

Empirical evaluation of generative fusion decoding for Indonesian general and medical speech recognition

Asril Jarin, Lyla Ruslana Aini, Agung Santosa, Gunarso, Mohammad Teduh Uliniansyah,
Elvira Nurfadhilah, Siska Pebiana

Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics,
National Research and Innovation Agency (BRIN), Bandung, Indonesia

Article Info

Article history:

Received Nov 5, 2025

Revised Aug 3, 2026

Accepted Aug 12, 2026

Keywords:

Automatic speech recognition

Domain adaptation

Generative fusion decoding

Indonesian speech

Large language models

ABSTRACT

This study empirically evaluates generative fusion decoding (GFD) with open-weight large language models for Indonesian automatic speech recognition (ASR) under two controlled offline conditions: general read speech (FLEURS-ID) and medical dictation (Harkit). The training-free byte-level fusion integrates three frozen language-model backends—SEA-LION, Sahabat-AI, and Mistral—with frozen Whisper-large-v3 during decoding. A development partition was used to select the static fusion weight and prompting condition, and a held-out test partition was used for final evaluation. On FLEURS-ID, the best GFD configuration reduced corpus-level word error rate (WER) from 5.91% to 4.17%, corresponding to a 29.44% relative reduction. On Harkit, it reduced WER from 15.89% to 12.30%, a 22.59% relative reduction. Paired utterance-level tests yielded $p < 0.01$ for both datasets. Several non-prompted configurations already improved over the unfused Whisper-large-v3 baseline, suggesting that the external language-model prior contributed more consistently than explicit prompting. Prompting produced limited changes on FLEURS-ID and larger but backend-dependent changes on Harkit. The results support GFD as a training-free decoding strategy under the evaluated conditions, while broader validation is required for noisy, spontaneous, streaming, and clinical deployment settings.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Asril Jarin

Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics

National Research and Innovation Agency

Bandung, Indonesia

Email: asri003@brin.go.id

1. INTRODUCTION

Automatic speech recognition (ASR) has advanced substantially over the past decade, transitioning from hybrid hidden Markov model (HMM)-based systems to fully neural, end-to-end approaches [1], [2]. Pioneering architectures such as deep recurrent neural networks [3] and the attention-based encoder-decoder model listen, attend, and spell [4] paved the way for direct acoustic-to-text mapping. Further breakthroughs in data augmentation [5], self-supervised representation learning [6]–[8], and architectural innovations such as the conformer [9] have substantially improved robustness and scalability. These advances, combined with massive training datasets, have enabled state-of-the-art systems such as Whisper [10] to achieve near-human transcription accuracy under clean, high-resource conditions.

Despite this progress, Indonesian ASR continues to face unique challenges. The Indonesian language exhibits agglutinative morphology and productive derivational affixation, which substantially

increase lexical variability and surface-form diversity in spoken language [11], [12]. In addition, frequent Indonesian–English code-switching—a common phenomenon in both informal and professional communication—introduces structural and phonological variability, further complicating acoustic and language modeling for ASR systems [13]–[15]. These linguistic characteristics, combined with the limited availability of annotated domain-specific corpora, degrade recognition performance in specialized contexts such as medical transcription.

Prior studies on medical ASR consistently report substantial increases in error when confronted with terminology-dense utterances and domain-specific expressions, even with large-scale neural models. Although multilingual models such as cross-lingual speech representation (XLS-R) and Whisper-large-v3 demonstrate strong cross-lingual generalization, their accuracy consistently declines under domain-mismatch conditions, particularly when encountering rare or morphologically complex word forms and code-switched utterances that fall outside their pre-training distributions [16]–[18]. Recent progress in multilingual datasets—including few-shot learning evaluation of universal representations of speech (FLEURS) [19], multilingual LibriSpeech (MLS) [20], and massively multilingual speech (MMS) [21]—have significantly advanced cross-lingual transfer for ASR. However, these corpora still lack sufficient coverage of domain-specific linguistic phenomena, particularly in low-resource domains like healthcare. In such settings, end-to-end models often misrecognize specialized terms, abbreviations, and context-dependent expressions that deviate from their pre-training distribution.

External language-model integration in end-to-end ASR spans several distinct paradigms. Conventional shallow fusion performs log-linear interpolation of independently trained ASR and language-model scores during beam-search inference and does not require retraining the ASR model [22]. Training-time fusion methods, such as cold fusion, instead train the sequence-to-sequence recognizer together with a fixed pretrained language model [23]. Contextual approaches such as contextual listen, attend and spell (CLAS) jointly optimize the recognizer with embeddings of task- or user-specific context phrases [24], whereas second-pass and synchronous methods such as RescoreBERT [25] and speedy ASR-large language model (LLM) synchronous aggregation (SALSA) [26] introduce linguistic information at different stages of the recognition process. These approaches differ in their training requirements, contextual resources, score compatibility, architectural assumptions, and inference-time cost. Hsu *et al.* [27] introduced generative fusion decoding (GFD) as a generative, shallow-fusion framework for combining autoregressive recognition models with large language models whose token vocabularies may differ. GFD maps model likelihoods into a shared byte-level representation and synchronizes ASR and LLM progression during beam-search decoding. This design permits instruction-conditioned LLM priors to be integrated without retraining either model.

The present study evaluates the effectiveness of this training-free byte-level fusion for Indonesian ASR in both general read speech and terminology-dense medical dictation. Accordingly, this study conducts an empirical evaluation of GFD for Indonesian ASR across two representative domains: i) general speech: representing controlled, clean read-speech conditions; and ii) medical speech: characterized by specialized terminology and stylistic variability. We employ three open-weight LLMs—Llama-SEA-LION-v3.5-8B-R [28], Gemma2-9B-cpt-sahabatai-v1-base [29], and Mistral-7B [30]—fused with Whisper-large-v3 using the GFD framework.

The investigation systematically analyzes the effects of the fusion weight parameter (r), instruction prompting, and domain sensitivity on decoding improvements across general and medical ASR. To guide this investigation, we address the following research questions:

- RQ1: does GFD improve ASR performance compared to a strong Whisper baseline in Indonesian general and medical domains?
- RQ2: how does the fusion weight parameter (r) influence recognition accuracy across domain conditions, and what values yield optimal trade-offs between acoustic evidence and linguistic priors?
- RQ3: how do the choice of open-weight LLM backend and fixed instruction prompting affect GFD performance across the two evaluated speech domains?

These research questions collectively examine the effectiveness of GFD across different domains, fusion settings, and LLM backends.

This study does not propose a new fusion algorithm. Instead, it adopts the original byte-level GFD framework proposed by Hsu *et al.* and investigates how external LLM priors interact with Whisper-large-v3 under Indonesian ASR conditions. The contributions of this paper are as follows:

- i) Domain-focused empirical evaluation: present a controlled empirical evaluation of GFD for Indonesian ASR across two representative speech conditions: general read speech (FLEURS-ID) and controlled, terminology-dense medical dictation (Harkit).

- ii) Fusion-weight characterization: show that the fusion weight parameter r governs the balance between acoustic grounding and linguistic prior strength, and that its optimal value is domain-dependent rather than universal.
- iii) Statistical confirmation of gains: using paired one-tailed t-tests and effect-size analysis, we confirm that GFD yields statistically significant and practically meaningful word error rate (WER) reductions relative to a strong Whisper-large-v3 baseline.
- iv) LLM-prior and prompting ablation: distinguish the contribution of the external LLM prior from explicit task conditioning by comparing whisper-only baseline, non-prompted GFD, and instruction-prompted GFD across three open-weight LLM backends. The results show that prompting produces a secondary and configuration-dependent effect relative to the contribution of the external LLM prior.

Through this study, empirical evidence is provided that GFD can improve Indonesian ASR under the evaluated offline conditions. This improvement is achieved by integrating external LLM priors during decoding without retraining either model. The results further characterize how the effectiveness of this integration depends on speech condition, static fusion weight, prompting configuration, and LLM backend, providing a foundation for future domain-sensitive decoding strategies and deployment-oriented evaluation.

2. METHODS

2.1. Overall experimental design

This study evaluates the effectiveness of GFD [27] for Indonesian ASR in two representative domains: general and medical. The experimental pipeline consists of four key components: i) dataset selection and preparation for domain-general and domain-specific tasks; ii) baseline ASR decoding setup using Whisper-large-v3; iii) implementation of GFD with systematic variation of the fusion weight parameter (r); and iv) comprehensive evaluation and statistical validation of performance improvements. All experiments were conducted using offline utterance-level decoding. Each utterance was evaluated as a complete input instance, and the study did not implement or assess chunk-wise audio processing, incremental transcript delivery, endpoint detection, or latency-constrained streaming inference.

Figure 1 presents the complete offline experimental workflow. For each dataset, complete utterances from FLEURS-ID and Harkit are decoded using the same frozen Whisper-large-v3 model, either as the unfused baseline or through GFD with one of three frozen LLM backends. The GFD experiments evaluate prompted and non-prompted conditions across the predefined static fusion weights r . The resulting transcripts are evaluated using corpus-level WER and paired utterance-level statistical analysis.

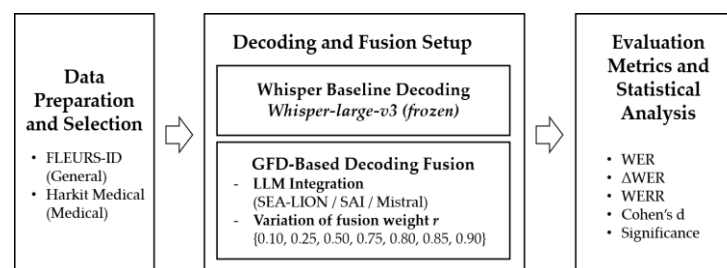


Figure 1. Offline experimental workflow for evaluating GFD for Indonesian ASR

2.2. Datasets and scenarios

Two datasets were used in this study: FLEURS-ID [19] and Harkit medical [31]. Table 1 summarizes the duration, number of utterances, and utterance-length characteristics of the two datasets used in the evaluation.

- i) FLEURS-ID is the Indonesian subset of the multilingual FLEURS benchmark [19] and consists of short, controlled read-speech utterances representing general-domain language. In this study, FLEURS-ID is used as a controlled, clean-speech reference condition to evaluate whether GFD improves recognition when the acoustic signal and speaking style are relatively constrained. It should therefore not be interpreted as representing spontaneous conversation, environmental noise, far-field speech, or other unconstrained real-world conditions.
- ii) Harkit medical is a domain-specific Indonesian clinical speech corpus developed under the National Research and Innovation Agency (BRIN). The corpus contains transcribed medical dictations involving

clinical terminology, abbreviations, and mixed Indonesian–English expressions commonly found in medical documentation [31]. In this study, Harkit is used as a controlled, terminology-dense, and domain-mismatched condition to evaluate decoding-time linguistic guidance. The corpus does not represent the complete diversity of medical specialties, spontaneous clinician–patient interaction, environmental noise, or institution-specific clinical workflows.

Both datasets were standardized to Unicode NFC form [32] to ensure canonical text equivalence while preserving punctuation and casing. Code-switched English tokens (e.g., “CT scan” and “check-up”) were retained to maintain lexical authenticity. No proprietary data were used.

Table 1. Summary of the Indonesian speech datasets used in the controlled offline evaluation

Dataset	Evaluation condition	Duration (h)	Utterances	Mean utterance duration (s)	Minimum (s)	Maximum (s)
FLEURS-ID	General	2.27	663	12.32	4.44	41.30
Harkit	Medical dictation	1.40	563	8.94	3.80	21.50

2.3. Baseline automatic speech recognition model

The baseline system is based on Whisper-large-v3 [10], a multilingual transformer encoder–decoder model trained on 1 million hours of weakly labeled audio and 4 million hours of pseudo-labeled audio, and used throughout this study as the fixed acoustic model. Whisper represents a strong state-of-the-art foundation for multilingual ASR. To isolate the contribution of GFD, whisper’s parameters were frozen and all modifications were restricted to the decoding stage. Both the baseline and GFD systems were evaluated using offline utterance-level decoding. The baseline used beam search with a beam width of 5 and no nucleus sampling, whereas the external LLM in GFD was decoded with a temperature of 1.7 and no top-p sampling. Accordingly, the reported performance differences reflect decoding-level fusion rather than model retraining or differences in streaming configuration.

2.4. Generative fusion decoding

GFD is a byte-level generative shallow-fusion method that incorporates external LLM priors into beam-search decoding. Like conventional shallow fusion, GFD combines independently obtained ASR and language-model probabilities during inference. Its distinguishing feature is the conversion of candidate-sequence probabilities into a shared byte-level representation, allowing the ASR model and the LLM to progress synchronously even when they use different token vocabularies [27]. Unlike cold fusion, which incorporates a fixed pretrained language model during sequence-to-sequence ASR training [23], and CLAS, which jointly learns contextual phrase embeddings together with the ASR model [24], GFD performs integration entirely during inference. Consequently, GFD enables external language model guidance without requiring retraining either the ASR model or the LLM.

In the present study, Whisper-large-v3 and all external LLMs were kept frozen, and integration was performed exclusively through GFD during decoding. No conventional shallow-fusion, cold-fusion, or CLAS system was implemented in the present study. Instead, the experimental baseline consists of the unchanged Whisper-large-v3 decoder, allowing the effect of GFD to be isolated under otherwise identical decoding conditions. This probabilistic fusion is controlled by a scalar fusion weight r , which determines the relative influence of the ASR and LLM during beam-search decoding. Accordingly, at each decoding step t , the best partial beam hypothesis $B_{1:t}^*$ is obtained by maximizing the following weighted combination of ASR and LLM log-probabilities as shown in (1).

$$B_{1:t}^* = \arg \max_{\{B_{1:t}\}} [(1 - r) \log P_{ASR}(B_{1:t}; Z_{TR}) + (r) \log P_{LLM}(B_{1:t-k}; Z_{LM})] \quad (1)$$

Where $B_{1:t}$ denotes the partial beam sequence up to time step t . $P_{ASR}(B_{1:t}; Z_{TR})$ is the probability assigned by the transcription model (ASR transcription model) parameterized by Z_{TR} . $P_{LLM}(B_{1:t-k}; Z_{LM})$ is the autoregressive language-model probability of the external LLM—parameterized by Z_{LM} —for continuing the hypothesis sequence, conditioned only on the previously generated $t - k$ text tokens. $r \in [0, 1]$ is the fusion weight controlling the balance between the acoustic model and the LLM contribution. k is a dynamic shifting value determined separately for each beam hypothesis. Following the original GFD implementation, k is set to the byte length of the last token proposed by the text-recognition model. This selection restricts the number of sequences rescored by the LLM to approximately the beam size while retaining the maximum available pre-expansion context.

In the present study, r is implemented as a static scalar interpolation coefficient. For each dataset, LLM backend, and prompting condition, a single r value is applied to all utterances and autoregressive

decoding steps. Candidate values are evaluated using the development split, and the selected configuration is then applied to the held-out test set. No utterance-level, token-level, confidence-conditioned, or automatically domain-conditioned adjustment of r is performed. Accordingly, r should be interpreted as an empirically selected coefficient for a particular experimental configuration rather than as a direct percentage measure of the linguistic contribution.

The dynamic shifting value k and the corresponding delayed corrective-feedback mechanism follow the original GFD algorithm and implementation without modification or independent tuning. Because the length of the most recent recognition token may differ across beam hypotheses, k is dynamically determined rather than fixed globally. This beam-level formulation (1) is fully equivalent to incremental token-level interpolation of ASR and LLM log-probabilities executed during autoregressive decoding; the byte-level representation ensures both models operate over the same token space, enabling training-free fusion without embedding projection or vocabulary remapping. This formulation allows domain priors to be incorporated through probability-space guidance rather than parameter updates, making GFD applicable across the evaluated general and specialized Indonesian ASR conditions without retraining.

Although GFD combines ASR and LLM probabilities incrementally at each autoregressive decoding step, this stepwise operation should not be interpreted as a streaming ASR implementation. In the present experiments, complete utterances were decoded offline, and no chunk-level audio processing, partial-hypothesis emission, endpointing, or latency measurement was performed. Therefore, the reported results characterize offline decoding performance rather than real-time ASR behavior.

The decoding framework nevertheless supports instruction-conditioned and domain-sensitive linguistic guidance without retraining. To characterize the sensitivity of this guidance, we evaluated $r \in \{0.10, 0.25, 0.50, 0.75, 0.80, 0.85, 0.90\}$. The optimal coefficient was not uniform across the evaluated conditions: FLEURS-ID achieved its best result at $r = 0.75$, whereas Harkit achieved its best result at $r = 0.50$. These values should be interpreted as configuration-specific empirical optima rather than universal measures of the required LLM contribution. The differing optima motivate future adaptive fusion mechanisms that condition r on recognition confidence, ASR–LLM agreement, or domain-related cues.

2.5. Large language model integration and ablation conditions

Three frozen open-weight LLMs were integrated into the GFD decoding pipeline: Mistral-7B, Llama-SEA-LION-v3.5-8B-R (SEA-LION), and Gemma2-9B-cpt-sahabatai-v1-base (SAI). These models represent general multilingual, Southeast Asian regional, and Indonesian-adapted language-model backgrounds, respectively. The comparison is empirical; the present study does not independently isolate the effects of pre-training data composition, tokenizer design, instruction tuning, or model architecture. To distinguish the external LLM prior from explicit task conditioning, the evaluation uses three decoding conditions:

- i) Whisper-only baseline ($r = 0$): the external LLM score is excluded from the interpolation, and decoding is performed using frozen Whisper-large-v3 only.
- ii) Non-prompted GFD ($r > 0$): the external LLM receives the partial ASR hypothesis without any task- or domain-specific instruction. This condition evaluates the contribution of the pretrained LLM prior alone.
- iii) Instruction-prompted GFD ($r > 0$): the same LLM backend and GFD procedure are used, but a fixed dataset-specific instruction is prepended to the partial hypothesis. The selected prompts are:
 - FLEURS-ID: “Berikut ini adalah transkripsi kalimat ucapan bahasa Indonesia:”
English translation: “Below is the transcription of an Indonesian spoken sentence:”
 - Harkit: “Anda adalah LLM dalam konteks medis. Pertahankan istilah medis, farmasi, anatomi, dan diagnosis apa adanya”.
English translation: “You are an LLM operating in a medical context. Preserve medical, pharmaceutical, anatomical, and diagnostic terms exactly as they are”.

The prompted and non-prompted conditions were evaluated using the same candidate values of the static fusion weight r , enabling direct same- r comparisons of prompt effects. The final prompt template and r value for each backend–condition pair were selected using the development split and subsequently fixed for evaluation on the held-out test set. No utterance-specific prompt selection or test-set prompt optimization was performed. Corpus-level sensitivity to r and utterance-level distributions of the development-selected configurations are reported in sections 3.2 and 3.4, respectively.

2.6. Evaluation metrics and statistical analysis

ASR accuracy was assessed using WER as the primary metric, computed as (2). Where S is the number of substitutions, D deletions, I insertions, and N the total number of reference words.

$$WER = \frac{S+D+I}{N} \quad (2)$$

Statistical significance was assessed using paired one-tailed t-tests on utterance-level WER differences between the unfused Whisper-large-v3 baseline and the best-performing GFD configuration for each evaluated dataset. The paired design compares the two decoding systems on identical utterances, thereby controlling for utterance-specific variability. Statistical testing was performed only after the final configuration had been selected on the development partition, ensuring that the held-out test data were used exclusively for independent evaluation. In addition to p-values, Cohen's d was computed to quantify the practical magnitude of the observed improvements.

2.7. Reproducibility and implementation

All experiments were implemented in Python 3.11 using PyTorch 2.2 and Hugging Face Transformers 4.43 and were executed on an NVIDIA A100 GPU with 80 GB of memory. The evaluation used offline utterance-level decoding and focused on transcription accuracy. Real-time factor, first-token latency, end-to-end latency, peak memory usage, computational throughput, and partial-transcript stability were not measured. Consequently, the present results should not be interpreted as evidence of real-time or streaming readiness.

To facilitate experimental reproducibility, this study follows the original open-source implementation of GFD released by Hsu *et al.* [27] while using the publicly available Whisper-large-v3 checkpoint. The evaluated FLEURS-ID and Harkit medical-dictation datasets were obtained from the official repositories listed in the data availability statement, with the evaluation subsets selected according to the criteria described in section 2.2. All model parameters remained frozen throughout the experiments, and only decoding-time configurations—including the static fusion weight, prompting condition, and other decoding parameters described in subsections 2.2–2.6—were varied.

3. RESULTS AND DISCUSSION

3.1. Overall performance across datasets

Table 2 reports corpus-level WER for the unfused Whisper-large-v3 baseline and the best GFD configuration selected for each dataset. WER is reported as a percentage, Δ WER denotes the absolute WER reduction in percentage points (pp), and WERR denotes the relative WER reduction. Lower WER indicates better recognition performance.

Table 2. Corpus-level WER of Whisper-large-v3 and the best GFD configurations on the held-out test sets

Dataset	Whisper-large-v3 WER (%)	Best GFD WER (%)	Δ WER (pp)	WERR (%)	Best GFD configuration
FLEURS-ID (general)	5.91	4.17	1.74	29.44	SEA-LION, prompted, $r = 0.75$
Harkit (medical)	15.89	12.30	3.59	22.59	SAI, prompted, $r = 0.50$

On FLEURS-ID, the lowest corpus-level WER is obtained by SEA-LION under the prompted configuration with a static fusion weight of $r = 0.75$, reducing WER from 5.91% to 4.17% (an absolute reduction of 1.74 pp and a relative reduction of 29.44%). On Harkit, the lowest corpus-level WER is obtained by SAI under the prompted configuration with $r = 0.50$, reducing WER from 15.89% to 12.30% (an absolute reduction of 3.59 pp and a relative reduction of 22.59%). These results identify the best-performing configurations under the evaluated experimental settings and establish consistent improvements over the unfused Whisper-large-v3 baseline.

Overall, the results indicate that GFD generalizes across both evaluated conditions. Still, its performance depends on the interaction among the LLM backend, the prompting condition, the speech domain, and the static fusion weight. No single backend or prompting configuration is uniformly optimal across the two datasets, highlighting the importance of configuration-specific evaluation rather than applying a single decoding strategy universally.

3.2. Sensitivity to fusion weight (r)

Figure 2 shows corpus-level WER across the evaluated static fusion weights for SEA-LION, SAI, and Mistral under prompted and non-prompted decoding. The same marker identifies each LLM backend in both panels, while solid and dashed lines denote prompted and non-prompted decoding, respectively. The horizontal dotted line represents the unfused Whisper-large-v3 baseline ($r = 0$). All configurations are compared using identical candidate values of r , enabling direct assessment of the contribution of explicit instruction prompting beyond the external LLM prior. Lower WER indicates better recognition performance.

Figure 2(a) shows FLEURS-ID, the lowest WER is obtained with prompted SEA-LION at $r = 0.75$. Across the evaluated configurations, improvements are generally maintained from the low-to-moderate fusion settings through $r = 0.80$, whereas performance deteriorates as r approaches 0.90. This pattern indicates that GFD benefits from substantial LLM guidance in the clean read-speech condition, but excessive weighting of the LLM score eventually weakens consistency with the recognition evidence.

Figure 2(b) shows Harkit medical dictation, the optimum shifts to $r = 0.50$, with prompted SAI producing the lowest corpus-level WER. This result indicates that terminology-dense speech benefits from LLM guidance, but the most effective configuration retains a more balanced interpolation between the ASR and LLM scores than the FLEURS-ID optimum. Performance deterioration beyond the optimum suggests that excessive linguistic bias may produce over-correction when the LLM prior conflicts with the recognition evidence. Instruction prompting provides additional domain conditioning for several backend-weight combinations, but its benefit is not uniform across all configurations and does not eliminate the need to preserve adequate ASR grounding.

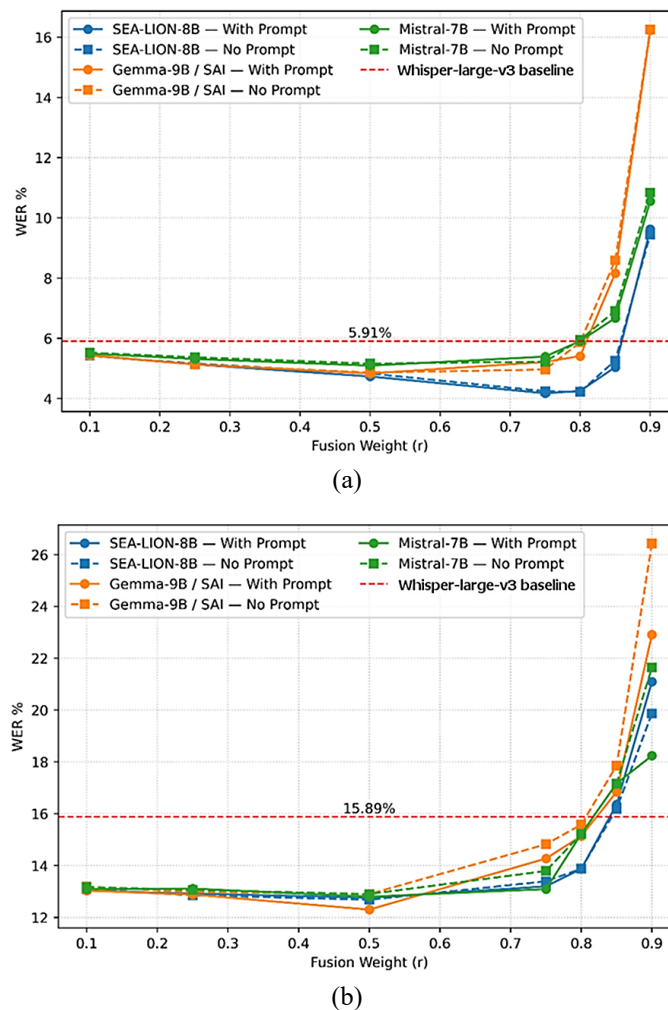


Figure 2. Sensitivity of corpus-level WER to the static fusion weight r on (a) FLEURS-ID and (b) Harkit

Taken together, these results show that r is a domain-sensitive interpolation coefficient rather than a universal constant. FLEURS-ID reaches its optimum at $r = 0.75$, whereas Harkit reaches its optimum at $r = 0.50$. Because the ASR and LLM score distributions are not necessarily calibrated to a common scale, these numerical values should not be interpreted as absolute measures of linguistic influence across domains or backends. Instead, the differing optima indicate that the effective balance depends jointly on acoustic reliability, lexical uncertainty, prompting condition, and the alignment of the LLM prior with the target speech. A single dataset-level value may therefore be suboptimal for individual utterances or decoding steps. This observation

motivates future adaptive fusion mechanisms that adjust r according to local recognition confidence, ASR–LLM agreement, or detected domain characteristics.

3.3. Statistical significance and effect size

Table 3 complements the corpus-level results in Table 2 with paired utterance-level statistical tests. The WER values are reported at the corpus level, ΔWER denotes the absolute reduction in pp, and the p -values and Cohen's d effect sizes are derived from paired utterance-level observations. Both datasets show statistically significant improvements ($p < 0.01$), confirming that the gains delivered by GFD are not attributable to statistical variance or random fluctuations in decoding. The effect size values (Cohen's $d \approx 0.30$ – 0.36) fall within the small effect sizes range, as expected, given that Whisper-large-v3 is already a robust multilingual baseline for Indonesian ASR. In such strong-baseline regimes, even moderate effect sizes can be considered materially meaningful due to diminishing returns behavior typical in low-WER regions. More importantly, the statistical results demonstrate that GFD benefits general-domain and domain-specific ASR consistently, without requiring model retraining. In practice, this indicates that decoding-level generative fusion operates as a robust improvement layer that can be expected to generalize across task conditions — even when the performance gap between domains differs substantially in absolute WER magnitude.

Table 3. Statistical comparison of Whisper-large-v3 and the best GFD configurations

Dataset	Whisper WER (%)	Best GFD WER (%)	ΔWER (pp)	p -value	Cohen's d
FLEURS-ID	5.91	4.17	1.74	1.63×10^{-19}	0.36
Harkit (medical)	15.89	12.30	3.59	3.50×10^{-13}	0.31

3.4. LLM-prior and prompt-conditioning ablation

Figure 3 compares utterance-level WER distributions across LLM backends and prompting conditions using the development-selected static r for each backend–prompt combination. Each violin represents the distribution across test utterances, with the black marker and adjacent value indicating the median WER. Unlike the corpus-level WER in Table 2, Figure 3 summarizes typical utterance-level behavior; therefore, the two measures are complementary rather than numerically equivalent. Because each configuration uses its own selected r , Figure 3 compares the best-performing configurations rather than isolating the effect of prompting; direct same- r prompt comparisons are provided in Figure 2.

Figure 3(a) shows that SEA-LION achieves the lowest median utterance-level WER on the FLEURS-ID. Figure 3(b) shows that prompted Mistral achieves the lowest median on Harkit, whereas prompted SAI, despite having a higher median and broader distribution, achieves the lowest Harkit corpus-level WER reported in Table 2. This difference reflects the aggregation methods: Figure 3 emphasizes typical utterance-level behavior through the median, whereas Table 2 aggregates errors across all reference words. Overall, the results indicate that backend performance depends on the speech condition and evaluation metric, while the external LLM prior provides a more consistent improvement than explicit instruction prompting, whose effect remains configuration-dependent.

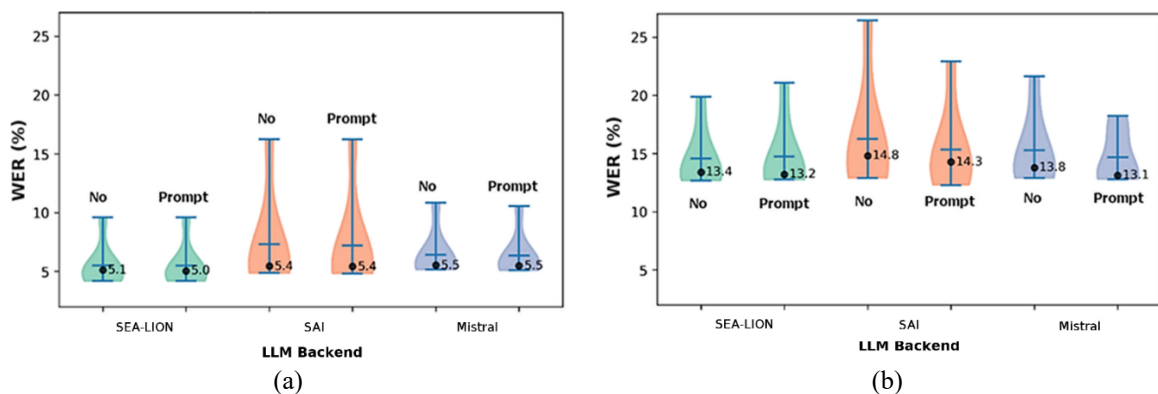


Figure 3. Utterance-level WER distributions for the development-selected GFD configurations on (a) FLEURS-ID and (b) Harkit

3.5. Discussion

Across the general-domain (FLEURS-ID) and medical-domain (Harkit) evaluations, the results collectively confirm that GFD improves Indonesian ASR performance relative to a strong Whisper baseline (RQ1). The improvements are obtained without retraining, purely through probabilistic steering at decoding time. This finding establishes GFD as a practical, scalable mechanism for exploiting linguistic priors in underrepresented languages such as Indonesian.

Second, fusion weight analysis demonstrates that r is not a universal constant (RQ2). Instead, it encodes a domain trade-off: higher r values ($\approx 0.70 - 0.80$) are optimal for general-domain Indonesian speech, where acoustic evidence is clean, and Whisper already models core structure well; whereas medical speech benefits most at moderate fusion (≈ 0.50), where terminology density requires linguistic prior support but excessive LLM dominance leads to over-correction. The observed domain-dependent optimum demonstrates empirically that a fixed coefficient is unlikely to be universally optimal across Indonesian ASR conditions [33].

An additional limitation concerns the use of a static fusion weight. In the present study, a single value of r , selected using development data, is applied uniformly throughout decoding, even though acoustic reliability and lexical ambiguity may vary substantially across utterances and decoding positions. The observed domain-dependent optimum therefore suggests that a fixed coefficient is unlikely to be universally optimal. Future work may instead investigate adaptive fusion mechanisms operating at multiple levels. At the utterance level, r could be conditioned on global ASR confidence, estimated acoustic quality, or detected domain cues. At the token or decoding-step level, it could be adjusted according to posterior entropy, the score margin between competing beam hypotheses, ASR–LLM disagreement, or uncertainty in next-token prediction. Such a strategy could increase LLM influence when acoustic evidence is uncertain but linguistic evidence is reliable, while preserving stronger ASR guidance when the language prior risks over-correction.

Implementing adaptive fusion would require careful calibration to ensure stable decoding behavior. In particular, the fusion coefficient should be smoothed and bounded to avoid abrupt fluctuations during beam search. All adaptation rules should be tuned exclusively on development data before evaluation on held-out test data to prevent selection bias. The present study does not implement such an adaptive mechanism; instead, the empirical observation that different domains favor different values of r provides a motivation for future confidence- and domain-conditioned fusion strategies.

The present results should not be interpreted as demonstrating that GFD outperforms conventional shallow fusion, cold fusion, or CLAS, because these systems were not implemented in the current experiments. The empirical comparison in this study is specifically between GFD and the unchanged Whisper-large-v3 decoder. The alternative methods also represent different integration assumptions. Conventional shallow fusion combines independently trained ASR and LM scores during beam search [22]. Cold fusion incorporates a fixed pretrained LM during recognizer training [23], whereas CLAS jointly learns contextual-phrase representations with the ASR system [24]. By contrast, GFD uses byte-level likelihood mapping to combine a frozen recognizer and a frozen autoregressive LLM during decoding [27]. This training-free property makes GFD particularly relevant when retraining data or model access is limited. Still, it does not by itself establish superiority over training-based or context-specific alternatives.

A direct comparative study should implement these methods under matched data partitions, text normalization, beam width, search budget, language-model capacity, contextual resources, and hyperparameter-selection procedures. In addition to WER, such evaluation should report training requirements, inference latency, memory consumption, and computational cost. This controlled comparison is left for future work.

The ablation analysis distinguishes the contribution of the external LLM prior from that of explicit instruction prompting (RQ3). The Whisper-only baseline ($r = 0$) excludes external LLM guidance; non-prompted GFD introduces the pretrained LLM prior without task instructions, and prompted GFD additionally incorporates a fixed dataset-specific instruction. Because several non-prompted configurations already improve over the Whisper baseline, the observed gains cannot be attributed solely to prompting. Instead, the external LLM prior provides the more consistently observed contribution. In contrast, prompt conditioning produces smaller, configuration-dependent changes whose effectiveness varies with the LLM backend, speech domain, and selected static fusion weight.

Backend comparisons likewise do not support a universal performance ranking. SEA-LION achieves the lowest median utterance-level WER on FLEURS-ID, prompted Mistral attains the lowest median on Harkit, and prompted SAI produces the lowest Harkit corpus-level WER. These complementary observations indicate that backend effectiveness depends on both the speech condition and the aggregation metric. Because the present experiments do not isolate the individual effects of pre-training corpus composition, tokenizer design, instruction tuning, or model architecture, attributing the observed differences to any single backend characteristic remains outside the scope of this study.

Beyond these mechanistic interpretations, the present findings should also be interpreted within the boundaries of the evaluated data and decoding conditions. FLEURS-ID represents controlled read speech, whereas Harkit represents a controlled, terminology-dense medical-dictation condition. The Harkit evaluation does not cover the full breadth of Indonesian clinical speech, including spontaneous doctor–patient conversations, multiparty clinical interactions, diverse medical specialties, varying documentation workflows, environmental noise, far-field recording, accent and speaking-style diversity, or continuously changing clinical contexts [24]. Moreover, the relatively limited duration of the two evaluation sets constrains broad deployment-level generalization [34]. Accordingly, the results provide controlled evidence that GFD can improve Indonesian ASR under the evaluated conditions, including the Harkit medical-dictation dataset. Still, they do not establish robustness across the broader Indonesian clinical-speech domain or unconstrained real-world environments.

From a deployment perspective, the evaluated configuration is therefore most directly applicable to offline, server-side transcription workflows, such as batch transcription of prerecorded Indonesian speech and post-encounter transcription of controlled medical dictation. In the latter scenario, the generated transcript should be reviewed and corrected by a qualified user before it is incorporated into a clinical record. Larger and more diverse Indonesian corpora covering multiple medical specialties, spontaneous and dictated speech, diverse speakers, interaction types, documentation workflows, and acoustic conditions are required for broader validation.

A further limitation concerns the decoding protocol. Although GFD integrates LLM probabilities at each autoregressive decoding step, the present implementation was evaluated only under offline full-utterance decoding and does not process incoming speech in streaming chunks or produce continuously updated partial transcripts. Consequently, the observed WER improvements should not be interpreted as evidence of real-time applicability, streaming readiness, on-device deployment, or interactive conversational ASR. Deployment in a new application domain would also require a representative development set to select the static fusion weight and prompting condition, after which the selected configuration would remain fixed during operation. Future deployment-oriented evaluations should therefore examine real-time factor, first-token and end-to-end latency, memory and computational overhead, computational throughput, endpointing behavior, and partial-hypothesis stability. Such studies should also determine whether the additional LLM scoring cost can be controlled without sacrificing the accuracy gains observed under offline decoding.

The integration of ASR and external LLMs also raises ethical and privacy considerations that are not captured by aggregate WER. The present study uses existing datasets obtained from the official repositories identified in the data availability statement and does not involve new participant recruitment or new speech-data collection. Nevertheless, prospective use with real clinical speech may involve sensitive personal and medical information. Operational implementation would therefore require data minimization, controlled access, secure storage and transmission, appropriate retention limits, auditability, and compliance with applicable institutional, ethical, and legal requirements. GFD should be regarded as a transcription-support method rather than a diagnostic, treatment-recommendation, or autonomous clinical-documentation system.

Fairness across speaker groups was not evaluated in the present experiments. Whisper-large-v3 and the external LLM backends may inherit representational biases from their pre-training data, including unequal coverage of Indonesian accents, regional language varieties, speaking styles, demographic groups, speech characteristics, medical specialties, and lexical conventions. Decoding-time fusion may correct some acoustically uncertain words while also favoring more frequent or standardized language forms. This behavior could disadvantage less represented pronunciations, personal names, regional expressions, speech impairments, or specialized terminology. Therefore, lower aggregate corpus-level or utterance-level WER should not be interpreted as evidence that errors are distributed equally across users or that the system is free from linguistic, demographic, or domain-related bias. Future work should include subgroup-level performance analysis, error-rate disparity assessment, examination of clinically consequential substitutions and omissions, and evaluation of the effects of fusion on underrepresented speakers and language varieties.

Taken together, these findings demonstrate that GFD behaves as a structured interface between acoustic evidence and language prior knowledge, where the magnitude of benefit is determined jointly by i) the strength of linguistic alignment, and ii) the match between r and domain signal characteristics. The general principle emerging is that ASR – LLM fusion for Indonesian is not simply a function of “more Indonesian tuning”, but depends on how linguistic density, domain specialization, and decoding-time weighting interact. This insight provides a principled foundation for future domain-sensitive decoding strategies in Indonesian ASR. At the same time, deployment in operational environments will require

additional validation of computational efficiency, acoustic robustness, and workflow integration beyond the offline evaluation presented in this study.

4. CONCLUSION

This study empirically evaluated GFD for Indonesian ASR across general read speech and medical dictation. GFD consistently reduced corpus-level WER relative to the unfused Whisper-large-v3 baseline, from 5.91% to 4.17% on FLEURS-ID and from 15.89% to 12.30% on Harkit, with statistically significant improvements in both conditions. The optimal static fusion weight was domain-dependent, with FLEURS-ID favoring $r = 0.75$ and Harkit $r = 0.50$, indicating that the balance between acoustic evidence and external linguistic priors is not universal across speech conditions. The results also show that no single LLM backend is uniformly optimal across domains and evaluation metrics, while the external LLM prior contributes more consistently than explicit instruction prompting, whose effect remains configuration-dependent. Overall, these findings demonstrate that GFD can improve Indonesian ASR without model retraining under the evaluated offline conditions, while also highlighting the importance of domain-sensitive fusion settings. The present results remain limited to controlled datasets and static fusion configurations and therefore should not be generalized to real-time, streaming, or broader clinical environments. Future work should investigate adaptive fusion, controlled comparisons with alternative fusion approaches, and deployment-oriented evaluation using larger and more diverse Indonesian speech datasets.

ACKNOWLEDGMENTS

The authors acknowledge institutional support from the National Research and Innovation Agency (BRIN) of the Republic of Indonesia.

FUNDING INFORMATION

This work was funded by the *Riset dan Inovasi untuk Indonesia Maju (RIIM) Kompetisi Batch-7* program, financed by the *Lembaga Pengelola Dana Pendidikan (LPDP)*, facilitated and organized by the National Research and Innovation Agency (BRIN), Republic of Indonesia, under contract no. B-5647/II.7.5/RT.01.01/4/2025 and contract no. B-6888/III.6/TK.01.03/4/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Asril Jarin	✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Lyla Ruslana Aini	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			
Agung Santosa	✓	✓	✓	✓	✓	✓	✓			✓		✓		
Gunarso				✓	✓	✓				✓				
Mohammad Teduh				✓	✓	✓				✓				
Uliniansyah														
Elvira Nurfadhilah				✓	✓	✓				✓	✓			
Siska Pebiana						✓		✓		✓				

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

DATA AVAILABILITY

The datasets supporting this study are available from their official repositories. FLEURS-ID is distributed as part of the Google FLEURS dataset at <https://huggingface.co/datasets/google/fleurs>. The Harkit medical-dictation resources are available through the BRIN RIN dataverse under the persistent identifiers hdl.handle.net/20.500.12690/RIN/OCIQHF and hdl.handle.net/20.500.12690/RIN/0DZTY8. The subsets used in this study were selected from these corpus sources according to the criteria and dataset characteristics described in subsection 2.2. The original GFD implementation is available at <https://github.com/mtkresearch/generative-fusion-decoding>, and Whisper-large-v3 is available at <https://huggingface.co/openai/whisper-large-v3>.

REFERENCES

- [1] G. Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.
- [2] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, “End-to-end speech recognition: a survey,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, pp. 325–351, 2024, doi: 10.1109/TASLP.2023.3328283.
- [3] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.
- [4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: a neural network for large vocabulary conversational speech recognition,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.
- [5] D. S. Park *et al.*, “SpecAugment: a simple data augmentation method for automatic speech recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2613–2617, doi: 10.21437/Interspeech.2019-2680.
- [6] S. Chen *et al.*, “WavLM: large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022, doi: 10.1109/JSTSP.2022.3188113.
- [7] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “HuBERT: self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.
- [8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pp. 12449–12460, 2020.
- [9] A. Gulati *et al.*, “Conformer: convolution-augmented transformer for speech recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, May 2020, pp. 5036–5040, doi: 10.21437/Interspeech.2020-3015.
- [10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *40th International Conference on Machine Learning*, pp. 28492–28518, Jul. 2023.
- [11] K. Denistia and R. H. Baayen, “The morphology of Indonesian: data and quantitative modeling,” in *The Routledge Handbook of Asian Linguistics*, Routledge: London, United Kingdom, 2022, pp. 605–634, doi: 10.4324/9781003090205-41.
- [12] D. S. Nugraha, “A morphological analysis of the Indonesian suffixation: a look at the different types of affixes and their semantic changes,” *GEMA Online Journal of Language Studies*, vol. 24, no. 4, pp. 109–132, 2024, doi: 10.17576/gema-2024-2404-07.
- [13] R. V. M. Tazakka, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, “Indonesian-English code-switching speech recognition using the machine speech chain based semi-supervised learning,” in *3rd Annual Meeting of the ELRA-ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2024 at LREC-COLING 2024*, 2024, pp. 143–148, doi: 10.63317/3hriruv4bj9u.
- [14] D. Setiawan, “English code switching in Indonesian language,” *Universal Journal of Educational Research*, vol. 4, no. 7, pp. 1545–1552, 2016, doi: 10.13189/ujer.2016.040707.
- [15] M. T. Agro, A. A. Kulkarni, K. Kadaoui, Z. Talat, and H. Aldarmaki, “Code-switching in end-to-end automatic speech recognition: a systematic literature review,” *Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pp. 9790–9812, 2026, doi: 10.63317/477upr9ikf9n.
- [16] P. Arisaputra, A. T. Handoyo, and A. Zahra, “XIS-R deep learning model for multilingual ASR on low-resource languages: Indonesian, Javanese, and Sundanese,” *ICIC Express Letters, Part B: Applications*, vol. 15, no. 6, pp. 551–559, 2024, doi: 10.24507/icicelb.15.06.551.
- [17] R. S. A. Pratama and A. Amrullah, “Analysis of Whisper automatic speech recognition performance on low-resource language,” *Jurnal Pilar Nusa Mandiri*, vol. 20, no. 1, pp. 1–8, 2024, doi: 10.33480/pilar.v20i1.4633.
- [18] A. Adila, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, “Enhancing Indonesian automatic speech recognition: evaluating multilingual models with diverse speech variabilities,” in *2024 27th Conference on the Oriental COCOSA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques, (O-COCOSA 2024)*, 2024, doi: 10.1109/O-COCOSA64382.2024.10800336.
- [19] A. Conneau *et al.*, “FLEURS: FEW-shot learning evaluation of universal representations of speech,” in *2022 IEEE Spoken Language Technology Workshop, (SLT 2022)*, May 2023, pp. 798–805, doi: 10.1109/SLT54892.2023.10023141.
- [20] V. Pratap *et al.*, “Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, vol. 2020-October, pp. 4751–4755, doi: 10.21437/Interspeech.2020-2831.
- [21] V. Pratap *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 4798–4849, May 2024, doi: 10.5555/3722577.3722674.
- [22] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5824–5828, doi: 10.1109/ICASSP.2018.8462682.
- [23] A. Sriram, H. Jun, S. Satheesh, and A. Coates, “Cold fusion: training Seq2seq models together with language models,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 387–391, doi: 10.21437/Interspeech.2018-1392.

- [24] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: end-to-end contextual speech recognition," in *2018 IEEE Spoken Language Technology Workshop, (SLT 2018)*, 2018, pp. 418–425, doi: 10.1109/SLT.2018.8639034.
- [25] L. Xu *et al.*, "RescoreBERT: discriminative speech recognition rescoring with BERT," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6117–6121, doi: 10.1109/ICASSP43922.2022.9747118.
- [26] A. Mittal, D. Prabhu, S. Sarawagi, and P. Jyothi, "SALSA: speedy ASR-LLM synchronous aggregation," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2024, pp. 3485–3489, doi: 10.21437/Interspeech.2024-2499.
- [27] C. J. Hsu *et al.*, "Let's fuse step by step: a generative fusion decoding algorithm with LLMs for robust and instruction-aware ASR and OCR," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 24959–24973, doi: 10.18653/v1/2025.findings-acl.1281.
- [28] R. Ng *et al.*, "SEA-LION: Southeast Asian languages in one network," *14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025, doi: 10.18653/v1/2025.ijcnlp-long.30.
- [29] W. Q. Leong, J. G. Ngui, Y. Susanto, H. Rengarajan, K. Sarveswaran, and W. C. Tjhi, "BHASA: a holistic Southeast Asian linguistic and cultural evaluation suite for large language models," 2023, *arXiv:2309.06085*.
- [30] A. Q. Jiang *et al.*, "Mistral 7B," 2023, *arXiv:2310.06825*.
- [31] E. Nurfadhilah *et al.*, "Evaluating the BPPT medical speech corpus for an ASR medical record transcription system," in *2021 9th International Conference on Information and Communication Technology, ICoICT 2021*, 2021, pp. 657–661, doi: 10.1109/ICoICT52021.2021.9527450.
- [32] Unicode Consortium, "Unicode Standard Annex #15: unicode normalization forms," *unicode.org*, 2025. Accessed: Oct. 30, 2025. [Online]. Available: <https://www.unicode.org/reports/tr15/>
- [33] D. Zhao *et al.*, "Shallow-fusion end-to-end contextual biasing," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1418–1422, doi: 10.21437/Interspeech.2019-1209.
- [34] D. Le *et al.*, "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, vol. 6, pp. 4086–4090, doi: 10.21437/Interspeech.2021-1566.

BIOGRAPHIES OF AUTHORS



Asril Jarin    holds a Doctor of Electrical Engineering from the Universitas Indonesia in 2017. He also received his Dipl.Ing. from GSO FH Nürnberg and M.Sc. from Hochschule Darmstadt, Germany, in 2001 and 2003, respectively. Presently, he serves as a senior researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2021, he has been an active researcher in artificial intelligence, specifically in speech and natural language processing. His research includes speech recognition, natural language processing, and large language models. He can be contacted at email: asri003@brin.go.id.



Lyla Ruslana Aini    earned a graduate degree in Computer Science from Universitas Indonesia in 2025. She is currently an associate researcher at the Research Center for Data and Information Sciences, part of the National Research and Innovation Agency (BRIN) in Indonesia. Her research interests include speech and natural language processing, with current projects focusing on ASR and retrieval-augmented generation in the medical and legal domains. She can be contacted at email: lyla001@brin.go.id.



Agung Santosa    received the Bachelor of Science in Electrical Engineering (B.S.E.E.) from the University of Toledo, United States, in 1992 and a Master of Computer Science from the Universitas Indonesia, Indonesia, in 2015. He is a senior researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2021, he has been an active researcher in artificial intelligence, specifically in speech and natural language processing. His research includes speech recognition, speech analytics, sentiment analysis, and natural language processing. He can be contacted at email: agun006@brin.go.id.



Gunarso     graduated in Electrical Engineering from Universitas Gadjah Mada, Indonesia in 1988. He is a senior researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2021, he has been an active researcher in artificial intelligence, specifically in speech and natural language processing. His research includes speech recognition, sentiment analysis, and natural language processing. He can be contacted at email: gunarso@brin.go.id.



Mohammad Teduh Uliniansyah     received a B.Eng. degree from Shibaura Institute of Technology, Japan, in 1991, a Master of Computer Science degree from Oklahoma State University, United States, in 1998, and a Ph.D. from Keio University, Japan, in 2007. He is a senior researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 1992, he has been an active researcher in artificial intelligence, specifically in speech and natural language processing. His research encompasses speech recognition and analysis, sentiment analysis, and natural language processing. He can be contacted at email: mtd001@brin.go.id.



Elvira Nurfadhilah     received the bachelor and master in Computer Science from IPB University, Indonesia, in 2011 and 2015. She is an associate researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2016, she has been an active researcher in artificial intelligence, specifically focusing on speech and natural language processing. Her research includes speech recognition and analysis, sentiment analysis, and natural language processing. She can be contacted at email: elvi003@brin.go.id.



Siska Pebiana     received the Bachelor of Information Engineering from Universitas Sriwijaya, Indonesia, in 2015 and the Master of Computer Science from Universitas Indonesia, Indonesia, in 2019. She is an associate researcher at the Research Center for Data and Information Sciences, affiliated with the National Research and Innovation Agency (BRIN) in Indonesia. Since 2019, she has been an active researcher in artificial intelligence, specifically focusing on speech and natural language processing. Her research includes natural language processing, speech recognition, sentiment analysis, and large language models. She can be contacted at email: siska.pebiana@brin.go.id.