results and bin sizes.

test instances in each bin, covering all 654 Arabic-script and 1,342 Roman-script test instances.

Overall, target words that occur more frequently in CAMELBERT-MIX are easier to recover (Table 5). Under *Word+Gloss*, average Acc@1 rises from 34.8 in the lowest-frequency bin to 73.0 in the highest-frequency bin for Arabic-script input, and from 37.2 to 70.8 for Roman-script input. Thus, glosses help throughout the frequency range but do not remove the difficulty of rare words. Although the averaged scores increase across the bins, the pattern is not equally smooth for every individual model. Moreover, CAMELBERT-MIX frequency is only an external measure of word commonness: rare words may also be less familiar, have less informative glosses, or occur with more misleading transcriptions.

Unsolved hard cases. We inspect cases where no model recovered the gold word. Errors are concentrated in rare targets: among 62 unresolved

cases, only 28 unique gold words appear, and ضَهْل

Dah.l ‘small quantity’ and قَمَح *qam~aH* ‘pay in installments; apportion’ alone account for 25 cases. The main pattern is a conflict between misleading transcriptions and rare or morphologically specific targets. For ضَهْل *Dah.l*, transcriptions often point to common words such as بَحْر *baH.r* ‘sea’ or نَهْر *nah.r* ‘river’, while the gloss is too generic.

For قَمَح *qam~aH*, models instead produce common semantic alternatives such as قَسَط *qas~aTa* ‘pay in installments’ or وَزَعَ *waz~aCa* ‘distribute’. Other errors show near-semantic recovery without exact-form recovery, such as مِيضَاء *miyDA* ‘ablu-

tion basin’ versus predicted variants such as مِيضَاءَ *miyDaÂaħ* ‘ablution basin’. Thus, the remaining failures mostly reflect rare lexical items, misleading transcriptions, and exact-form mismatches rather than complete semantic failure.

6 Conclusion and Future Work

We presented gloss-grounded intended-word recovery, a ranked recovery task in which models infer a standard word from an uncertain spelling, a meaning cue, or both. Across four LLMs and Arabic- and Roman-script ArabScribe transcriptions, combining form and meaning consistently produced the strongest results. Experts showed the same qualitative benefit from glosses, while the strongest model approached individual expert performance on sampled items. Our analyses further show that semantic support is especially valuable for distorted transcriptions and partially offsets proficiency-related differences. Nevertheless, rare targets, misleading forms, and exact-form confusions remain persistent challenges.

These findings are based on isolated MSA words and curated dictionary glosses, so future work should test more realistic settings, e.g. replacing clean glosses with free-form user descriptions, incorporating sentence and conversational context, extending the task to dialectal Arabic and other languages and writing systems, and evaluating interactive ranked suggestions with users. Comparisons with lexicon-constrained, retrieval-based, and hybrid approaches would also clarify which methods offer the best accuracy, efficiency, and controllability for practical assistive word-recovery systems.

Limitations

This study is limited to isolated MSA word lookup from ArabScribe learner transcriptions in Arabic and Roman script, so the results may not generalize to dialectal Arabic, sentence-level input, or real interactive dictionary use. Each item is presented without sentence or discourse context, although such context could substantially constrain the intended word in practice. Moreover, the *Word+Gloss* condition uses curated English dictionary glosses as a controlled upper bound on semantic support. Actual user descriptions are likely to be vaguer, messier, incomplete, or partially incorrect, so performance under this condition may overestimate what can be achieved with real user-provided meaning cues. We also evaluate prompting-only LLMs and exact-form retrieval, leaving fine-tuning, lexicon-constrained decoding, user studies, and model-update effects for future work.

Ethics Statement

This work uses existing learner transcriptions and lexical resources that contain no personal information. For the human baseline, we hired professional freelance annotators with extensive experience in Arabic annotation and compensated them fairly for their contributions.

The intended use of this research is assistive Arabic word recovery, not learner assessment, grading, or high-stakes decision-making. Because model outputs may be wrong, especially for rare words or misleading transcriptions, user-facing systems should present ranked candidates with definitions and uncertainty rather than a single authoritative correction.

We used AI writing assistance only under the “Assistance purely with the language of the paper” category defined in the ACL Policy on Publication Ethics.

Acknowledgments

We gratefully acknowledge the support and computational resources provided by the High Performance Computing Center at New York University Abu Dhabi, which supported the experiments in this work. This research is in part with support from Google.org and the Google Cloud Research Credits program for the Gemini Academic Program. We thank Mayar Nassar and Adel Wizani for their support in the human baseline.

References

- Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander, and Owen Rambow. 2014. Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Doha, Qatar.
- Kareem Darwish. 2014. Arabizi Detection and Conversion to Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 217–224, Doha, Qatar.
- Fadhl Eryani and Nizar Habash. 2021. [Automatic Romanization of Arabic bibliographic records](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 213–218, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- FANAR TEAM, Ummar Abbas, Mohammad Shahmeer Ahmad, Minhaj Ahmad, Abdulaziz Al-Homaid, Anas Al-Nuaimi, Enes Altinisik, Ehsaneddin Asgari, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Asim Ersoy, Masoomali Fatehkia, Mohammed Qusay Hashim, and 18 others. 2026. [Fanar 2.0: Arabic generative ai stack](#). *Preprint*, arXiv:2603.16397.
- David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.
- Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouni, Houada Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. [Unified guidelines and resources for Arabic dialect orthography](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.
- Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.
- Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. [Learning to understand phrases by embedding the dictionary](#). *Transactions of the Association for Computational Linguistics*, 4:17–30.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. [The interplay of variant, size, and task type in Arabic pre-trained language models](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Salam Khalifa, Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. [Camel Arabic Frequency Lists](#).

Meta. 2025. [The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation](#).

Fanchao Qi, Lei Zhang, Yanhui Yang, Zhiyuan Liu, and Maosong Sun. 2020. [WantWords: An open-source online reverse dictionary system](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–181, Online. Association for Computational Linguistics.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. [A unified model for Arabizi detection and transliteration using sequence-to-sequence models](#). In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 167–177, Barcelona, Spain (Online). Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 82 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1 others. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2019. [Multi-channel reverse dictionary model](#). *Preprint*, arXiv:1912.08441.

Lingliang Zhang, Nizar Habash, and Godfried Toussaint. 2017. [Robust dictionary lookup in multiple noisy orthographies](#). In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 119–129, Valencia, Spain. Association for Computational Linguistics.

A License

We list below the licenses of the data and tools used in this work, all of which are employed in accordance with their intended use.

- ArabScribe Dataset (Zhang et al., 2017): Non Commercial Use - NYUAD License.²
- Standard Arabic Morphological Analyzer (Graff et al., 2009): Linguistic Data Consortium User License Agreement.³
- CAMELBERT Frequency List (Khalifa et al., 2021): Creative Commons Attribution Share Alike 4.0
- *Llama-4-Scout-17B-16E-Instruct*: Llama 4 Community License Agreement.⁴
- *Fanar-2-27B-Instruct* (FANAR TEAM et al., 2026): Apache License 2.0.⁵

B Model Size and Computational Budget

All experiments use inference only; no model fine-tuning or training was performed.

GPT-5.4 and Gemini-3.1-Pro were accessed through provider APIs; their parameter counts and provider-side hardware are not publicly disclosed. For the open-weight models, Llama-4-Scout has 17B active parameters and 109B total parameters, and Fanar-2 has 27B parameters. The total closed-model API cost was approximately \$75. The open-weight models were run locally for inference using four NVIDIA A100 GPUs. The total local compute time was approximately 25 hours.

C Prompt Templates

We used three prompt templates corresponding to the three cue conditions: *Word*, *Gloss*, and *Word+Gloss*. For each test item, the relevant placeholders are filled with the learner transcription, the gloss, or both. In the few-shot setting, training examples are inserted before the test item using the same format as the target condition.

²<https://camel.abudhabi.nyu.edu/arabscribe/>

³<https://catalog.ldc.upenn.edu/LDC2010L01>

⁴<https://huggingface.co/meta-llama/>

Llama-4-Scout-17B-16E-Instruct/blob/main/LICENSE

⁵<https://huggingface.co/QCRI/>

Fanar-2-27B-Instruct

C.1 Word-only Prompt

Given the user transcription, which may be correct or wrong, predict the intended Arabic word.

Transcription: {transcription}

Respond with a JSON array of exactly 10 different possible Arabic words with diacritics, from most likely to least likely:

["word1", "word2", ..., "word10"]

C.2 Gloss-only Prompt

Given the word gloss, predict the intended Arabic word.

Gloss: {gloss}

Respond with a JSON array of exactly 10 different possible Arabic words with diacritics, from most likely to least likely:

["word1", "word2", ..., "word10"]

C.3 Word+Gloss Prompt

Given the user transcription, which may be correct or wrong, along with the word gloss, predict the intended Arabic word.

Transcription: {transcription}

Gloss: {gloss}

Respond with a JSON array of exactly 10 different possible Arabic words with diacritics, from most likely to least likely:

["word1", "word2", ..., "word10"]

D Dataset Statistics

Split	Arabic Script	Roman Script
Train	2,580	5,357
Test	654	1,342
Total	3,234	6,699

Table 6: Number of ArabScribe transcriptions by script and data split.

Arabic Experience Level	Users
No experience	26
0–1 years of Arabic education	45
1–2 years of Arabic education	14
2+ years of Arabic education	6
Heritage speaker	3
Native speaker	9
Total	103

Table 7: ArabScribe participant Arabic proficiency levels. Table taken from (Zhang et al., 2017).

E Human Baseline Details

Annotator experience. Expert 1 had three years of Arabic annotation experience focused on morphology and semantics, as well as experience in translation, subtitling, MSA and dialect localization, and teaching subtitling. Expert 2 had three years of experience in Arabic morphology, semantics, syntax, and lexicography annotation, together with two years of teaching Lebanese Arabic. Both therefore represent expert Arabic NLP annotators rather than typical dictionary users.

Procedure. Each expert received separate sheets for Arabic- and Roman-script items. They first inferred one intended word from each transcription alone. Only after completing all transcription-only sheets did they receive otherwise identical sheets containing English glosses, and they were instructed not to revisit their earlier answers. Clarifications specified that all targets were MSA, could include proper nouns, and could be rare or advanced. Responses and gold targets were evaluated after the same full normalization used for model Acc@1.

Outcome	Arabic		Roman	
	W	W+G	W	W+G
Agree	57	75	53	74
Both correct	36	62	34	60
Agree wrong	21	13	19	14
E1 only	7	15	6	11
E2 only	7	5	4	10
Wrong, differ	29	5	37	5

Table 8: Exact agreement decomposition for the two experts after full normalization. All values are percentages of the 100 items in the corresponding script and cue condition. W and W+G denote *Word* and *Word+Gloss*; “Wrong, differ” denotes cases where both responses are incorrect and differ from each other.

F Additional Retrieval Results

Model	Cue	0-shot			5-shot			20-shot		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	46.3	59.9	0.51	43.1	47.7	0.45	45.1	58.6	0.48
Gemini-3.1-Pro	Gloss	43.1	69.4	0.52	41.1	68.5	0.51	43.1	75.4	0.54
Gemini-3.1-Pro	Word+Gloss	76.0	93.1	0.83	76.3	87.9	0.81	77.5	93.1	0.83
GPT-5.4	Word	41.7	58.0	0.47	44.7	58.1	0.49	44.5	57.8	0.49
GPT-5.4	Gloss	36.9	57.6	0.43	37.3	59.5	0.44	37.3	62.5	0.45
GPT-5.4	Word+Gloss	67.1	80.4	0.72	72.0	86.2	0.77	69.4	85.2	0.75
Llama-4-Scout	Word	37.3	45.9	0.40	38.2	47.7	0.41	36.9	45.9	0.40
Llama-4-Scout	Gloss	22.3	36.5	0.27	20.2	36.1	0.25	20.5	38.2	0.26
Llama-4-Scout	Word+Gloss	49.2	66.2	0.55	48.8	61.8	0.53	48.0	61.3	0.52
Fanar-2	Word	34.7	43.6	0.38	38.4	43.6	0.40	35.9	44.5	0.39
Fanar-2	Gloss	12.2	21.4	0.16	15.1	34.1	0.22	21.6	28.3	0.24
Fanar-2	Word+Gloss	40.5	54.4	0.45	43.9	55.4	0.48	45.3	56.4	0.49

Table 9: **Arabic-script** retrieval results under **full normalization** for 0, 5, and 20 shots. Bold indicates the best result in each column.

Model	Cue	0-shot			5-shot			20-shot		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	38.2	62.0	0.46	41.4	65.3	0.49	41.5	68.0	0.50
Gemini-3.1-Pro	Gloss	42.8	73.5	0.52	47.7	73.5	0.57	48.7	72.6	0.58
Gemini-3.1-Pro	Word+Gloss	82.3	93.7	0.87	85.8	95.6	0.89	84.7	95.3	0.89
GPT-5.4	Word	25.3	49.9	0.33	31.6	55.0	0.39	32.0	57.5	0.41
GPT-5.4	Gloss	33.8	57.9	0.43	37.7	65.9	0.47	38.0	65.6	0.46
GPT-5.4	Word+Gloss	70.4	84.4	0.75	75.0	88.2	0.80	75.0	89.5	0.80
Llama-4-Scout	Word	19.1	33.0	0.24	24.1	34.4	0.28	24.7	35.6	0.28
Llama-4-Scout	Gloss	16.8	34.6	0.22	18.6	38.3	0.25	17.8	35.5	0.24
Llama-4-Scout	Word+Gloss	42.0	59.1	0.48	42.3	57.4	0.48	43.9	58.6	0.49
Fanar-2	Word	15.9	27.4	0.20	16.3	28.6	0.21	16.9	29.1	0.21
Fanar-2	Gloss	12.2	26.4	0.18	22.4	36.3	0.27	19.6	37.3	0.26
Fanar-2	Word+Gloss	33.8	47.5	0.39	36.0	51.0	0.42	36.3	50.8	0.42

Table 10: **Roman-script** retrieval results under **full normalization** for 0, 5, and 20 shots. Bold indicates the best result in each column.

Model	Cue	0-shot			5-shot			20-shot		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	12.1	20.5	0.15	16.1	25.8	0.20	21.7	37.8	0.27
Gemini-3.1-Pro	Gloss	20.8	31.7	0.25	27.2	39.9	0.32	31.7	53.2	0.40
Gemini-3.1-Pro	Word+Gloss	27.7	45.6	0.35	39.9	62.2	0.48	49.4	74.6	0.59
GPT-5.4	Word	13.6	25.1	0.17	22.8	37.3	0.27	24.5	41.6	0.30
GPT-5.4	Gloss	15.4	24.0	0.19	22.5	40.5	0.29	23.9	41.1	0.30
GPT-5.4	Word+Gloss	34.9	45.6	0.39	43.4	61.3	0.50	48.3	67.7	0.55
Llama-4-Scout	Word	6.1	13.3	0.08	15.8	29.7	0.20	15.4	28.1	0.20
Llama-4-Scout	Gloss	2.3	4.1	0.03	7.2	15.6	0.09	9.3	19.0	0.12
Llama-4-Scout	Word+Gloss	7.7	20.8	0.11	21.6	35.0	0.26	23.4	38.7	0.29
Fanar-2	Word	3.4	7.5	0.05	9.0	14.7	0.11	11.0	18.5	0.13
Fanar-2	Gloss	0.0	1.8	0.01	5.2	12.1	0.07	6.7	9.6	0.08
Fanar-2	Word+Gloss	5.2	11.5	0.07	13.6	21.7	0.16	17.3	24.6	0.20

Table 11: **Arabic-script** retrieval results under **exact matching**. Bold indicates the best result in each column.

Model	Cue	0-shot			5-shot			20-shot		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	13.9	24.4	0.18	17.4	30.0	0.22	25.7	47.4	0.33
Gemini-3.1-Pro	Gloss	12.0	29.5	0.18	19.0	35.5	0.25	38.1	58.9	0.46
Gemini-3.1-Pro	Word+Gloss	30.9	46.1	0.37	36.1	54.8	0.43	51.4	76.4	0.62
GPT-5.4	Word	8.9	14.2	0.10	15.2	20.5	0.16	22.4	28.8	0.24
GPT-5.4	Gloss	6.9	13.0	0.08	13.9	22.0	0.16	26.1	36.0	0.28
GPT-5.4	Word+Gloss	26.5	32.4	0.28	33.4	40.5	0.35	59.4	66.0	0.61
Llama-4-Scout	Word	5.7	11.5	0.07	13.0	18.3	0.15	13.2	19.2	0.15
Llama-4-Scout	Gloss	0.9	3.0	0.01	4.8	13.9	0.08	5.0	17.1	0.10
Llama-4-Scout	Word+Gloss	9.9	21.7	0.14	20.5	31.2	0.24	23.0	33.9	0.27
Fanar-2	Word	1.4	4.5	0.02	2.9	7.6	0.04	4.4	10.0	0.06
Fanar-2	Gloss	1.9	6.9	0.04	4.8	13.9	0.08	4.9	14.1	0.08
Fanar-2	Word+Gloss	6.3	10.4	0.08	11.1	18.9	0.14	12.4	22.1	0.16

Table 12: **Roman-script** retrieval results under **exact matching**. Bold indicates the best result in each column.

Model	Cue	0-shot			5-shot			20-shot		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	45.1	59.3	0.50	42.2	47.7	0.44	43.6	58.6	0.47
Gemini-3.1-Pro	Gloss	42.1	69.4	0.52	40.1	68.5	0.51	42.1	75.4	0.53
Gemini-3.1-Pro	Word+Gloss	75.7	92.8	0.82	75.5	87.8	0.80	76.8	92.8	0.83
GPT-5.4	Word	40.5	57.3	0.46	42.7	57.3	0.48	42.8	56.9	0.48
GPT-5.4	Gloss	35.8	57.7	0.43	37.3	59.5	0.44	36.2	62.5	0.45
GPT-5.4	Word+Gloss	66.4	80.3	0.71	71.4	85.6	0.76	68.4	84.9	0.74
Llama-4-Scout	Word	35.8	45.1	0.39	37.3	47.3	0.41	36.2	45.7	0.39
Llama-4-Scout	Gloss	22.3	36.5	0.27	20.2	36.1	0.25	20.5	38.2	0.26
Llama-4-Scout	Word+Gloss	49.1	65.6	0.54	47.4	61.5	0.52	47.7	61.2	0.52
Fanar-2	Word	32.7	41.9	0.36	37.5	42.7	0.39	34.4	43.6	0.38
Fanar-2	Gloss	12.2	21.4	0.16	13.3	32.3	0.20	20.5	27.2	0.23
Fanar-2	Word+Gloss	39.6	53.4	0.44	43.1	54.6	0.47	43.6	55.2	0.48

Table 13: **Arabic-script** retrieval results under **undiacritized matching**. Bold indicates the best result in each column.

Model	Cue	0-shot			5-shot			20-shot		
		Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Gemini-3.1-Pro	Word	37.8	61.6	0.46	41.1	64.9	0.49	41.1	67.7	0.50
Gemini-3.1-Pro	Gloss	41.7	72.5	0.51	46.7	72.5	0.56	47.6	72.6	0.58
Gemini-3.1-Pro	Word+Gloss	81.5	93.5	0.86	85.0	95.3	0.89	83.8	95.3	0.88
GPT-5.4	Word	24.7	33.8	0.27	30.9	41.0	0.33	31.7	42.2	0.34
GPT-5.4	Gloss	32.7	46.8	0.36	36.7	51.0	0.40	37.0	48.9	0.40
GPT-5.4	Word+Gloss	69.3	77.1	0.71	74.1	81.0	0.76	74.3	82.6	0.76
Llama-4-Scout	Word	17.7	31.3	0.22	23.9	34.1	0.27	24.4	35.2	0.28
Llama-4-Scout	Gloss	15.7	33.5	0.21	17.6	38.3	0.24	16.8	35.5	0.23
Llama-4-Scout	Word+Gloss	40.4	57.5	0.47	41.3	56.1	0.47	43.0	57.7	0.48
Fanar-2	Word	14.9	26.3	0.19	15.2	27.2	0.19	16.1	28.2	0.20
Fanar-2	Gloss	11.1	25.4	0.17	21.4	36.3	0.27	18.6	35.3	0.25
Fanar-2	Word+Gloss	32.9	46.5	0.38	34.9	49.6	0.40	34.6	49.9	0.40

Table 14: **Roman-script** retrieval results under **undiacritized matching**. Bold indicates the best result in each column.

G Analysis Tables

Model	Arabic			Roman		
	Word only	Gloss only	Word+Gloss only	Word only	Gloss only	Word+Gloss only
Gemini-3.1-Pro	2.6	6.6	16.8	1.3	3.1	23.3
GPT-5.4	6.6	2.8	16.8	3.2	1.5	28.0
Llama-4-Scout	4.7	4.7	9.3	3.3	3.0	18.2
Fanar-2	9.9	3.1	11.3	4.7	3.3	18.1

Table 15: Exclusive 0-shot Acc@1 successes (%) under full normalization. Each cell counts items for which only the named cue condition retrieves the target at rank 1 and the other two conditions do not; the three columns within each script are therefore mutually exclusive.

Cue	Edit distance bin	Gemini-3.1-Pro			GPT-5.4			Llama-4-Scout			Fanar-2		
		A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR
Word	[0, 0]	96.3	98.8	0.97	84.8	98.8	0.91	89.3	98.0	0.93	82.8	93.9	0.87
Word	(0, 0.25]	33.1	65.5	0.44	35.1	58.8	0.42	16.9	29.7	0.21	15.5	28.4	0.19
Word	(0.25, 0.5]	10.3	29.3	0.15	7.5	28.2	0.14	0.6	9.8	0.03	1.1	7.5	0.03
Word	(0.5, 1]	1.1	3.4	0.01	1.1	2.3	0.01	0.0	0.0	0.00	0.0	0.0	0.00
Word+Gloss	[0, 0]	92.2	99.2	0.95	77.9	90.2	0.82	80.7	94.7	0.85	67.2	85.7	0.73
Word+Gloss	(0, 0.25]	70.9	96.6	0.82	70.3	83.1	0.75	43.2	64.2	0.49	35.1	50.7	0.40
Word+Gloss	(0.25, 0.5]	73.6	93.1	0.80	65.5	80.5	0.70	32.2	51.7	0.39	21.3	33.3	0.25
Word+Gloss	(0.5, 1]	44.3	70.5	0.54	35.2	48.9	0.40	5.7	19.3	0.10	13.6	15.9	0.14

Table 16: **Arabic-script** 0-shot performance by normalized edit-distance bin under full normalization. Rows are grouped by cue setup, and columns report A@1, A@10, and MRR for each model. A@1 and A@10 abbreviate Acc@1 and Acc@10.

Cue	Edit distance bin	Gemini-3.1-Pro			GPT-5.4			Llama-4-Scout			Fanar-2		
		A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR
Word	[0, 0]	80.7	99.1	0.89	63.2	93.0	0.74	52.6	78.9	0.61	42.1	65.8	0.51
Word	(0, 0.25]	51.7	82.7	0.62	31.3	65.3	0.42	30.0	45.7	0.35	24.0	42.7	0.30
Word	(0.25, 0.5]	35.5	62.2	0.44	23.5	49.9	0.32	15.6	31.3	0.21	13.6	24.3	0.17
Word	(0.5, 1]	15.1	28.0	0.19	9.3	19.3	0.12	3.5	7.4	0.05	2.9	4.5	0.03
Word+Gloss	[0, 0]	93.9	100.0	0.97	76.3	91.2	0.83	67.5	78.1	0.72	49.1	72.8	0.57
Word+Gloss	(0, 0.25]	89.3	99.3	0.93	77.7	90.0	0.81	45.0	66.3	0.52	41.7	57.0	0.47
Word+Gloss	(0.25, 0.5]	83.6	94.7	0.88	73.6	86.7	0.78	41.7	59.2	0.49	32.9	46.7	0.38
Word+Gloss	(0.5, 1]	68.8	83.9	0.74	55.0	71.7	0.61	30.5	45.0	0.35	22.5	30.5	0.26

Table 17: **Roman-script** 0-shot performance by normalized edit-distance bin under full normalization. Rows are grouped by cue setup, and columns report A@1, A@10, and MRR for each model. A@1 and A@10 abbreviate Acc@1 and Acc@10.

Cue	Level	Gemini-3.1-Pro			GPT-5.4			Llama-4-Scout			Fanar-2		
		A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR
Word	Beginner	39.5	54.1	0.44	35.1	52.2	0.41	29.5	38.0	0.32	27.4	35.8	0.30
Word	Advanced	72.6	82.2	0.76	67.4	80.0	0.73	67.4	76.3	0.70	63.0	72.6	0.66
Word+Gloss	Beginner	74.6	92.3	0.81	66.3	79.4	0.71	44.7	63.4	0.51	37.0	50.7	0.42
Word+Gloss	Advanced	81.5	96.3	0.87	70.4	84.4	0.75	66.7	77.0	0.70	54.1	68.9	0.59

Table 18: **Arabic-script** 0-shot performance by learner proficiency group under full normalization. Beginner combines the no-experience, 0–1-year, and 1–2-year categories; Advanced combines the 2+-year, heritage-speaker, and native-speaker categories. Rows are grouped by cue setup, and columns report A@1, A@10, and MRR for each model. A@1 and A@10 abbreviate Acc@1 and Acc@10.

Cue	Level	Gemini-3.1-Pro			GPT-5.4			Llama-4-Scout			Fanar-2		
		A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR	A@1	A@10	MRR
Word	Beginner	35.4	58.7	0.43	23.2	47.7	0.31	17.9	31.2	0.22	14.4	25.3	0.18
Word	Advanced	54.8	81.2	0.64	37.6	62.9	0.46	26.4	43.7	0.32	24.4	39.1	0.29
Word+Gloss	Beginner	81.6	93.4	0.86	69.0	83.6	0.74	41.4	58.5	0.48	32.5	45.9	0.37
Word+Gloss	Advanced	86.8	95.4	0.90	78.7	88.8	0.83	45.7	62.4	0.52	41.6	56.9	0.47

Table 19: **Roman-script** 0-shot performance by learner proficiency group under full normalization. Beginner combines the no-experience, 0–1-year, and 1–2-year categories; Advanced combines the 2+-year, heritage-speaker, and native-speaker categories. Rows are grouped by cue setup, and columns report A@1, A@10, and MRR for each model. A@1 and A@10 abbreviate Acc@1 and Acc@10.

Script	Frequency	Types	Items	Gemini-3.1-Pro			GPT-5.4			Llama-4-Scout			Fanar-2		
				G	W	W+G	G	W	W+G	G	W	W+G	G	W	W+G
Arabic	< 1k	23	148	21.6	32.4	53.4	9.5	25.0	38.5	5.4	23.0	26.4	3.4	20.3	20.9
Arabic	1k–10k	20	127	29.9	47.2	79.5	29.9	36.2	64.6	20.5	35.4	47.2	4.7	32.3	37.0
Arabic	10k–100k	28	187	35.3	50.8	82.9	32.1	45.5	75.9	19.3	40.1	50.3	4.3	40.6	40.1
Arabic	≥ 100k	29	192	76.0	52.1	84.4	67.2	54.7	82.3	39.6	46.9	67.2	31.8	41.7	58.3
Roman	< 1k	18	245	33.5	29.0	66.5	22.9	15.5	42.4	5.7	11.0	23.7	0.0	4.5	16.3
Roman	1k–10k	25	340	31.5	40.3	85.0	27.6	20.3	66.5	8.2	17.6	34.4	0.0	15.9	28.8
Roman	10k–100k	33	439	48.7	37.6	84.1	35.5	28.2	74.9	17.8	18.2	48.3	22.6	17.1	36.9
Roman	≥ 100k	24	318	53.8	44.0	89.3	46.2	34.3	89.9	33.0	28.3	55.7	20.1	23.0	48.4

Table 20: 0-shot Acc@1 (%) under full normalization by target frequency in the CAMELBERT-MIX list. G, W, and W+G denote *Gloss*, *Word*, and *Word+Gloss*. Types and Items give the number of target words and test transcriptions in each bin.