Zero-shot ASR Error Correction for Low-resource Spoken Question Answering

Sherwynn Clarence Angeles[1], *Bagus Tris Atmaja*[2], *Nathaniel Oco*[1], *Sakriani Sakti*[2]

[1] De La Salle University, Manila, Philippines
[2] Nara Institute of Science and Technology, Ikoma, Nara, Japan

sherwynn_clarence_angeles@dlsu.edu.ph, nathaniel.oco@dlsu.edu.ph, bagus.tris@naist.ac.jp, ssakti@is.naist.jp

Abstract

Cascaded spoken question answering (SQA) remains practical for low-resource languages where end-to-end systems remain limited, because it can reuse stronger pretrained ASR and text QA models. ASR errors, however, can propagate to QA and reduce accuracy. This study evaluates prompting-based LLM ASR error correction as post-processing for Filipino multiple-choice SQA. On Filipino 2M-BELEBELE, we compare Gemma-3 models and Gemini-2.5-Flash for correction and SQA in clean and noise-augmented settings. Correction improves transcription quality and generally improves QA accuracy, but does not fully recover performance with ground-truth transcripts. Gains remain consistent under noise, although correction adds about a 2.5× latency increase over direct QA on uncorrected ASR output. E2E speech models such as Gemma-3n-E4B perform worse than cascaded systems. Overall, LLM-based ASR error correction is a promising zero-training strategy for improving cascaded SQA robustness in low-resource settings.

Index Terms: automatic speech recognition, error correction, spoken question answering, Filipino language

1. Introduction

Spoken question answering (SQA) requires answering a question given a spoken passage and spoken question [1]. SQA extends text-based question answering (QA) to settings where users interact with systems through speech, enabling more natural interactions and applications [2], such as digital assistants or voice interfaces.

Recent work has increasingly focused on multimodal and end-to-end (E2E) SQA systems that directly process speech inputs to generate answers [3]. However, these approaches typically require significant computational resources to train and deploy [4]. For low-resource languages, performance and robustness to noise remain open problems [5]. This study therefore focuses on cascaded SQA systems, which first transcribe speech inputs using automatic speech recognition (ASR) and then use a text-based QA model. Cascaded systems remain attractive in this setting because they can reuse pretrained ASR and text QA models, such as Whisper [6] and Gemma-3 [7], and can work with more general speech recognition and text-based QA datasets, when task-specific SQA training data is not available.

A major challenge with cascaded SQA systems is that ASR errors can propagate to the downstream QA model, which can significantly degrade performance [1, 8]. Figure 1 shows how correcting an ASR error can change the downstream answer.

Inspired by recent work on LLM-based ASR error correction as a post-processing step to improve transcription [9], this study explores whether prompting-based correction can improve cascaded SQA without training a specialized correction model. This is particularly relevant where task-specific training data are scarce. We make the following contributions and research objectives:

- We provide an in-depth evaluation of zero-shot ASR error correction and its effects on transcription quality and downstream QA accuracy.
- We evaluate robustness under noise-augmented speech and compare cascaded SQA with E2E models.
- We analyze how ASR errors and correction relate to QA performance, and also measure its effect on additional inference latency.

Without correction

Passage: The Great Pyramid of Giza ...
Question: What is the goal of the Great Pyramid?
Choices:
A. To honor the Thero
...
D. To place tombs for the deceased.

QA Answer:
D. To place tombs for the deceased.

With correction

Passage: (no change)
Question: (no change)
Choices:
A. To honor the ~~Thero~~ Pharaoh
(no change)

QA Answer:
A. To honor the Pharaoh

Figure 1: *Motivating example of ASR error correction improving downstream QA. 'A', the correct answer, was chosen after correcting the input text from 'Thero' to 'Pharaoh'.*

2. Related Work

Cascaded systems have simple and modular architectures that make them easy to develop. However, work has shown that ASR errors significantly degrade the performance of text-based QA models [1, 8]. Prior work has improved ASR-error robustness using subword or phonetic representations, including augmenting phoneme, character, syllable, and pinyin-based embeddings to word embeddings [1, 8]. Another line of work trains QA models using real ASR hypotheses and text-to-speech and ASR pipelines for synthetic data generation [1, 10]. However, these approaches require retraining or adaptation of the QA model, which requires domain-specific data that may not be available for low-resource languages.

To bypass the problem of cascading ASR errors, recent work has studied E2E SQA systems that directly process speech inputs to generate answers [3]. These include specialized SQA

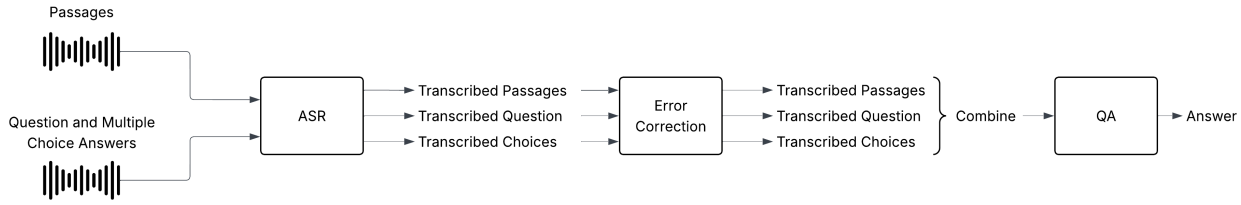


Figure 2: *Error correction pipeline.*

models like [11] that train BERT to process speech inputs. Speech large language models have also been evaluated on SQA, such as [12, 13]. However, aside from being difficult to deploy [4], these models also do not always outperform cascaded systems [14, 13, 15]. This motivates revisiting cascaded systems, especially for low-resource languages where training E2E SQA models may be impractical.

Meanwhile, LLM-based ASR error correction (ASR-EC) has been explored as a post-processing step to improve ASR transcription quality. [9] evaluated zero-shot prompting using ASR n-best lists and reported relative average WER reductions of 2.6% to 25.2% with Whisper and Transducer respectively. Error correction has also been applied to SQA. [15] proposes a spoken language understanding evaluation framework containing ASR-EC plus LLM pipeline as a method, finding that cascaded ASR-based systems outperform similarly sized multimodal models, but larger E2E multimodal models achieve stronger downstream QA performance. [10] studied different data augmentation strategies for ASR error robustness and includes a custom-trained ASR-EC model as an experimental setting; [16] fine-tune an LLM to perform both error correction and medical SQA using knowledge graphs of medical terms, while [17] trains a multitask model that jointly performs error correction and slot filling.

Prior ASR-EC studies have included correction as one strategy in broader SQA evaluations [15, 10], without making it the main focus, or used specialized correction models whose training requirements can limit their use in low-resource languages [10, 16]. What remains missing is an in-depth evaluation of zero-shot, prompting-based ASR error correction as a practical approach to low-resource SQA, where task-specific training data are scarce and robustness and latency matter for deployment. This study addresses that gap for Filipino by examining transcription quality, downstream QA accuracy, noise robustness, and latency together, supported by comparisons with ground-truth text and E2E systems and analysis of how ASR errors and correction relate to QA performance.

3. Approach

Our pipeline has three stages. First, the passage, question, and answer-option audio are transcribed into separate ASR n-best lists. Second, a prompting-based error-correction model produces one corrected transcription for each list. Third, the corrected fields are combined and passed to a text-based QA model, which selects an answer option. Fig. 2 summarizes this process.

3.1. Task Definition

We focus on machine reading comprehension style multiple-choice SQA, where the system is given a passage, a question, and a set of multiple choice answer options in speech form, and must select the correct answer option. The passage, question, and answer options remain separate throughout ASR transcription and error correction, and only the resulting corrected text fields are combined into the QA input. This is to keep the correction component generic and not specific to SQA.

3.2. Prompting-based ASR Error Correction

The LLM model is prompted to process each ASR n-best hypothesis and generate the corrected transcription considering the agreement between hypotheses, grammar, and vocabulary, such that the correction is phonetically similar to the ASR hypotheses. We also implement reasoning where the model first produces a brief analysis before generating the corrected transcription. Both are returned through structured and constrained output implemented with JSON for deterministic parsing.

3.3. Cascaded Question Answering

After correction, the passage, question, and answer options are combined into a single structured QA prompt. The model is prompted to select the most appropriate answer option. Reasoning is implemented in the same way, where the model first produces a brief analysis, followed by the selected answer, through structured and constrained output implemented with JSON. Our main setting reuses the same LLM for ASR-EC and QA, adding error correction as a single additional prompting stage without a separately trained correction model. The correction and QA roles remain separate prompts so that the correction component remains generic. Fig. 11 shows the ASR error correction and QA prompts with reasoning enabled.

4. Experimental Setup

4.1. Dataset and Models

We use the 2M-BELEBELE dataset [18], a highly multilingual SQA dataset with 4 answer options per question. The dataset is a read-speech extension of the BELEBELE benchmark [19], a reading comprehension dataset derived from Flores-200 passages [20]. We focus on the Filipino test set, with 801 complete SQA examples (99 are removed due to missing audio files). When there are multiple audio files for a text item, the first one in the list is chosen to form a single QA instance, as it is the main unit of evaluation. The first recording is chosen as a heuristic as speaker IDs are not provided.

The ASR model is limited to just `Whisper-large-v3`. A variety of LLM sizes is tested, using Gemma-3 [7] as the main LLM for evaluation, because it performs strongly on Filipino benchmarks [21, 22] at the time of implementation[1]. The 270M, 4B, and 27B instruction-tuned Gemma-3 variants are used, to evaluate performance on very resource-constrained set-

[1] https://leaderboard.sea-lion.ai/detailed/TL

tings (270M), and on medium-sized to larger models (4B and 27B), referred to as Gemma 270M, Gemma 4B, and Gemma 27B respectively. We additionally evaluate Gemini-2.5-Flash as a reference for commercial LLMs, and E2E audio LLMs as a baseline for cascaded systems, which output the answer option from the audio input in a single step. Reasoning is implemented in the same way, with a brief analysis before the answer, through structured and constrained output implemented with JSON.

Additionally, we fine-tune an encoder-only model for question answering, using MMBERT-base [23] for its low latency and computational efficiency. Since we are unable to find a large-scale Filipino multiple choice QA dataset, we train the model on a translated version of the 'all' subset of the RACE dataset [24], since its multiple-choice format closely mirrors BELEBELE. Despite potential domain mismatch and translation errors, this acts as a reference for evaluating the impact of ASR error correction on a more efficient encoder-only QA model.

Translation used Gemma-3-27B with the entire QA sample and relied on few-shot prompting, instructing the LLM to preserve the original meaning, maintain answerability, use natural phrasing with transliterated technical terms, and output structured JSON. We fine-tune MMBERT-base with an effective batch size of 8, 5 epochs, a 10% learning rate warmup, 0.01 weight decay, and a maximum token length of 2048, achieving an accuracy of 69.89% on the test set. We additionally test performance on XLM-RoBERTa-large [25], but it fails to converge with the same dataset.

We use vLLM[2] for the LLM latency measurements, and a single NVIDIA A100 GPU with a batch size of 1 for all experiments. All LLM-prompting configurations are set to use greedy decoding. N-best hypotheses are constructed with beam search with a beam-width of 100. The MMBERT-base QA model uses the Hugging Face transformers model implementation and the same GPU for inference.

4.2. Experimental Variables

A range of LLM-prompting-based ASR error correction inference setups are studied. For the Gemma models, we vary the number of n-best hypothesis examples $(1, 3, 5)$ to evaluate how much adding information improves ASR-EC performance, and enable or disable reasoning to evaluate how much the added tokens for reasoning affect latency and correction performance. Gemini-2.5-Flash error correction is evaluated only with 5-best hypotheses and reasoning enabled to limit API cost.

To evaluate the impact of ASR error correction, we test on the product of all Gemma ASR-EC configurations, QA models, and reasoning enabled or disabled. Comparisons using Gemini-2.5-Flash error correction use only its fixed 5-best configuration with reasoning enabled.

Table 1 summarizes the experimental variables for each phase of the evaluation.

4.3. Noise Simulation

To evaluate robustness of the error correction pipeline under noise, we use the MUSAN noise dataset [26] to augment the original dataset with noise. A random background noise sample is selected for each utterance of 2M-BELEBELE, which is repeated then truncated (tiled) such that the noise is the same length as the original audio. The noise sample is then mixed with the original audio at signal-to-noise ratios (SNR) of 10 dB and 0 dB, to simulate moderate and high noise conditions. To simulate the process of using a weaker ASR model and analyze robustness to ASR errors, we also take the 50th-best ASR hypothesis with no noise augmentation. Here, the n-best list starts from the 50th rank (i.e., 50th, 51st, 52nd). Although this is an imperfect simulation of noisy or weak-ASR conditions, it gives some insight on how these pipelines would perform under noisier real-world conditions.

Table 1: *Experimental variables.*

Phase	Variable	Values
ASR-EC	EC Model	Gemma-3 (270M[3], 4B[4], 27B[5]), Gemini-2.5-Flash[6]
ASR-EC	Configuration	n-best: 1, 3, 5; reasoning: enabled, disabled. (except Gemini: 5-best and reasoning enabled only)
Cascade-QA	Input	All ASR-EC configurations, no correction
Cascade-QA	QA Model	Gemma-3 (270M, 4B, 27B), Gemini-2.5-Flash, MMBERT
Cascade-QA	Configuration	reasoning: enabled, disabled
End-to-End	Audio LLM	Gemma-3n-E4B (Gemma E4B), Gemini-2.5-Flash
Cascade & EC	Noise	Rank 50, SNR 10 dB, SNR 0 dB, no noise

4.4. Evaluation Metrics

ASR error correction performance is evaluated using word error rate (WER) and word error rate reduction (WERR), defined as:

$$WER = \frac{S + D + I}{N}$$

$$WERR = \frac{WER_{ASR} - WER_{corrected}}{WER_{ASR}}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcription. WER_{ASR} is the WER of the original ASR output, and $WER_{corrected}$ is the WER of the corrected ASR output.

QA performance is evaluated using accuracy or the percentage of correctly answered questions out of the total questions. We additionally evaluate latency of both ASR error correction and QA inference, measured as the average time taken to perform each step, from 100 randomly sampled instances from the dataset.

5. Results and Discussion

5.1. ASR Error Correction Performance

We first discuss error correction performance on the original unaugmented dataset. Table 2 shows the WER of Whisper on the original dataset (16.56%), and the best configuration (n-best count and reasoning) for each Gemma LLM variant, along with the added latency of performing error correction. The Gemini-2.5-Flash row uses only the fixed 5-best configuration with reasoning enabled. Among open models, Gemma-3-27B has the best WERR of 16.72%, but has the highest added latency of 3.11s. Gemma-3-4B achieves a similar corrected WER and also a much lower added latency of 0.34s, but also a lower WERR of 10.73%. Using Gemma-3-270M, however, degrades performance, with a WERR of -32.35%, indicating that the model

[2]https://github.com/vllm-project/vllm

[3]https://huggingface.co/google/gemma-3-270m-it

[4]https://huggingface.co/google/gemma-3-4b-it

[5]https://huggingface.co/google/gemma-3-27b-it

[6]API version on June 15, 2026

is not strong enough to perform error correction. The commercial reference Gemini-2.5-Flash achieves the best WERR of 21.00%, indicating room for improvements for open-weight models. Fig. 3 and Fig. 4 illustrate examples where WER degrades due to model errors and improves through contextual information respectively. Here, and in succeeding examples, strikethrough marks deletions, underline marks insertions, red indicates errors, and green indicates correct text. Aside from Gemma 270M, all Gemma models perform best with the highest number of n-best examples (5), suggesting that added information benefits correction, but enabling reasoning lets open-weight models overthink and over-correct the ASR output especially in grammar or some omitted words.

We additionally use a regression model to analyze the contribution of each feature in the experimental grid (LLM model, n-best example count, reasoning enabled or disabled). Features are normalized from 0 to 1. The correction model is a binary classifier that predicts whether the WER of each individual utterance improved (decreased), while the latency model is a regression model that predicts the added latency (also scaled from 0 to 1). The regression coefficients are shown in Table 3. For correction, the n-best example count has the strongest positive effect, while it is the LLM model for the latency.

Table 2: *ASR error correction performance and latency across different models. Brackets show 95% percentile bootstrap confidence intervals for WER, using 10,000 resamples.*

Model	N-best examples	Reasoning Allowed	WER (WERR) [95% CI]	Add. Latency
Whisper-large-v3 (uncorrected baseline)	N/A	N/A	16.56% (N/A) [16.25%-16.87%]	N/A
Gemma-3-270M-it	1	False	21.91% (-32.35%) [20.80%-23.39%]	+0.0558s
Gemma-3-4B-it	5	False	14.78% (10.73%) [14.47%-15.09%]	+0.3395s
Gemma-3-27B-it	5	False	13.79% (16.72%) [13.49%-14.09%]	+3.1121s
Gemini-2.5-Flash	5	True	13.08% (21.00%) [12.79%-13.37%]	N/A

Table 3: *Regression coefficients for error correction.*

Feature	Correction Coefficient	Latency Coefficient
LLM model (4B as 0, 27B as 1)	0.0966	0.5756
N-best example count	0.3356	0.1444
Reasoning Allowed	-0.0809	0.2604

Next, Table 4 shows the WER and WERR of applying error correction on noise-augmented and weaker ASR hypothesis conditions. WER values for each model listed are from the best configurations, all having an n-best count of 5 and reasoning disabled. Gemma-3-270M is not included since it degrades performance on the original dataset. WER increases significantly with noise augmentation. When noise is equal to the signal (0 dB), WER increases to 23.70% from 16.56%. When the 50th-best ASR hypothesis is used, WER increases to 24.70%. Error correction remained effective across noise augmentation with similar WERRs per model, where Gemma-3-27B also has a stronger effect than Gemma-3-4B. Error correction is particularly effective when the 50th-best ASR hypothesis is used. These findings suggest that LLM-based error correction

Ground Truth: Isa sa mga pinakakaraniwang pamamaraang ginagamit upang ipakita ang kahalagahan ng sosyalisasyon ay pagbatayan ang ilang hindi kanais-nais na kalagayan ng mga batang hindi naturuan ng mabuting asal ng matatanda habang lumalaki sila sa pamamagitan ng pagpapabaya, kasawiang-palad, o sinadyang pang-aabuso.

Corrected: ... kalagayan ng mga ~~bata. Hindi~~ batang hindi naturuan ng mabuting asal ng matatanda habang lumalaki, Hindi naturuan ng mabuting asal ng matatanda habang lumalaki sila sila sa pamamagitan ng pagpapabalik.

Translation: One of the most common ways to show the importance of socialization is to cite the unfortunate cases of children who, due to neglect, misfortune, or deliberate abuse, were not taught proper behavior by adults as they grew up.

Figure 3: *Example of a failed correction due to model error. In reasoning traces, the LLM is able to infer that the correct text should contain 'pabaya', but it is not included, and the last phrase of the ground truth is omitted, while the last phrase of the ASR hypothesis is repeated.*

is useful for improving transcription quality under noise. However, higher WER improvements do not necessarily translate to downstream QA performance improvements, which we discuss in the next section.

Table 4: *WER and WERR under simulated noise before and after correction. Brackets show 95% percentile bootstrap confidence intervals for WER, using 10,000 resamples.*

Model	Original (WERR) [95% CI]	Rank=50 (WERR) [95% CI]	SNR=10 dB (WERR) [95% CI]	SNR=0 dB (WERR) [95% CI]
Whisper-large-v3 (uncorrected)	16.56% (N/A) [16.25-16.87]	24.70% (N/A) [24.26-25.15]	18.17% (N/A) [17.84-18.49]	23.70% (N/A) [23.27-24.14]
Gemma-3-4B-it	14.78% (10.73%) [14.47-15.09]	17.05% (30.96%) [16.73-17.40]	15.56% (14.34%) [15.25-15.89]	20.24% (14.56%) [19.83-20.66]
Gemma-3-27B-it	13.79% (16.72%) [13.49-14.09]	15.72% (36.37%) [15.40-16.04]	14.61% (19.58%) [14.30-14.92]	19.06% (19.54%) [18.66-19.48]

5.2. Downstream SQA Performance

Table 6 summarizes downstream QA accuracy for each model, and Table 7 reports the corresponding configurations and latencies. Overall, ground-truth text inputs substantially outperform ASR-transcribed inputs, with Gemma-3-270M as an exception because its accuracy remains close to random guessing. Among open models, Gemma-3-27B achieves the highest accuracy but also incurs the highest latency, whereas MMBERT-base performs competitively relative to its size and runtime. Larger QA models generally achieve stronger absolute performance on both ground-truth and ASR-transcribed inputs. The largest correction gains are observed for Gemma-3-4B and Gemma-3-27B; however, correction still does not close the gap to ground-truth transcription performance. Although reasoning has a negative EC coefficient, the best downstream QA configurations enable it for all models except Gemma-3-270M, and the difference is small and not statistically significant. Table 7 also shows that error correction increases latency by about 2.5× for Gemma-3-4B and Gemma-3-27B relative to directly passing

Ground Truth: Gayunpaman, iminumungkahi ng mga bagong ebidensya na ang Moriori ay isang grupo ng mainland Moriori na ibang bayan mula sa New Zealand papunta sa Chatham Islands na pinasusulong ang kanilang sariling kakaiba, mapayapang kultura.

Corrected: ~~Gayon pa man~~ Gayunpaman*, ... ebidensya ~~ng ma-riori~~ na ang Moriori ay isang grupo ng mainland ~~ma-ri~~ Moriori na ~~ng~~ ... papunta sa ~~Katam~~ Chatham Islands na ~~i~~pinasusulong ang kanilang sariling ~~kakaibang upayapang~~ kakaiba, mapayapang kultura.

Translation: However, new evidence suggests that the Moriori were a group of mainland Māori who migrated from New Zealand to the Chatham Islands, where they developed their own distinct, peaceful culture.

Figure 4: *Example of a successful correction. Gayunpaman appears in the 3rd-best ASR hypothesis, while the other corrections especially with proper names are inferred from real-world knowledge and grammar.*

Table 5: *Downstream spoken question answering accuracy under simulated noise. For cascaded ASR-EC rows, parentheses show correction gains in percentage points; other baselines show accuracy without error correction or performance degradation. Noise augmentation on API-based methods is evaluated on fewer settings to limit cost. Statistically significant differences ($p < 0.05$) under McNemar's test are marked with an asterisk (*).*

Model	No Noise	Rank=50	SNR=10 dB	SNR=0 dB
Cascaded LLM QA with ASR-EC				
Gemma-3-270M-it	26.84% (+0.12)	26.84% (+0.25)	27.47% (-0.62)	25.47% (-0.12)
Gemma-3-4B-it	75.41% (+3.50)*	68.04% (+7.24)*	73.03% (+2.87)	60.80% (+1.25)
Gemma-3-27B-it	89.39% (+2.87)*	87.02% (+9.11)*	86.39% (+1.37)	76.28% (+1.62)
Gemini-2.5-Flash (cascade)	93.63% (+0.88)	–	–	–
Non-EC baselines				
MMBERT-base	58.18%	36.70%	53.56%	47.44%

uncorrected ASR hypotheses to the QA model. We further compare cascaded systems with E2E audio LLM baselines in Table 8. For Gemini-2.5-Flash, direct audio and Whisper-transcribed inputs achieve the same accuracy before correction, while error correction improves accuracy and ground-truth text input performs slightly better. For Gemma-3n-E4B (different from Gemma 4B), the cascaded system is much closer to its ground-truth text performance than the direct-audio setting, which performs poorly. This suggests that current E2E audio LLMs remain limited for Filipino SQA, with recognition failures still propagating into incorrect reasoning as illustrated in Fig. 10.

Table 9 further examines how WER reduction (improved or worsened), model, and reasoning are associated with downstream QA performance. All features are binary variables, and the binary classifier predicts whether each individual QA instance would be correctly answered. The table shows that model choice has a stronger association with answer correctness than WER reduction or enabling reasoning. However, a reduction in WER is also positively associated with a correct answer, followed by enabling reasoning, while an increase in WER has the strongest association with an incorrect answer. Error correction helps most when it preserves or recovers words needed to identify the correct answer, but it can hurt performance when it changes the correct option or removes relevant phrases. Fig. 5, Fig. 6, and Fig. 7 illustrate cases where error correction improves passage comprehensibility, corrects question interpretation, or incorrectly changes the answer choice. Errors that occur in Gemma-3-4B but not Gemma-3-27B often reflect downstream reading-comprehension failures rather than transcription failures. As shown in Fig. 8 and Fig. 9, the reasoning trace can identify relevant evidence but still contradict itself or misinterpret the passage before selecting the final answer.

Table 6: *Downstream spoken question answering accuracy across different QA models. Parentheses show correction gains in percentage points. Statistically significant differences ($p < 0.05$) under McNemar's test are marked with an asterisk (*).*

Model	Self EC Acc. (Δ After Correction)	Gemini EC Acc. (Δ After Correction)	Ground Truth Acc.
Gemma-3-270M-it	26.84% (+0.12)	24.97% (-1.75)	27.22%
Gemma-3-4B-it	75.41% (+3.50)*	75.03% (+3.12)*	81.15%
Gemma-3-27B-it	89.39% (+2.87)*	88.01% (+1.50)	90.76%
MMBERT-base (307m)	N/A	58.93% (+0.75)	63.30%
Gemini-2.5-Flash	93.63% (+0.88)	93.63% (+0.88)	95.26%

Table 7: *Self-error correction configurations and downstream QA latency.*

Model	ASR-EC Config.	QA Reasoning?	QA Latency (s) (Total)
Gemma-3-270M-it	5-best + reasoning	No	0.02331 (0.1909)
Gemma-3-4B-it	3-best + reasoning	Yes	0.4799 (1.2251)
Gemma-3-27B-it	5-best + reasoning	Yes	3.4517 (8.4440)
MMBERT-base (307m)	N/A	N/A	0.04444 (N/A)
Gemini-2.5-Flash	5-best + reasoning	Yes	N/A

We similarly evaluate QA performance under ASR transcriptions from noise-augmented and weaker ASR hypothesis conditions. Table 5 shows the accuracy and accuracy improvement after error correction for cascaded LLM models, along with separate MMBERT and E2E baselines to evaluate performance degradation. Comparing the cascaded Gemini to the end-to-end version, we see that performance pre-correction is similar, but error correction slightly improves accuracy (+0.88 percentage points).

Gemma E4B has roughly 4B active parameters and 8B total parameters and is not directly comparable to the 4B or 27B cascaded models; however, it achieves much lower accuracy than the cascaded 4B and 27B models. For Gemma 4B and 27B, error correction consistently improves accuracy across different levels of noise simulations, while Gemma 270M remained similar to random guessing.

6. Conclusion

This study evaluated prompting-based LLM ASR error correction for low-resource spoken question answering. Results show that end-to-end performance is still limited with small models like Gemma-3n-E4B, and that ASR error correction can improve transcription quality and downstream QA accuracy in cascaded SQA systems, with around a 2.5x factor increase in inference latency compared to directly using ASR-transcribed text as input. Among open-weight models, Gemma-3-27B achieved the strongest correction performance, while Gemma-3-4B provided a more latency-efficient trade-off.

Error correction is able to partially recover downstream QA performance, with larger gains especially for Gemma-3-4B and Gemma-3-27B. However, it is unable to fully recover the performance obtained with ground truth transcriptions. The pipeline

Ground Truth: Alin sa mga sumusunod na pahayag ang tumpak na naglalarawan sa mga pananakop sa Cambodia?

Sa panahon ng ika-18 na siglo naipit ang Cambodia sa pagitan ng dalawang malalakas na mga karatig bansa, ang Thailand at Vietnam. Ilang beses nilusob ng mga Thai ang Cambodia noong ika-18 siglo at noong 1772 winasak nila ang Phnom Phen. Noong mga taon sa dulo ng ika-18 na siglo sinakop din ng mga taga-Vietnam ang Cambodia.

Answers: A. Sinakop ng Vietnam ang Cambodia nang maraming beses B. Sinakop ng Vietnam at Thailand ang Cambodia noong 1772 C. Sinira ng mga taga-Vietnam ang Phnom Phen D. Sinakop ng Thailand ang Cambodia nang maraming beses

Corrected: ... pananakop sa ~~Cambodia~~ Kambodia?
Sa panahon ng 18 na siglo, naipit ang ~~kod~~ Code Kambodya sa ... karating na bansa ... Ilang beses ~~nilusog~~ nilusob ng mga sayang Kambodya ... noong 1772 ~~minasak~~ binasak nila ang ~~prompel~~ Phom Phen. ... sa dulo ng ika-80 na siglo sinakapitin ng mga taga-Vietnam ...
Answers: A. ... ~~na~~ ng ... B. Sinira ng mga taga-Vietnam ang ~~pinomfen~~ Phom Phen ... D. ... ng ...

Translation: Which of the following statements accurately describes the invasions of Cambodia?
During the 18th century, Cambodia was caught between two powerful neighboring countries, Thailand and Vietnam. The Thai invaded Cambodia several times in the 18th century, and in 1772 they destroyed Phnom Penh. Toward the end of the 18th century, the Vietnamese also occupied Cambodia.
Answers: A. Vietnam occupied Cambodia many times, B. Vietnam and Thailand occupied Cambodia in 1772, C. The Vietnamese destroyed Phnom Penh, D. Thailand occupied Cambodia many times

Figure 5: *Example of a correction improving QA accuracy. After correction, the model changes its prediction from A to D. Despite multiple errors still being present in the correction, error correction improved comprehensibility by lessening non-existent words, and the model is able to correctly infer that Cambodia is involved in multiple conflicts.*

Ground Truth: Ano ang ikalawa sa huling pagtatanghal ng gabi? ...
Answers: A. Sina Sezen Aksu, Alessandro Safina, at Haris Alexiou, B. Apoy ni Anatolia, C. Umiikot-ikot na mga Dervish, D. Fatih Erkoç at Müslüm Gürses

Corrected: Ano ang ikalawa sa ~~kuling~~ huling pagtatanghal ng gabi? ...
Answers: A. ... ~~Haris~~ Harris Alexuo B. ~~APOY~~ Apo'y ni Anatolia, C. Umiikot ikot ng mga dervish, D. Fatih eko at Muslim ~~guses~~ Gooses

Translation: What was the second-to-last performance of the evening? Answers: A. Sezen Aksu, Alessandro Safina, and Haris Alexiou, B. Fire of Anatolia, C. Whirling Dervishes, D. Fatih Erkoç and Müslüm Gürses

Figure 6: *Example of a correction improving QA accuracy. The ASR error involves the question text, where the non-existent word 'kuling' caused the model to interpret the question as 'the second' instead of 'second to the last' (A). Corrections involved proper names, and minor spelling variations.*

Table 8: *Comparison of End-to-End and Cascaded QA models*

Model	End to End	Ground Truth Text	Cascaded (EC)
Gemini-2.5-Flash	92.76%	95.26%	93.63%
Gemma-3n-E4B-it	37.20%	84.52%	80.52%

Table 9: *QA regression coefficients.*

Feature	Coef.	Feature	Coef.
WER reduced	0.0888	WER increased	-0.7882
Gemma 27B	0.8722	Gemma 4B	0.1703
Reasoning	0.0740	MMBERT	-0.3398

is particularly effective when the correct answer choice is correctly predicted, or when error correction corrected words relevant to answering the passage. Results also remain consistent under noise-augmented conditions, and the cascaded systems outperform similarly sized E2E audio LLMs. These findings suggest that LLM-based ASR error correction is a promising approach for improving the robustness of cascaded SQA systems to ASR errors, especially in zero-training settings.

This study is limited to the Filipino language, a single ASR model, Whisper-large-v3, and to the 801 complete QA examples from the 2M-BELEBELE Filipino subset. The results may not generalize to other datasets, languages, or ASR models with different error patterns. Noise is synthetically added to read speech, which may not fully reflect real-world recording conditions or spontaneous speech. We also evaluate only one paid API model, Gemini-2.5-Flash. Other commercial models and other LLM families are not tested. Future work can explore different spoken language understanding tasks like extractive or generative question answering, where the presence of ASR errors causes further degradation due to not having answer choices, expand to other low-resource languages, and evaluate different error correction prompting strategies to mitigate degradations.

7. Acknowledgements

This work was supported by J-PEAKS funding. The computing resources used to run the models and access to Gemini were provided by Prof. Sakriani Sakti and Prof. Charibeth Cheng respectively.

8. AI Use Declaration

Generative AI tools were used solely to refine the grammar of this manuscript. The authors reviewed all AI-assisted revisions and take full responsibility for the final content.

Ground Truth: Saang bansa nagmula ang armada ni Trygvasson? ...
Answers: A. England, B. Denmark, C. Germany, D. Norway

Corrected: ~~Sa ang~~ Saang bansa nagmula ang armada ni Trigvason? ...
Answers: A. England, B. Denmark, C. Germany, D. ~~Norway~~ Thank you

Translation: Which country did Trygvasson's armada come from?
Answers: A. England, B. Denmark, C. Germany, D. Norway

Figure 7: *Example of a failed correction. Norway turned into 'Thank you', the 2nd best prediction which caused a failed prediction of B instead of D.*

Ground Truth Question: Alin sa mga sumusunod ang hindi tumpak na naglalarawan sa Pureland origami?

Model Analysis:
The question asks what is not a characteristic of Pure Land Origami.
A. Para ito sa mga matagal nang nagtutupi ng origami: The passage states it's for beginners, so this is incorrect.
B. Hindi kabilang sa paraang ito ang pabaliktad na tupi: The passage explicitly states that reverse folds are not allowed, so this is correct.
C. Dapat ay may malinaw lokasyon ang mga topi: The passage says that all folds have clear locations, so this is correct.
D. Magandang paraan ito para sa mga nagtutupi na may limitadong paghilos ng katawan: The passage states it's designed for those with limited motor skills, so this is correct.

Model Answer: B.

Figure 8: *Example of Gemma 4B contradicting itself when selecting the final answer. The model is able to recognize misspellings, but despite correctly identifying the correct answer choice D in reasoning, it chose B as the final answer.*

Ground Truth Question: Ayon sa sipi, alin sa mga sumusunod ang hindi pokus ng Mistisismo?

Model Analysis:
The question asks what is not a focus of mysticism.
A. Pagkahanap ng kaalaman na hindi nakabatay sa kamalayan para sa banal na katotohanan: This aligns with the passage's description of mysticism as seeking knowledge not based on awareness of sacred truth. This is a focus of mysticism.
B. Paghahanap ng kabatiran sa pamamagitan ng ibinahagi at kumunal na mga karanasan: The passage states that mysticism involves 'shared experiences,' so this is a focus. C. Pagdidiin tungkol sa maligayang kalagayan ng kamalayan: The passage explicitly mentions a 'peaceful, understanding, joyful' state of consciousness as characteristic of mysticism. This is a focus.
D. Paghanap ng may malay na kabatiran ng Diyos: This directly reflects the definition of mysticism as seeking unity with the consciousness of God. This is a focus.

Model answer: A.

Figure 9: *Example of a Gemma 4B interpretation failure. The passage is about the pursuit of mysticism, or seeking direct and personal experiences with a higher being. However, the model incorrectly interpreted the passage as about 'shared experiences' and chose B as the answer, which contradicts one of the passages emphasizing direct and personal experiences.*

Recognition failure:

Passage excerpt: ... Tommy Dreamer said, "Luna was the first Queen of Extreme. My first manager. Luna passed away on the night of two moons. Truly unique, just like her. A brave woman." Dustin "Goldust" Runnels said that "Luna was strange like me... perhaps even more so... I loved her and I miss her... I hope she is in a better place."

Question: Who called Luna the "Queen of Extreme"? ... C. Luna's manager, D. Tommy Dreamer

Model Analysis: The question asks who was declared the first queen of Cordillera by Dustin Ramos. Passage 3 states, "This told Dustin Ramos that Luna was the first queen of Cordillera." Therefore, Luna is the answer. The other passages discuss the recovery of a black box and the respect shown to Luna, but they do not state that she was the first queen of Cordillera. The answer choice C directly states that Luna was the first queen of Cordillera according to Dustin Ramos.

Model Answer: C - Ang manager ni Luna

Figure 10: *Example of an E2E recognition failure. The model preserves some relevant names, but misrecognizes the key phrase "Queen of Extreme" and related entities as "Cordillera" and "Dustin Ramos," leading it to choose Luna's manager instead of Tommy Dreamer.*

ASR error correction prompt

You are an expert in the Filipino language.

You are given an ASR (Automatic Speech Recognition) hypothesis for an audio utterance. Your task is to produce the best corrected transcription by:
1. Using your knowledge of Filipino grammar, vocabulary, and common ASR error patterns to identify and correct any errors in the ASR hypothesis.
2. Ensuring that the final transcription is phonetically and *structurally* similar to the hypothesis. Preserve the original wording of the hypothesis. If the word is a number or an abbreviation, do not spell it out.
3. Outputting only the corrected transcription, nothing else.

ASR hypothesis:
{hypotheses_text}

Your output must be in strict JSON format with the following structure:
analysis: a detailed analysis of the ASR hypothesis, identifying possible errors or substitutions, and ensuring that the final transcription is phonetically and structurally similar to the hypothesis.
correction: the single best corrected transcription based on your analysis.

Corrected transcription:

Question answering prompt

You are given a question and multiple answer choices in Filipino. Choose the best answer based on the provided passages. You are free to reason through the question step-by-step before giving your final answer.
Answer is in strict JSON format with the following structure:
analysis: a detailed analysis of the passages and question. Identify key information, relevant passages, and explain your reasoning process step-by-step.
final_answer: the single best answer choice (A, B, C, or D).

Passages:
{passages_text}

Question: {question_text}

Answer Choices:
{answer_choices_text}

Figure 11: *ASR error correction and question answering prompts with reasoning enabled. Braced placeholders are replaced with the corresponding input text.*

9. References

[1] C.-H. Li, S.-L. Wu, C.-L. Liu, and H.-y. Lee, "Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension," *arXiv preprint arXiv:1804.00320*, 2018.

[2] Y. Wu, S. Rallabandi, R. Srinivasamurthy, P. P. Dakle, A. Gon, and P. Raghavan, "Heysquad: A spoken question answering dataset," *arXiv preprint arXiv:2304.13689*, 2023.

[3] J. Peng, Y. Wang, B. Li, Y. Guo, H. Wang, Y. Fang, Y. Xi, H. Li, X. Li, K. Zhang *et al.*, "A survey on speech large language models for understanding," *IEEE Journal of Selected Topics in Signal Processing*, 2025.

[4] Y. Jin, J. Li, T. Gu, Y. Liu, B. Zhao, J. Lai, Z. Gan, Y. Wang, C. Wang, X. Tan *et al.*, "Efficient multimodal large language models: A survey," *Visual Intelligence*, vol. 3, no. 1, p. 27, 2025.

[5] M. Trenton, "A comprehensive survey on multilingual and multimodal automatic speech recognition systems," *Journal of Computer Technology and Software*, vol. 4, no. 10, 2025.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[7] G. T. et al., "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[8] C.-H. Lee, S.-M. Wang, H.-C. Chang, and H.-Y. Lee, "Odsqa: Open-domain spoken question answering dataset," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 949–956.

[9] R. Ma, M. Qian, M. Gales, and K. Knill, "Asr error correction using large language models," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[10] L. Feng, J. Yu, Y. Wang, S. Liu, D. Cai, and H. Zheng, "Asr-robust natural language understanding on asr-glue dataset," in *Proc. Interspeech 2022*, 2022, pp. 1101–1105.

[11] Y.-S. Chuang, C.-L. Liu, H.-Y. Lee, and L.-s. Lee, "Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering," *arXiv preprint arXiv:1910.11559*, 2019.

[12] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 16 607–16 629.

[13] Y. He, Z. Liu, G. Lin, S. Sun, B. Wang, W. Zhang, X. Zou, N. Chen, and A. Aw, "Meralion-audiollm: Advancing speech and language understanding for singapore," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2025, pp. 22–30.

[14] Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," *arXiv preprint arXiv:2410.17196*, 2024.

[15] Z. Wan, C.-H. H. Yang, Y. Yu, J. Tian, S. Li, K. Hu, Z. Chen, S. Watanabe, F. Cheng, C. Chu *et al.*, "Speechiq: Speech-agentic intelligence quotient across cognitive levels in voice understanding by large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 30 381–30 398.

[16] Y. Song, S. Shrestha, C. Lyu, E. Khatibi, P. Zhang, H. Xu, N. Dutt, and A. Rahmani, "Medspeak: A knowledge graph-aided asr error correction framework for spoken medical qa," *arXiv preprint arXiv:2602.00981*, 2026.

[17] Y. Weng, S. S. Miryala, C. Khatri, R. Wang, H. Zheng, P. Molino, M. Namazifar, A. Papangelis, H. Williams, F. Bell *et al.*, "Joint contextual modeling for asr correction and language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6349–6353.

[18] M. R. Costa-Jussa, B. Yu, P. Andrews, B. Alastruey, N. C. Camgöz, J. Chuang, J. Maillard, C. Ropers, A. Turkatenko, and C. Wood, "2m-belebele: Highly multilingual speech and american sign language comprehension dataset," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 10 893–10 904.

[19] L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Khabsa, "The belebele benchmark: a parallel reading comprehension dataset in 122 language variants," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 749–775.

[20] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[21] J. R. Montalan, J. P. Layacan, D. D. Africa, R. I. S. Flores, M. T. L. Ii, T. D. Magsajo, A. Cayabyab, and W. C. Tjhi, "Batayan: A filipino nlp benchmark for evaluating large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 31 239–31 273.

[22] Y. Susanto, A. V. Hulagadri, J. R. Montalan, J. G. Ngui, X. Yong, W. Q. Leong, H. Rengarajan, P. Limkonchotiwat, Y. Mai, and W. C. Tjhi, "Sea-helm: Southeast asian holistic evaluation of language models," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 12 308–12 336.

[23] M. Marone, O. Weller, W. Fleshman, E. Yang, D. Lawrie, and B. Van Durme, "mmbert: A modern multilingual encoder with annealed language learning," *arXiv preprint arXiv:2509.06888*, 2025.

[24] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-scale reading comprehension dataset from examinations," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 785–794.

[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 8440–8451.

[26] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.