WhisperX-LLM-TermFixer: Enhancing Whisper with LLM-Guided Post-Decoding for Domain-Specific Terminology

Anonymous ACL submission

Abstract

Large-scale ASR models such as Whisper perform well in general domains but struggle with specialized terminology, particularly in sensitive domains like medicine. We propose *WhisperX-LLM-TermFixer*, a lightweight post-decoding framework that improves term recognition without retraining or modifying the acoustic model. The method combines N-best decoding, phonetic n-gram correction, and LLM-based contextual filtering. We evaluate our approach on a multilingual synthetic dataset covering English, Spanish, Catalan, and Basque, using the same set of medical terms across languages. Results show consistent gains across all Whisper model sizes, with term-level accuracy improvements exceeding +30 percentage points and reductions in WER. Additionally, perplexity analysis indicates improved sentence-level coherence. Despite a moderate increase in inference time (1.4×–2.4×), the method achieves a favorable trade-off between accuracy, coherence, and efficiency, demonstrating the effectiveness of post-decoding strategies for domain-specific ASR. Our code and data are publicly available at `https://anonymous.4open.science/status/WhisperX-LLM-TermFixer-7F82`.

1 Introduction

Automatic Speech Recognition (ASR) systems have undergone significant advances in recent years, driven by End-To-End (E2E) neural architectures (Chorowski et al., 2015; Prabhavalkar et al., 2023), such as Whisper (Radford et al., 2023), Wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022). Among these systems, Whisper has emerged as the foundational model for many current tools and frameworks due to its strong performance across multiple languages, acoustic conditions, and tasks. Its open-vocabulary formulation weakly supervised training on large volumes of data enable robust recognition in general-domain scenarios without the need for explicit adaption.

Despite its strong overall performance, however, Whisper exhibits limitations in domain-specific scenarios. In specialized fields such as medicine, pharmacology, or technical discourse, recognition errors frequently occur on rare or morphologically complex terms that are poorly represented in Whisper's training data (Hwang et al., 2025). These errors typically manifest as phonetic confusion, partial matches, or substitutions with more frequent general-domain words, leading to semantically incorrect transcriptions and limiting the applicability of the system in a sensitive context (Lin et al., 2024; Ito et al., 2005). In particular, ASR models often struggle to recognize proper nouns and entities that occur infrequently in the training data. To address this issue, numerous studies have proposed lexical and contextual adaptation mechanisms.

A common approach involves constructing lists of *hot words* or domain-specific entity vocabularies and incorporating them during decoding through shallow fusion (Gourav et al., 2021; Zhao et al., 2019) or deep fusion techniques (Chang et al., 2021; Han et al., 2022; Munkhdalai et al., 2022). Shallow fusion boosts the scores of hot words during beam search decoding, although it requires careful tuning of the fusion weight to avoid degrading overall performance. In contrast, deep fusion introduces a contextual encoder that is jointly trained with the ASR model from scratch, mapping both entity words and the speech signal into a shared feature space, thereby enabling the decoder to leverage both contextual and acoustic information. To adapt existing ASR models, other approaches employ adapters (Sathyendra et al., 2022) to modify intermediate representations or pointer networks to revise output distributions in favor of domain-specific entities (Sun et al., 2023). Moreover, combining shallow and deep fusion strategies has been shown to further enhance recognition performance

in strongly biased scenarios (Le et al., 2021).

In the context of Whisper-based systems, some works integrate keyword spotting modules or auxiliary signals during decoding to explicitly bias the recognition of rare terms (Li et al., 2024). Other lines of research explore the integration of external language models via techniques such as shallow fusion or hypothesis re-scoring, providing additional linguistic knowledge beyond the acoustic signal, which is particularly beneficial in low-resource settings (de Zuazo et al., 2025).

The existing solutions require, however, supervised adaptation of Whisper through fine-tuning or introduce modifications to the model architecture or decoding process, increasing computational requirements and deployment costs in Whisper-based frameworks. Furthermore, correction approaches relying solely on direct string or phonetic similarity replacements often fail for multi-word terms and may introduce contextual inconsistencies, including incorrect partial substitutions or duplications.

In this paper, we propose a lightweight and modular alternative to improve the recognition of specialized terminology in Whisper-based ASR systems without requiring retraining or intrusive modifications to the acoustic model. Our approach is implemented as an extension of WhisperX (Bain et al., 2023), a widely used framework for practical Whisper deployment, and operates exclusively at the hypothesis generation and post-processing stages. It combines multiple decoding hypotheses, phonetic- and text-based lexical correction mechanisms, and contextual re-scoring using a Large Language Model (LLM), aiming to enhance linguistic coherence and accurate recovery of domain-specific terms while preserving the overall system performance.

We evaluate our proposed method, **WhisperX-LLM-TermFixer**, using a diverse set of LLMs, including Gemma-3 (1B and 4B) (Kamath et al., 2025), Gemma-2 (9B) (Riviere et al., 2024), BioMistral-7B and its variants (DARE and SLERP) (Labrak et al., 2024), LLaMA-3.1 (8B) (Grattafiori et al., 2024), LLaMA-3.2 (1B and 3B) (Grattafiori et al., 2024), and Salamandra (2B and 7B) (Gonzalez-Agirre et al., 2025), across multiple languages: English, Spanish, Catalan, and Basque. Our experiments focus on domain-specific evaluation, selecting the medical domain as a representative case study, with an emphasis on the recognition and correction of medical terminology in a multilingual setting.

Our contributions can be summarized as follows:

1. We propose **WhisperX-LLM-TermFixer**, a lightweight and modular framework for improving domain-specific term recognition in Whisper-based ASR systems without requiring model retraining or architectural modifications.

2. We introduce a hybrid correction strategy that combines N-best hypothesis re-ranking, phonetic and textual similarity matching, and LLM-based contextual re-scoring to ensure both accurate term recovery and sentence-level coherence.

3. We conduct a comprehensive multilingual evaluation across English, Spanish, Catalan, and Basque using a controlled synthetic dataset of medical terminology, analyzing the impact of different LLMs and demonstrating improvements over baseline WhisperX.

2 Methodology

2.1 System overview

As illustrated in Figure 1, the proposed approach consists of three key components designed to improve the recognition of specialized terminology without modifying Whisper's architecture or performing domain adaption through fine-tuning.

First, WhisperX is used to perform the initial decoding of the input audio into speech segments using default Voice Activity Detection (VAD). Second, we extend and modify WhisperX to generate multiple decoding hypotheses per segments via beam search, enabling the capture of acoustically plausible variants that may contain partially recognized domain-specific terms. Among these hypotheses, the one that maximizes the presence of terms from the specialized lexicon is selected, using the number of lexical matches as the primary criterion and, in the case of ties, the decoding score produced by the Whisper model. Next, a lexical detection and correction module based on N-gram phonetic and textual similarity is applied, enabling the identification of specialized terms even when they appear fragmented or acoustically distorted. Finally, a contextual re-scoring step is introduced using a LLM, which sequentially evaluates candidate replacements and selects the globally most coherent option at the sentence level. This step incorporates domain-specific constraints to prevent partial replacements and duplicate insertions.

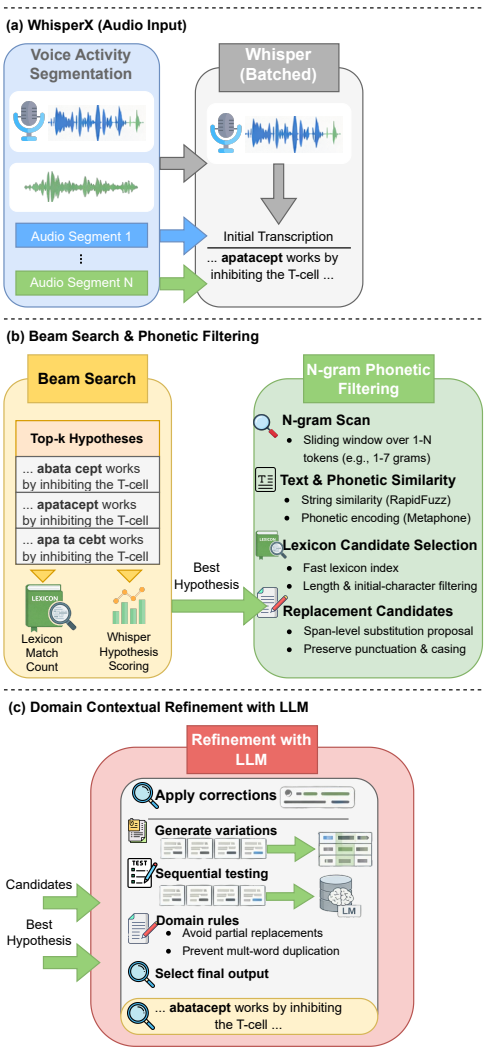

Figure 1: System architecture: (a) Audio inputs are converted into an initial transcription using WhisperX; (b) beam search and phonetic filtering are applied to the initial transcription to generate a set of lexicon-based candidate hypotheses; and (c) domain contextual refinement with a LLM is performed over the candidate hypotheses to select the most plausible final transcription.

2.2 WhisperX

WhisperX is used as the inference backbone due to its efficient segmentation and decoding capabilities. Given an input audio signal, the system applies WhisperX's default VAD to identify speech regions and construct temporal segments. Each segment is then decoded with Whisper to produce segment-level textual hypotheses.

2.3 Beam search and phonetic filtering

By default, WhisperX, when used with FastWhisper, does not support beam search, limiting the generation of alternative decoding hypotheses for each audio segment. To address this limitation, we modify the WhisperX architecture to enable beam search decoding, producing multiple hypotheses per segment and returning an N-best list along with their corresponding decoding scores. This extension allows the system to capture acoustically plausible variants, in which domain-specific terms may be partially recognized in some hypotheses while being absent in others.

2.3.1 Lexicon-guided hypothesis selection

Let $\{h_k\}_{k=1}^{K}$ denote the set of decoding hypotheses for a given segment, and let $\mathcal{L}$ be a normalized domain-specific lexicon. For each hypothesis h_k, we compute the number of lexical matches as follows:

$$c(h_k) = \sum_{w \in \text{words}(h_k)} \mathbf{I}[w \in \mathcal{L}],$$

where $\text{words}(\cdot)$ denotes word-level tokenization and $\mathcal{L}$ is the indicator function. The selected hypothesis is then defined as:

$$h^* = \arg\max_{k} \big(c(h_k),\, s(h_k)\big),$$

where $s(h_k)$ is the decoding score produced by Whisper model and is used as a tie-breaking criterion. This selection strategy prioritizes hypotheses that contain a higher number of lexicon terms while still accounting for the acoustic confidence of the underlying ASR model.

2.3.2 N-gram-based phonetic filtering

After selecting h^*, a detection and correction module is applied that explores word n–grams of length $n \leq N$ over the hypothesis text. For each candidate n-gram, a combined similarity score is computed between the recognized fragment and a lexicon term, integrating (i) textual similarity and (ii) phonetic similarity using *metaphone* encoding and fuzzy matching.

To ensure efficiency with large lexicons, the system employs an index based on the first character and approximate string length to restrict the search space. In addition, a length-dependent dynamic threshold is applied to reduce false positives, with stricter thresholds for shorter strings. The module also incorporates heuristic rules to avoid trivial or

harmful corrections, such as preventing replacements when the lexicon term already appears as a subtoken within the candidate n-gram.

2.4 Domain contextual refinement with LLM

Although phonetic and textual correction improves the recognition of specialized or rare terms, it may introduce ambiguity when multiple similar candidates exist or when substitutions affect sentence-level coherence. To mitigate this effect, we introduce a contextual re-scoring module based on a LLM, which evaluates the global plausibility of a sentence after applying a replacement.

Given a candidate sentence y, obtained by applying a substitution over a *span*, we compare a cost $J(y)$ based on the LLM's negative log-likelihood. A correction is accepted only if it yields a sufficient improvement over the current sentence:

$$\Delta = J(y_{\text{old}}) - J(y_{\text{new}}) \geq \tau,$$

where τ denotes a minimum improvement threshold. This criterion prevents overly aggressive corrections that increase lexical matching but degrade fluency or semantic coherence, potentially leading to higher Word Error Rate (WER).

Unlike re-scoring strategies that evaluate only full-span replacement, our method performs sequential re-scoring by considering multiple replacement variants for the same lexical candidate. Specifically, it evaluates incremental replacements over span prefixes (e.g., `[start:start+1]`, `[start:start+2]`, ..., `[start:end]`). This strategy enables the selection of the variant that best preserves contextual coherence, without enforcing unnecessary substitutions when parts of the recognized fragment are already correct. In addition, domain-specific constraints are incorporated to prevent common errors in multi-word terms, such as partial replacements or duplications. In particular, substitutions are discarded if replacing a prefix would leave an immediate residual token to the right that belongs to the target term "(e.g., retaining '*fructu*' after replacing it with '*agni casti fructus*' , thereby yielding '*agni casti fructus fructus*')", as such cases typically result in duplicated insertions or incomplete corrections. Among the valid candidates, the variant with the lowest LLM loss is selected, provided that it exceeds the improvement threshold τ.

Table 1: Summary statistics of the synthetic multilingual medical speech dataset. Each language contains 1,005 utterances with domain-specific medical terminology.

Language	Total / Words	Words / Utterance	Avg / Duration (s)
English	16,190	16.10	8.10
Spanish	19,579	19.48	7.75
Catalan	19,323	19.23	9.47
Basque	13,123	13.06	9.36

3 Experimental Setup

3.1 Dataset

To conduct a controlled and exhaustive evaluation of medical terminology recognition, we construct a synthetic speech dataset comprising 1,005 domain-specific terms per language, including English, Spanish, and lower-resource languages such as Catalan and Basque, drawn from the medical and pharmaceutical domains and extracted from the *Centro de Información online de Medicamentos de la AEMPS (CIMA)*[1]. The use of a synthetic dataset allows precise control over pronunciation quality, lexical coverage, and speaker variability, enabling a focused analysis of term detection error independently of acoustic noise or disfluencies. For this purpose, contextualized medical sentences containing the target terms are automatically generated using Gemini (Team et al., 2023). The corresponding speech signals are synthesized using multiple Text-To-Speech (TTS) systems. In particular, for Spanish, Catalan, and Basque, we use the Microsoft Text-to-Speech API, which provides high-quality neural voices with robust support for these languages, ensuring accurate phonetic realization and consistent prosody. For English, speech is generated with `Coqui-TTS`[2] using multiple speakers, ensuring clearly articulated and consistently pronounced medical terminology.

Table 1 summarizes the main statistics of the constructed synthetic dataset. Each language contains 1,005 utterances built from the same set of medical terms, ensuring a controlled and comparable evaluation across English, Spanish, Catalan, and Basque. We observe notable differences in lexical density and duration across languages. In particular, Spanish and Catalan exhibit higher average word count per utterance, while Basque shows shorter utterances in terms of words count but comparable audio durations, reflecting language-specific morphologi-

cal and syntactic properties.

3.2 Parameter Settings

All experiments are conducted using WhisperX with the Faster-Whisper backend, running on GPU when available (`float16`, batch size 16). Audio is segmented using VAD (`vad_onset=0.15`, `vad_offset=0.20`) and decoded in English with a maximum chunk size of 30s. Beam search is enabled with `beam_size=10`, generating an N-best list of `num_hypotheses=10` per segment; the selected hypothesis maximizes the number of matches against the medical lexicon, with Whispers decoding score used as a tie-breaker. After hypothesis selection, a phonetic n-gram correction module is applied (up to 7-grams) using combined fuzzy textual and metaphone-based phonetic similarity with a base threshold of 85. Finally, LLM-based contextual re-scoring is performed using a LLM; a replacement is accepted only if it improves sentence likelihood by at least `lm_threshold=0.01`, ensuring conservative, contextually consistent corrections. A detailed description of all configurable parameters is provided in Appendix B.

3.3 Evaluation Metrics

We evaluate the proposed approach using WER and term-level accuracy, which capture complementary aspects of recognition performance. WER measures overall transcription quality and reflects the balance between substitutions, deletions, and insertions at the word level.

Since this work focuses on the recovery of specialized medical terminology, we additionally report term accuracy, defined as the percentage of target medical terms correctly recognized in the transcription. A term is considered correct only when the full term, including multi-word expressions, is accurately recovered. Reporting both metrics allows us to assess global ASR performance while explicitly measuring improvements in domain-specific term recognition.

4 Experimental Results

4.1 Main Results

Table 2 reports the best-performing LLM within each model family (Gemma, BioMistral, LLaMA, and Salamandra) for each Whisper configuration and language. Overall, all LLM families yield consistent improvements over the baseline across all languages and model sizes, indicating that the proposed post-decoding framework is robust to the choice of LLM. While the performance differences between families are relatively small, Salamandra models frequently achieve the best results, especially for smaller Whisper models and in lower-resource languages such as Catalan and Basque. This trend indicates that compact or specialized LLMs can effectively leverage contextual information for terminology correction without requiring large-scale models.

4.2 Multilingual Analysis

As shown in Table 3, improvements are consistently observed across all languages and Whisper model sizes, with particularly large gains in term-level accuracy. Across languages, we observe that improvements are more pronounced when the baseline performance is lower. In particular, Catalan and Basque exhibit substantially higher WER and lower term accuracy in the baseline setting (e.g., Catalan Tiny: 6.57% accuracy; Basque Tiny: 6.67%), leaving greater room for improvement. In these scenarios, the proposed method yields substantial gains, increasing term-level accuracy by more than +30 percentage points in many cases while also reducing WER by up to 4 points.

In contrast, English and Spanish present stronger baseline performance, resulting in more moderate but still consistent improvements. For English, gains in term-level accuracy typically range between +22 and +30 percentage points, while maintaining steady reductions in WER across all model sizes. This indicates that even when the ASR system already performs well, the proposed method can effectively refine domain-specific terminology without degrading overall transcription quality. Similarly, Spanish achieves some of the highest absolute accuracies after correction (e.g., up to 93.83%), with improvements exceeding +35 percentage points for smaller models. This highlights the ability of the approach to enhance already strong ASR systems by correcting specialized terminology.

These results indicate that the proposed approach is especially effective in low-resource and linguistically challenging settings. In Basque, a morphologically rich and agglutinative language, the method consistently improves term recovery across all model sizes (e.g., +30.85 accuracy gain for

[1] https://cima.aemps.es/cima/publico/home.html
[2] https://github.com/coqui-ai/TTS

Table 2: Best-performing LLM within each model family (Gemma, BioMistral, LLaMA, and Salamandra) for each Whisper model size and language. Results are reported in terms of WER and term-level accuracy (%). Best results per row are highlighted in bold. *Large-t* denotes the *large-v3-turbo* model, and *Large* denotes the *large-v3* model.

Whisper	**Gemma**		**BioMistral**		**LLaMA**		**Salamandra**	
	WER	Acc	WER	Acc	WER	Acc	WER	Acc
English								
Tiny	16.85	35.52	16.80	36.12	16.81	36.12	**16.79**	**36.22**
Base	11.08	49.85	**11.02**	**49.95**	11.05	49.95	11.04	49.95
Small	7.22	62.29	7.21	62.09	7.23	62.19	**7.17**	**62.59**
Medium	4.97	70.55	4.97	70.55	4.99	70.45	**4.93**	**70.95**
Large-t	4.61	72.74	4.61	72.74	4.55	73.73	**4.54**	**73.83**
Large	3.61	78.81	3.59	79.00	3.61	79.00	**3.55**	**79.20**
Spanish								
Tiny	14.57	67.06	14.54	67.06	14.58	66.87	**14.63**	**67.16**
Base	8.77	77.81	**8.72**	**77.91**	8.75	77.61	8.73	77.81
Small	**4.46**	**85.97**	4.52	85.47	4.46	85.87	4.51	85.67
Medium	**2.50**	**90.95**	2.50	90.95	2.50	90.85	2.51	90.95
Large-t	**1.85**	**93.63**	1.86	93.23	1.87	93.33	1.87	93.23
Large	1.81	93.73	1.79	93.73	**1.81**	**93.83**	**1.81**	**93.83**
Catalan								
Tiny	41.78	30.75	41.79	30.75	41.75	31.24	**41.71**	**31.34**
Base	27.91	39.30	28.01	38.91	27.94	39.20	**27.78**	**39.80**
Small	15.66	50.05	15.62	49.85	15.63	50.45	**15.51**	**50.65**
Medium	9.55	60.00	9.59	59.30	**9.54**	**60.20**	9.50	60.10
Large-t	6.85	64.58	6.95	63.78	6.83	65.07	**6.80**	**65.17**
Large	6.28	70.05	6.30	70.05	6.30	70.15	**6.22**	**70.35**
Basque								
Tiny	94.18	31.64	94.38	33.03	94.15	33.33	**93.76**	**34.93**
Base	87.72	36.62	87.80	39.50	87.63	40.00	**87.32**	**41.89**
Small	67.92	42.69	68.04	43.88	67.59	45.87	**67.47**	**46.27**
Medium	53.44	45.57	53.82	45.97	53.21	46.87	**53.05**	**47.76**
Large-t	32.59	59.40	33.01	58.81	32.62	58.81	**32.44**	**60.60**
Large	34.21	54.53	34.47	55.82	34.12	55.52	**33.95**	**56.02**

Small), demonstrating its ability to recover complex or infrequent terms that are typically misrecognized by the base ASR system. Similarly, in Catalan, which exhibits moderate resource availability and linguistic variation, large improvements are observed across all configurations, confirming the robustness of the approach in handling domain-specific terminology under varying linguistic conditions.

From a model scaling perspective, the gains remain stable across different Whisper model sizes. As shown in Table 3, smaller models benefit from larger relative improvements, particularly in low-resource settings (e.g., Spanish Tiny: +35.22 accuracy), while larger models also show consistent gains despite their already strong baseline performance (e.g., English Large: +23.58 accuracy). This indicates that the proposed method is complementary to model scaling and can enhance both low- and high-capacity ASR systems.

4.3 Impact of LLM

Domain-specific lexicons, particularly in specialized domains such as medicine, often contain multi-word expressions and morphologically complex terms. In these scenarios, relying solely on phonetic-based replacement can lead to incorrect substitutions. While phonetic similarity is effective for recovering partially recognized terms, it may also introduce errors by replacing acoustically similar but contextually incorrect tokens, especially in multi-word expressions.

Our N-gram-based phonetic correction enables the recovery of longer terms, but may still produce errors due to overlapping span or partial replacements that disrupt context. To address this, we incorporate an LLM-based validation step that filters candidate corrections based on sentence-level coherence, accepting only those that improve overall plausibility.

Table 4 illustrates representative examples of this

Table 3: Comparison between baseline WhisperX and the proposed WhisperX-LLM-TermFixer using the best-performing LLM per configuration. Results are reported for each language and Whisper model size in terms of WER and term-level accuracy (%), along with absolute improvements (Δ).

Whisper	Baseline		WhisperX-LLM-TermFixer		Δ	
	WER	Acc	WER	Acc	WER ↓	Acc ↑
English						
Tiny	18.91	14.23	16.79	36.22	**-2.12**	**+21.99**
Base	13.66	19.90	11.02	49.95	**-2.64**	**+30.05**
Small	9.41	33.33	7.17	62.59	**-2.24**	**+29.26**
Medium	6.88	43.28	4.93	70.95	**-1.95**	**+27.67**
Large-t	6.50	44.88	4.54	73.83	**-1.96**	**+28.95**
Large	5.24	55.62	3.55	79.20	**-1.69**	**+23.58**
Spanish						
Tiny	17.61	31.94	14.63	67.16	**-2.98**	**+35.22**
Base	11.53	42.59	8.72	77.91	**-2.81**	**+35.32**
Small	6.54	56.82	4.46	85.97	**-2.08**	**+29.15**
Medium	3.74	70.85	2.50	90.95	**-1.24**	**+20.10**
Large-t	2.85	76.42	1.85	93.63	**-1.00**	**+17.21**
Large	2.69	77.71	1.81	93.83	**-0.88**	**+16.12**
Catalan						
Tiny	43.91	6.57	41.71	31.34	**-2.20**	**+24.77**
Base	30.54	11.64	27.78	39.80	**-2.76**	**+28.16**
Small	18.02	20.20	15.51	50.65	**-2.51**	**+30.45**
Medium	11.91	26.07	9.54	60.20	**-2.37**	**+34.13**
Large-t	9.03	32.84	6.80	65.17	**-2.23**	**+32.33**
Large	8.53	37.01	6.22	70.35	**-2.31**	**+33.34**
Basque						
Tiny	97.32	6.67	93.76	34.93	**-3.56**	**+28.26**
Base	90.99	12.54	87.32	41.89	**-3.67**	**+29.35**
Small	71.24	15.42	67.47	46.27	**-3.77**	**+30.85**
Medium	57.06	17.51	53.05	47.76	**-4.01**	**+30.25**
Large-t	35.28	31.24	32.44	60.60	**-2.84**	**+29.36**
Large	36.61	28.26	33.95	56.02	**-2.66**	**+27.76**

behavior. Phonetic matching alone may introduce incorrect, incomplete, or duplicated substitutions, particularly in multi-word terms. In contrast, LLM-based filtering successfully removes inconsistent candidates and preserves coherent and semantically valid outputs.

In addition to WER and term-level accuracy, we evaluate sentence-level coherence using perplexity, as reported in Table 5. Perplexity is computed using a causal language model as a proxy for fluency and contextual plausibility. Specifically, we use *GPT-oss-20B* (OpenAI, 2025) for English, Spanish, and Catalan, and *latxa-13b-v1.2* (Etxaniz et al., 2024) for Basque, which act as evaluation models to score sentence-level likelihood, ensuring appropriate multilingual coverage.

Results show a consistent reduction in perplexity from the baseline to the phonetic-only configuration, and further improvements when applying the full *WhisperX-LLM-TermFixer* approach. This trend is particularly evident in English, Spanish, and Catalan, where LLM-based filtering systematically achieves the lowest perplexity across all model sizes. These results indicate that, beyond improving lexical accuracy, the proposed method enhances sentence-level fluency and contextual coherence.

Table 4: Examples of phonetic corrections and LLM-based filtering. The LLM prevents incorrect or redundant substitutions and preserves contextual coherence.

ASR Output	Phonetic Correction	LLM-Filtered Output
The herbalist recommended AgniCosti fructus to help alleviate the symptoms of premenstrual syndrome.	"AgniCosti fructus to" → "agni casti fructus"	The herbalist recommended **agni casti fructus** to help alleviate the symptoms of premenstrual syndrome.
Anhydrocitric acid is used as an acidifying agent in various effervescent tablet formulations.	"'Anhydrocitric acid is" → "anhydrous citric acid"	**Anhydrous citric acid** is used as an acidifying agent in various effervescent tablet formulations.
Apriclonidine I drops are used to control or prevent elevations in intraocular pressure after laser surgery.	"Apriclonidine I" → "apraclonidine"	**Apraclonidine** I drops are used to control or prevent elevations in intraocular pressure after laser surgery.

While phonetic correction alone already reduces perplexity by correcting acoustically distorted tokens, it may still introduce contextually inconsistent substitutions. The LLM-based filtering stage refines these candidates by rejecting replacements that degrade overall sentence plausibility. As a result, the final system achieves a better balance between lexical recovery and linguistic coherence.

In Basque, improvements in perplexity are more limited and, in some cases, phonetic-only correction slightly outperforms the LLM-filtered variant. This behavior can be attributed to the increased morphological complexity of the language and the relatively lower coverage of the language model, which may lead to more conservative filtering decisions. Nevertheless, the overall trend still shows that phonetic correction substantially improves fluency over the baseline, while the LLM component prioritizes contextual consistency over aggressive substitutions.

4.4 Efficiency Analysis

We evaluate the computational overhead of the proposed *WhisperX-LLM-TermFixer* in terms of average inference time per audio segment. Figure 2 shows the relative latency compared to the baseline WhisperX system.

Table 5: Perplexity comparison across Whisper model sizes for different configurations. *Base* denotes the baseline WhisperX system, *Phon* corresponds to phonetic-only correction, and *LLM* refers to the full *WhisperX-LLM-TermFixer* approach. Lower values indicate better fluency and contextual coherence.

Model	**English**			**Spanish**			**Catalan**			**Basque**		
	Base	Phon	LLM	Base	Phon	LLM	Base	Phon	LLM	Base	Phon	LLM
Tiny	123.33	108.21	**106.26**	85.69	68.76	**66.99**	282.89	249.31	**248.58**	344.22	**305.65**	306.85
Base	84.05	67.93	**66.01**	50.52	38.59	**37.26**	158.90	137.26	**136.96**	267.11	230.57	**229.93**
Small	50.12	40.54	**38.85**	28.53	24.10	**23.18**	76.18	66.39	**66.09**	112.60	**98.77**	99.58
Medium	34.57	28.59	**26.94**	20.26	18.60	**17.46**	52.39	44.98	**44.67**	65.03	**56.97**	57.32
Large-t	32.09	27.27	**25.75**	18.03	17.09	**15.92**	40.81	35.73	**35.03**	28.99	**26.30**	26.36
Large	26.32	22.53	**20.95**	17.39	16.67	**15.53**	38.27	33.29	**32.64**	29.49	26.96	**26.86**

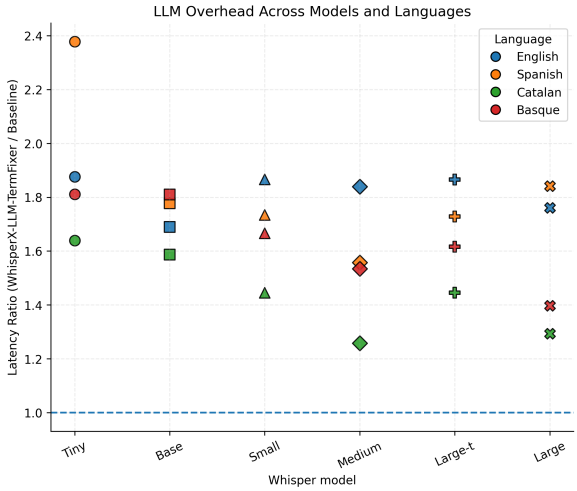

Figure 2: Relative inference cost (latency ratio) of the *WhisperX-LLM-TermFixer* compared to the baseline WhisperX system across languages and Whisper model sizes. Each point represents an independent configuration. Values above 1 indicate increased latency due to LLM-based filtering.

Results indicate a consistent overhead across all configurations, typically ranging between 1.4× and 2.4×. This increase is expected due to the additional LLM-based validation step. The overhead remains relatively stable across model sizes, indicating that the proposed method scales proportionally with the underlying ASR model.

We also observe some variation across languages, with higher overhead in Spanish for smaller models and lower ratios in Catalan. Despite this increase, inference remains within practical limits, and the additional cost is offset by the significant improvements in term accuracy and coherence observed in previous sections.

5 Conclusions

In this work, we proposed *WhisperX-LLM-TermFixer*, a lightweight and modular framework to improve the recognition of specialized terminology in Whisper-based ASR systems without requiring retraining or architectural modifications. The approach combines N-best decoding, phonetic n-gram correction, and LLM-based contextual filtering to enhance domain-specific term recovery while preserving overall transcription quality.

Experimental results across four languages (English, Spanish, Catalan, and Basque) demonstrate consistent improvements in term-level accuracy, with gains exceeding +30 percentage points in many cases, alongside systematic reductions in WER. The proposed method is particularly effective in low-resource and morphologically complex languages, where baseline performance is weaker and terminology recognition is more challenging.

In addition to accuracy improvements, we showed that LLM-based filtering enhances sentence-level coherence, as reflected by reductions in perplexity across most configurations. While phonetic correction increases recall by aggressively recovering candidate terms, the LLM acts as a precision-oriented filter that prevents contextually implausible substitutions, leading to more fluent and consistent transcriptions.

Finally, we analyzed the computational cost of the proposed approach and observed a moderate increase in inference time, typically between 1.4× and 2.4× relative to the baseline. This overhead remains stable across model sizes and languages, resulting in a favorable trade-off between accuracy, coherence, and efficiency.

Overall, our results show that post-decoding strategies combining phonetic matching and LLM-based validation provide an effective and practical solution for improving domain-specific ASR, and can be readily extended to other specialized domains without requiring model retraining.

Limitations

Despite the consistent improvements observed across languages and model sizes, our approach presents three main limitations.

First, the evaluation is conducted on a synthetic dataset specifically designed to isolate term recognition performance. This controlled setup allows us to analyze the behavior of Whisper on domain-specific terminology without the influence of external factors such as background noise, speaker variability, or disfluencies. However, this also means that the results do not directly reflect performance in real-world conditions, where such factors may interact with lexical errors.

Second, the method relies on the availability of a predefined domain-specific lexicon. While this enables targeted correction of specialized terminology, it limits applicability in open-domain settings or in scenarios where comprehensive lexicons are not available. Extending the approach to dynamically extract or update lexicons remains an open challenge.

Third, while perplexity is used as a proxy for sentence-level coherence, it depends on the underlying language model and may not fully capture semantic correctness, particularly in specialized domains.

Ethical Considerations

The aim of this work is to improve the recognition of specialised medical terminology in ASR systems, which could help to reduce transcription errors in sensitive areas. However, since incorrect corrections could still introduce clinically relevant mistakes, the proposed method should be considered an assistive tool rather than a fully autonomous clinical system. Our experiments were conducted using fully synthetic speech generated from publicly available medical terminology, with no patient data or personal information used, thereby minimising privacy concerns. Nevertheless, further evaluation of real-world clinical speech under appropriate ethical safeguards is necessary.

Acknowledgments

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech 2023*, pages 4489–4493.

Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. 2021. Context-aware transformer transducer for speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 503–510. IEEE.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518.

Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 577–585.

Xabier de Zuazo, Eva Navas, Ibon Saratxaga, and Inma Hernáez Rioja. 2025. Whisper-LM: Improving ASR Models with Language Models for Low-Resource Languages. *Preprint*, arXiv:2503.23542.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An Open Language Model and Evaluation Suite for Basque. *Preprint*, arXiv:2403.20266.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 4 others. 2025. Salamandra Technical Report. *Preprint*, arXiv:2502.08489.

Aditya Gourav, Linda Liu, Ankur Gandhe, Yile Gu, Guitang Lan, Xiangyang Huang, Shashank Kalmane, Gautam Tiwari, Denis Filimonov, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2021. Personalization strategies for end-to-end speech recognition systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7348–7352. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten,

Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Minglun Han, Linhao Dong, Zhenlin Liang, Meng Cai, Shiyu Zhou, Zejun Ma, and Bo Xu. 2022. Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8532–8536. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Haeeul Hwang, Eric Jordan, Deok-Hee Kim-Dufor, Christophe Lemey, and Motasem Alrahabi. 2025. Evaluating ASR in a Clinical Context : What Whisper Misses. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 374–378, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Taisuke Ito, Kazuya Takeda, and Fumitada Itakura. 2005. Analysis and recognition of whispered speech. *Speech Communication*, 45(2):139–152.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. Gemma 3 Technical Report. *ArXiv*, abs/2503.19786.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *Preprint*, arXiv:2402.10373.

Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L. Seltzer. 2021. Deep shallow fusion for RNN-T personalization. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 251–257. IEEE.

Yuang Li, Min Zhang, Chang Su, Yinglu Li, Xiaosong Qiao, Mengxin Ren, Miaomiao Ma, Daimeng Wei, Shimin Tao, and Hao Yang. 2024. A multitask training approach to enhance whisper with open-vocabulary keyword spotting. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Yeh-Sheng Lin, Shu-Chuan Tseng, and Jyh-Shing Roger Jang. 2024. Leveraging phonemic transcription and whisper toward clinically significant indices for automatic child speech assessment. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA.

Tsendsuren Munkhdalai, Khe Chai Sim, Angad Chandorkar, Fan Gao, Mason Chua, Trevor Strohman, and Françoise Beaufays. 2022. Fast contextual adaptation with neural associative memory for on-device personalized speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6632–6636. IEEE.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-End Speech Recognition: A Survey. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:325351.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 28492–28518. PMLR.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 176 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *ArXiv*, abs/2408.00118.

Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8537–8541. IEEE.

Guangzhi Sun, Chao Zhang, and Philip C. Woodland. 2023. Minimising biasing word errors for contextual ASR with the tree-constrained pointer generator. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:345–354.

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

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-fusion end-to-end contextual biasing. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz,*

Austria, September 15-19, 2019, pages 1418–1422.
ISCA.

A Synthetic Dataset Used for Evaluation

A.1 Dataset example

The medical terminology used in the dataset was extracted from the *Centro de Información online de Medicamentos de la AEMPS (CIMA)* [3], which provides official information on pharmaceutical compounds and active ingredients. The selected terms include both single-word entities (e.g., drug names) and multi-word expressions (e.g., compound formulations or Latin-derived terminology), which are known to be challenging for ASR systems.

Special emphasis was placed on including morphologically complex and phonetically ambiguous terms, as well as low-frequency entities that are typically underrepresented in ASR training data. The same set of 1,005 terms was adapted across all languages to ensure consistent cross-lingual evaluation.

Table 6: Examples of a medical term with its cross-lingual variants and contextualized sentences.

Term (ES)	**Translations**	**Example Contexts**
Adalimumab	EN: Adalimumab CA: Adalimumab EU: Adalimumab	(ES) El gastroenterólogo inició la terapia con adalimumab para el paciente con enfermedad de Crohn moderada a grave. (EN) The gastroenterologist initiated adalimumab therapy for the patient with moderate-to-severe Crohn's disease. (CA) El gastroenteròleg va iniciar la teràpia amb adalimumab per al pacient amb malaltia de Crohn moderada a greu. (EU) Gastroenterologoak adalimumab terapia hasi zuen Crohn gaixotasun moderatua edo larria duen pazientearentzat.
Alfa-1-antitripsina	EN: Alpha-1 antitrypsinb CA: Alfa-1-antitripsina EU: Alfa-1 antitripsina	(ES) Los resultados del laboratorio confirmaron una deficiencia de alfa-1 antitripsina, lo que explica el enfisema de inicio precoz del paciente. (EN) The lab results confirmed a deficiency in alpha-1 antitrypsin, explaining the patient's early-onset emphysema. (CA) Els resultats del laboratori van confirmar una deficiència d'alfa-1 antitripsina, explicant l'emfisema d'aparició precoç del pacient. (EU) Laborategiko emaitzek alfa-1 antitripsinaren gabezia baieztatu zuten, pazientearen enfisema goiztiarra azalduz.
Etinilestradiol	EN: Ethinylestradiol CA: Etinilestradiol EU: Etinilestradioa	(ES) La mayoría de los anticonceptivos orales combinados contienen etinilestradiol como el componente estrogénico principal. (EN) Most combined oral contraceptives contain ethinylestradiol as the primary estrogenic component. (CA) La majoria dels anticonceptius orals combinats contenen etinilestradiol com a component estrogènic principal. (EU) Ahozko antisorgailu konbinatu gehienek etinilestradioa dute osagai estrogeniko nagusi gisa.

B Detailed Parameter Settings

This section provides a detailed overview of the configurable parameters used in our approach. Table 7 summarizes the key components involved in lexicon-aware decoding, phonetic correction, and language model re-scoring.

[3] https://cima.aemps.es/cima/publico/home.html

Table 7: Configuration parameters of the proposed system, including lexicon-aware decoding and LLM-based re-scoring.

Parameter	Type	Default	Description
`lexicon_terms`	List[str]	None	List of domain-specific terms used for lexicon-aware re-ranking and phonetic correction. Lowercased format is recommended.
`num_hypotheses`	int	1	Number of N-best hypotheses generated per segment. Higher values enable more effective lexicon-based re-ranking at the cost of increased computation.
`use_lm`	bool	False	Enables contextual re-scoring using a causal language model to validate phonetic replacements.
`lm_name`	str	None	Path or Hugging Face model identifier of the language model used for re-scoring (e.g., `google/gemma-2-9b`).
`deep_lm`	bool	False	Enables prefix-span search during LM validation to avoid partial token mismatches. Recommended for multi-word terms.
`lm_threshold`	float	0.01	Minimum improvement in LM score required to accept a phonetic replacement. Higher values enforce more conservative corrections.
`quantization`	str	none	Controls model quantization for memory efficiency. Options include `none`, `4bit`, and `8bit`.

C Detailed Results

C.1 English

Table 8: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **English**. Improvements relative to the baseline are shown in parentheses.

Configuration	Model	WER ↓	Accuracy (%) ↑	Avg. Time / segment (s)
		Baseline		
Baseline	Tiny	18.91	14.23	0.234
	Base	13.66	19.90	0.235
	Small	9.41	33.33	0.269
	Medium	6.88	43.28	0.366
	Large-v3-turbo	6.50	44.88	0.282
	Large-v3	5.24	55.62	0.426
		WhisperX-LLM-TermFixer		
Gemma-3-1B	Tiny	16.84 (-2.07)	35.22 (+20.99)	0.371
	Base	11.08 (-2.58)	48.96 (+29.09)	0.392
	Small	7.24 (-2.17)	61.79 (+28.46)	0.432
	Medium	5.02 (-1.86)	69.45 (+26.17)	0.537
	Large-v3-turbo	4.62 (-1.88)	72.54 (+27.66)	0.435
	Large-v3	3.64 (-1.60)	78.21 (+22.59)	0.604
Gemma-3-4B	Tiny	16.82 (-2.09)	34.93 (+20.70)	0.398
	Base	11.12 (-2.54)	48.46 (+28.56)	0.424
	Small	7.26 (-2.15)	61.59 (+28.26)	0.460
	Medium	5.06 (-1.82)	68.96 (+25.68)	0.568
	Large-v3-turbo	4.66 (-1.84)	72.14 (+27.26)	0.462
	Large-v3	3.63 (-1.61)	78.41 (+22.79)	0.642
Gemma-2-9B	Tiny	16.85 (-2.06)	35.52 (+21.29)	0.390
	Base	11.08 (-2.58)	49.85 (+29.95)	0.410
	Small	7.22 (-2.19)	62.29 (+28.96)	0.544
	Medium	4.97 (-1.91)	70.55 (+27.27)	0.567
	Large-v3-turbo	4.61 (-1.89)	72.74 (+27.86)	0.643
	Large-v3	3.61 (-1.63)	78.81 (+23.19)	0.818
BioMistral-7B	Tiny	16.85 (-2.06)	35.82 (+21.59)	0.372
	Base	11.09 (-2.57)	49.95 (+30.05)	0.397
	Small	7.21 (-2.20)	62.09 (+28.76)	0.434
	Medium	4.98 (-1.90)	70.45 (+27.17)	0.545
	Large-v3-turbo	4.61 (-1.89)	72.74 (+27.86)	0.442
	Large-v3	3.58 (-1.66)	78.71 (+23.09)	0.608
BioMistral-7B-dare	Tiny	16.80 (-2.11)	36.12 (+21.89)	0.381
	Base	11.02 (-2.64)	49.95 (+30.05)	0.397
	Small	7.22 (-2.19)	62.09 (+28.76)	0.438
	Medium	4.97 (-1.91)	70.55 (+27.27)	0.540
	Large-v3-turbo	4.62 (-1.88)	72.54 (+27.66)	0.444
	Large-v3	3.59 (-1.65)	79.00 (+23.38)	0.606
BioMistral-7B-slerp	Tiny	16.83 (-2.08)	35.82 (+21.59)	0.382
	Base	11.09 (-2.57)	49.25 (+29.35)	0.400
	Small	7.21 (-2.20)	62.09 (+28.76)	0.439
	Medium	5.00 (-1.88)	70.15 (+26.87)	0.538
	Large-v3-turbo	4.64 (-1.86)	72.44 (+27.56)	0.438
	Large-v3	3.61 (-1.63)	78.51 (+22.89)	0.601

Table 9: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **English**. Improvements relative to the baseline are shown in parentheses.

Configuration	**Model**	**WER ↓**	**Accuracy (%) ↑**	**Avg. Time / segment (s)**
		Baseline		
Baseline	Tiny	18.91	14.23	0.234
	Base	13.66	19.90	0.235
	Small	9.41	33.33	0.269
	Medium	6.88	43.28	0.366
	Large-v3-turbo	6.50	44.88	0.282
	Large-v3	5.24	55.62	0.426
		WhisperX-LLM-TermFixer		
LLaMA-3.2-1B	Tiny	16.85 (-2.06)	35.92 (+21.69)	0.354
	Base	11.05 (-2.61)	49.95 (+30.05)	0.370
	Small	7.23 (-2.18)	62.19 (+28.86)	0.412
	Medium	4.99 (-1.89)	70.35 (+27.07)	0.517
	Large-v3-turbo	4.55 (-1.95)	73.73 (+28.85)	0.418
	Large-v3	3.63 (-1.61)	78.61 (+22.99)	0.577
LLaMA-3.2-3B	Tiny	16.81 (-2.10)	36.12 (+21.89)	0.366
	Base	11.15 (-2.51)	49.45 (+29.55)	0.383
	Small	7.24 (-2.17)	61.89 (+28.56)	0.422
	Medium	5.00 (-1.88)	70.15 (+26.87)	0.535
	Large-v3-turbo	4.59 (-1.91)	73.13 (+28.25)	0.424
	Large-v3	3.61 (-1.63)	79.00 (+23.38)	0.592
LLaMA-3.1-8B	Tiny	16.89 (-2.02)	35.52 (+21.29)	0.386
	Base	11.11 (-2.55)	49.65 (+29.75)	0.403
	Small	7.28 (-2.13)	61.39 (+28.06)	0.443
	Medium	4.99 (-1.89)	70.45 (+27.17)	0.536
	Large-v3-turbo	4.62 (-1.88)	73.13 (+28.25)	0.448
	Large-v3	3.64 (-1.60)	78.91 (+23.29)	0.601
Salamandra-2B	Tiny	16.83 (-2.08)	35.62 (+21.39)	0.427
	Base	11.12 (-2.54)	48.96 (+29.06)	0.441
	Small	7.27 (-2.14)	61.29 (+27.96)	0.488
	Medium	5.03 (-1.85)	69.45 (+26.17)	0.654
	Large-v3-turbo	4.63 (-1.87)	72.24 (+27.36)	0.513
	Large-v3	3.65 (-1.59)	78.21 (+22.59)	0.734
Salamandra-7B	Tiny	16.79 (-2.12)	36.22 (+21.99)	0.439
	Base	11.04 (-2.62)	49.95 (+30.05)	0.457
	Small	7.17 (-2.24)	62.59 (+29.26)	0.502
	Medium	4.93 (-1.95)	70.95 (+27.67)	0.673
	Large-v3-turbo	4.54 (-1.96)	73.83 (+28.95)	0.526
	Large-v3	3.55 (-1.69)	79.20 (+23.58)	0.750

C.2 Spanish

Table 10: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **Spanish**. Improvements relative to the baseline are shown in parentheses.

Configuration	Model	WER ↓	Accuracy (%) ↑	Avg. Time / segment (s)
		Baseline		
Baseline	Tiny	17.61	31.94	0.246
	Base	11.53	42.59	0.252
	Small	6.54	56.82	0.298
	Medium	3.74	70.85	0.400
	Large-v3-turbo	2.85	76.42	0.283
	Large-v3	2.69	77.71	0.478
		WhisperX-LLM-TermFixer		
Gemma-3-1B	Tiny	14.60 (-3.01)	66.67 (+34.73)	0.433
	Base	8.86 (-2.67)	76.32 (+33.73)	0.449
	Small	4.52 (-2.02)	84.98 (+28.16)	0.494
	Medium	2.51 (-1.23)	90.35 (+19.50)	0.614
	Large-v3-turbo	1.91 (-0.94)	92.84 (+16.42)	0.465
	Large-v3	1.84 (-0.85)	93.33 (+15.62)	0.687
Gemma-3-4B	Tiny	14.57 (-3.04)	67.06 (+35.12)	0.464
	Base	8.81 (-2.72)	77.61 (+35.02)	0.477
	Small	4.46 (-2.08)	85.97 (+29.15)	0.517
	Medium	2.50 (-1.24)	90.95 (+20.10)	0.623
	Large-v3-turbo	1.87 (-0.98)	93.13 (+16.71)	0.484
	Large-v3	1.81 (-0.88)	93.73 (+16.02)	0.689
Gemma-2-9B	Tiny	14.60 (-3.01)	66.77 (+34.83)	0.470
	Base	8.77 (-2.76)	77.81 (+35.22)	0.479
	Small	4.48 (-2.06)	85.97 (+29.15)	0.516
	Medium	2.52 (-1.22)	90.95 (+20.10)	0.624
	Large-v3-turbo	1.85 (-1.00)	93.63 (+17.21)	0.489
	Large-v3	1.81 (-0.88)	93.73 (+16.02)	0.708
BioMistral-7B	Tiny	14.57 (-3.04)	66.77 (+34.83)	0.448
	Base	8.83 (-2.70)	76.52 (+33.93)	0.458
	Small	4.55 (-1.99)	84.78 (+27.96)	0.501
	Medium	2.50 (-1.24)	90.95 (+20.10)	0.610
	Large-v3-turbo	1.86 (-0.99)	93.23 (+16.81)	0.478
	Large-v3	1.79 (-0.90)	93.73 (+16.02)	0.677
BioMistral-7B-dare	Tiny	14.60 (-3.01)	66.47 (+34.53)	0.450
	Base	8.72 (-2.81)	77.81 (+35.22)	0.461
	Small	4.52 (-2.02)	85.17 (+28.35)	0.502
	Medium	2.51 (-1.23)	90.75 (+19.90)	0.611
	Large-v3-turbo	1.86 (-0.99)	93.03 (+16.61)	0.475
	Large-v3	1.82 (-0.87)	93.43 (+15.72)	0.691
BioMistral-7B-slerp	Tiny	14.54 (-3.07)	67.06 (+35.12)	0.442
	Base	8.72 (-2.81)	77.91 (+35.32)	0.448
	Small	4.52 (-2.02)	85.47 (+28.65)	0.483
	Medium	2.54 (-1.20)	90.55 (+19.70)	0.612
	Large-v3-turbo	1.86 (-0.99)	92.94 (+16.52)	0.475
	Large-v3	1.82 (-0.87)	93.33 (+15.62)	0.680

Table 11: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **Spanish**. Improvements relative to the baseline are shown in parentheses.

Configuration	**Model**	**WER ↓**	**Accuracy (%) ↑**	**Avg. Time / segment (s)**
		Baseline		
Baseline	Tiny	17.61	31.94	0.246
	Base	11.53	42.59	0.252
	Small	6.54	56.82	0.298
	Medium	3.74	70.85	0.400
	Large-v3-turbo	2.85	76.42	0.283
	Large-v3	2.69	77.71	0.478
		WhisperX-LLM-TermFixer		
LLaMA-3.2-1B	Tiny	14.67 (-2.94)	65.97 (+34.03)	0.421
	Base	8.91 (-2.62)	76.02 (+33.43)	0.434
	Small	4.65 (-1.89)	84.38 (+27.56)	0.481
	Medium	2.56 (-1.18)	90.75 (+19.90)	0.592
	Large-v3-turbo	1.92 (-0.93)	92.64 (+16.22)	0.454
	Large-v3	1.84 (-0.85)	93.53 (+15.82)	0.662
LLaMA-3.2-3B	Tiny	14.64 (-2.97)	66.47 (+34.53)	0.437
	Base	8.79 (-2.74)	77.41 (+34.82)	0.450
	Small	4.55 (-1.99)	84.98 (+28.16)	0.496
	Medium	2.50 (-1.24)	90.85 (+20.00)	0.600
	Large-v3-turbo	1.90 (-0.95)	92.74 (+16.32)	0.470
	Large-v3	1.83 (-0.86)	93.43 (+15.72)	0.679
LLaMA-3.1-8B	Tiny	14.58 (-3.03)	66.87 (+34.93)	0.452
	Base	8.75 (-2.78)	77.61 (+35.02)	0.463
	Small	4.46 (-2.08)	85.87 (+29.05)	0.503
	Medium	2.50 (-1.24)	90.85 (+20.00)	0.618
	Large-v3-turbo	1.87 (-0.98)	93.33 (+16.91)	0.482
	Large-v3	1.81 (-0.88)	93.83 (+16.12)	0.684
Salamandra-2B	Tiny	14.63 (-2.98)	67.16 (+35.22)	0.585
	Base	8.80 (-2.73)	77.61 (+35.02)	0.595
	Small	4.50 (-2.04)	85.57 (+28.75)	0.663
	Medium	2.50 (-1.24)	90.85 (+20.00)	0.774
	Large-v3-turbo	1.85 (-1.00)	93.13 (+16.71)	0.572
	Large-v3	1.82 (-0.87)	93.63 (+15.92)	0.867
Salamandra-7B	Tiny	14.52 (-3.09)	66.77 (+34.83)	0.611
	Base	8.73 (-2.80)	77.81 (+35.22)	0.613
	Small	4.51 (-2.03)	85.67 (+28.85)	0.676
	Medium	2.51 (-1.23)	90.95 (+20.10)	0.786
	Large-v3-turbo	1.87 (-0.98)	93.23 (+16.81)	0.580
	Large-v3	1.81 (-0.88)	93.83 (+16.12)	0.880

C.3 Catalan

Table 12: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **Catalan**. Improvements relative to the baseline are shown in parentheses.

Configuration	Model	WER ↓	Accuracy (%) ↑	Avg. Time / segment (s)
		Baseline		
Baseline	Tiny	43.91	6.57	0.288
	Base	30.54	11.64	0.307
	Small	18.02	20.20	0.377
	Medium	11.91	26.07	0.535
	Large-v3-turbo	9.03	32.84	0.382
	Large-v3	8.53	37.01	0.691
		WhisperX-LLM-TermFixer		
Gemma-3-1B	Tiny	41.88 (-2.03)	30.25 (+23.68)	0.461
	Base	28.02 (-2.52)	38.61 (+26.97)	0.473
	Small	15.72 (-2.30)	49.05 (+28.85)	0.527
	Medium	9.67 (-2.24)	58.51 (+32.44)	0.655
	Large-v3-turbo	6.99 (-2.04)	63.48 (+30.64)	0.499
	Large-v3	6.39 (-2.14)	69.35 (+32.34)	0.733
Gemma-3-4B	Tiny	41.80 (-2.11)	30.75 (+24.18)	0.485
	Base	27.91 (-2.63)	39.30 (+27.66)	0.497
	Small	15.61 (-2.41)	49.85 (+29.65)	0.552
	Medium	9.54 (-2.37)	59.80 (+33.73)	0.677
	Large-v3-turbo	6.91 (-2.12)	63.78 (+30.94)	0.518
	Large-v3	6.33 (-2.20)	69.25 (+32.24)	0.754
Gemma-2-9B	Tiny	41.78 (-2.13)	30.75 (+24.18)	0.489
	Base	28.01 (-2.53)	38.21 (+26.57)	0.496
	Small	15.66 (-2.36)	50.05 (+29.85)	0.662
	Medium	9.55 (-2.36)	60.00 (+33.93)	0.678
	Large-v3-turbo	6.85 (-2.18)	64.58 (+31.74)	0.523
	Large-v3	6.28 (-2.25)	70.05 (+33.04)	0.752
BioMistral-7B	Tiny	41.84 (-2.07)	30.35 (+23.78)	0.475
	Base	28.01 (-2.53)	38.31 (+26.67)	0.488
	Small	15.76 (-2.26)	48.96 (+28.76)	0.544
	Medium	9.63 (-2.28)	58.81 (+32.74)	0.670
	Large-v3-turbo	6.95 (-2.08)	63.78 (+30.94)	0.517
	Large-v3	6.34 (-2.19)	69.55 (+32.54)	0.743
BioMistral-7B-dare	Tiny	41.80 (-2.11)	30.45 (+23.88)	0.478
	Base	27.93 (-2.61)	38.51 (+26.87)	0.488
	Small	15.62 (-2.40)	49.85 (+29.65)	0.539
	Medium	9.59 (-2.32)	59.30 (+33.23)	0.665
	Large-v3-turbo	6.95 (-2.08)	63.48 (+30.64)	0.512
	Large-v3	6.30 (-2.23)	70.05 (+33.04)	0.742
BioMistral-7B-slerp	Tiny	41.79 (-2.12)	30.75 (+24.18)	0.472
	Base	28.01 (-2.53)	38.91 (+27.27)	0.488
	Small	15.68 (-2.34)	49.55 (+29.35)	0.541
	Medium	9.59 (-2.32)	59.20 (+33.13)	0.666
	Large-v3-turbo	6.97 (-2.06)	63.18 (+30.34)	0.512
	Large-v3	6.31 (-2.22)	69.75 (+32.74)	0.736

Table 13: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **Catalan**. Improvements relative to the baseline are shown in parentheses.

Configuration	**Model**	**WER ↓**	**Accuracy (%) ↑**	**Avg. Time / segment (s)**
		Baseline		
Baseline	Tiny	43.91	6.57	0.288
	Base	30.54	11.64	0.307
	Small	18.02	20.20	0.377
	Medium	11.91	26.07	0.535
	Large-v3-turbo	9.03	32.84	0.382
	Large-v3	8.53	37.01	0.691
		WhisperX-LLM-TermFixer		
LLaMA-3.2-1B	Tiny	41.88 (-2.03)	30.25 (+23.68)	0.452
	Base	28.03 (-2.51)	38.01 (+26.37)	0.465
	Small	15.70 (-2.32)	49.65 (+29.45)	0.517
	Medium	9.61 (-2.30)	59.30 (+33.23)	0.625
	Large-v3-turbo	7.06 (-1.97)	63.38 (+30.54)	0.493
	Large-v3	6.40 (-2.13)	69.55 (+32.54)	0.731
LLaMA-3.2-3B	Tiny	41.79 (-2.12)	30.65 (+24.08)	0.467
	Base	27.99 (-2.55)	39.00 (+27.36)	0.476
	Small	15.69 (-2.33)	49.55 (+29.35)	0.529
	Medium	9.54 (-2.37)	60.00 (+33.93)	0.659
	Large-v3-turbo	6.92 (-2.11)	64.68 (+31.84)	0.498
	Large-v3	6.35 (-2.18)	69.95 (+32.94)	0.731
LLaMA-3.1-8B	Tiny	41.75 (-2.16)	31.24 (+24.67)	0.476
	Base	27.94 (-2.60)	39.20 (+27.56)	0.493
	Small	15.63 (-2.39)	50.45 (+30.25)	0.545
	Medium	9.54 (-2.37)	60.20 (+34.13)	0.673
	Large-v3-turbo	6.83 (-2.20)	65.07 (+32.23)	0.520
	Large-v3	6.30 (-2.23)	70.15 (+33.14)	0.745
Salamandra-2B	Tiny	41.71 (-2.20)	30.55 (+23.98)	0.456
	Base	27.90 (-2.64)	39.40 (+27.76)	0.469
	Small	15.56 (-2.46)	50.45 (+30.25)	0.523
	Medium	9.54 (-2.37)	59.90 (+33.83)	0.647
	Large-v3-turbo	6.93 (-2.10)	63.68 (+30.84)	0.492
	Large-v3	6.29 (-2.24)	70.15 (+33.14)	0.726
Salamandra-7B	Tiny	41.71 (-2.20)	31.34 (+24.77)	0.472
	Base	27.78 (-2.76)	39.80 (+28.16)	0.487
	Small	15.51 (-2.51)	50.65 (+30.45)	0.545
	Medium	9.50 (-2.41)	60.10 (+34.03)	0.658
	Large-v3-turbo	6.80 (-2.23)	65.17 (+32.33)	0.552
	Large-v3	6.22 (-2.31)	70.35 (+33.34)	0.894

C.4 Basque

Table 14: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **Basque**. Improvements relative to the baseline are shown in parentheses.

Configuration	Model	WER ↓	Accuracy (%) ↑	Avg. Time / segment (s)
		Baseline		
Baseline	Tiny	97.32	6.67	0.264
	Base	90.99	12.54	0.274
	Small	71.24	15.42	0.324
	Medium	57.06	17.51	0.447
	Large-v3-turbo	35.28	31.24	0.315
	Large-v3	36.61	28.26	0.542
		WhisperX-LLM-TermFixer		
Gemma-3-1B	Tiny	94.36 (-2.96)	29.85 (+23.18)	0.462
	Base	87.91 (-3.08)	35.82 (+23.28)	0.481
	Small	68.02 (-3.22)	40.80 (+25.38)	0.525
	Medium	53.74 (-3.32)	43.48 (+25.97)	0.664
	Large-v3-turbo	32.93 (-2.35)	56.82 (+25.58)	0.493
	Large-v3	34.53 (-2.08)	52.64 (+24.38)	0.739
Gemma-3-4B	Tiny	94.30 (-3.02)	30.65 (+23.98)	0.488
	Base	87.81 (-3.18)	36.52 (+23.98)	0.507
	Small	67.86 (-3.38)	41.39 (+25.97)	0.553
	Medium	53.50 (-3.56)	44.38 (+26.87)	0.693
	Large-v3-turbo	32.74 (-2.54)	57.61 (+26.37)	0.516
	Large-v3	34.30 (-2.31)	52.94 (+24.68)	0.768
Gemma-2-9B	Tiny	94.18 (-3.14)	31.64 (+24.97)	0.492
	Base	87.72 (-3.27)	36.62 (+24.08)	0.513
	Small	67.92 (-3.32)	42.69 (+27.27)	0.556
	Medium	53.44 (-3.62)	45.57 (+28.06)	0.697
	Large-v3-turbo	32.59 (-2.69)	59.40 (+28.16)	0.520
	Large-v3	34.21 (-2.40)	54.53 (+26.27)	0.771
BioMistral-7B	Tiny	94.56 (-2.76)	32.14 (+25.47)	0.477
	Base	87.95 (-3.04)	38.41 (+25.87)	0.492
	Small	68.10 (-3.14)	43.58 (+28.16)	0.540
	Medium	53.90 (-3.16)	44.68 (+27.17)	0.687
	Large-v3-turbo	33.06 (-2.22)	58.11 (+26.87)	0.507
	Large-v3	34.67 (-1.94)	54.13 (+25.87)	0.756
BioMistral-7B-dare	Tiny	94.38 (-2.94)	33.03 (+26.36)	0.475
	Base	87.80 (-3.19)	39.50 (+26.96)	0.492
	Small	68.04 (-3.20)	43.88 (+28.46)	0.541
	Medium	53.82 (-3.24)	45.97 (+28.46)	0.684
	Large-v3-turbo	33.01 (-2.27)	58.81 (+27.57)	0.506
	Large-v3	34.47 (-2.14)	55.82 (+27.56)	0.757
BioMistral-7B-slerp	Tiny	94.40 (-2.92)	32.54 (+25.87)	0.477
	Base	87.92 (-3.07)	38.41 (+25.87)	0.494
	Small	68.10 (-3.14)	43.28 (+27.86)	0.539
	Medium	53.87 (-3.19)	44.88 (+27.37)	0.686
	Large-v3-turbo	33.05 (-2.23)	58.41 (+27.17)	0.508
	Large-v3	34.68 (-1.93)	54.63 (+26.37)	0.756

Table 15: WER and term-level accuracy for Whisper models of different sizes using WhisperX under a baseline configuration and the proposed approach **WhisperX-LLM-TermFixer** in **Basque**. Improvements relative to the baseline are shown in parentheses.

Configuration	**Model**	**WER ↓**	**Accuracy (%) ↑**	**Avg. Time / segment (s)**
		Baseline		
Baseline	Tiny	97.32	6.67	0.264
	Base	90.99	12.54	0.274
	Small	71.24	15.42	0.324
	Medium	57.06	17.51	0.447
	Large-v3-turbo	35.28	31.24	0.315
	Large-v3	36.61	28.26	0.542
		WhisperX-LLM-TermFixer		
LLaMA-3.2-1B	Tiny	94.15 (-3.17)	33.33 (+26.66)	0.447
	Base	87.63 (-3.36)	40.00 (+27.46)	0.465
	Small	67.59 (-3.65)	45.87 (+30.45)	0.513
	Medium	53.53 (-3.53)	46.27 (+28.76)	0.654
	Large-v3-turbo	32.80 (-2.48)	57.91 (+26.67)	0.483
	Large-v3	34.41 (-2.20)	54.83 (+26.57)	0.729
LLaMA-3.2-3B	Tiny	94.11 (-3.21)	32.14 (+25.47)	0.464
	Base	87.63 (-3.36)	39.30 (+26.76)	0.480
	Small	67.50 (-3.74)	45.37 (+29.95)	0.528
	Medium	53.21 (-3.85)	46.87 (+29.36)	0.673
	Large-v3-turbo	32.67 (-2.61)	58.81 (+27.57)	0.497
	Large-v3	34.30 (-2.31)	54.23 (+25.97)	0.746
LLaMA-3.1-8B	Tiny	94.09 (-3.23)	32.44 (+25.77)	0.481
	Base	87.54 (-3.45)	39.30 (+26.76)	0.500
	Small	67.53 (-3.71)	45.27 (+29.85)	0.544
	Medium	53.20 (-3.86)	46.77 (+29.26)	0.689
	Large-v3-turbo	32.62 (-2.66)	58.81 (+27.57)	0.510
	Large-v3	34.12 (-2.49)	55.52 (+27.26)	0.758
Salamandra-2B	Tiny	93.89 (-3.43)	33.83 (+27.16)	0.455
	Base	87.42 (-3.57)	40.60 (+28.06)	0.472
	Small	67.64 (-3.60)	44.98 (+29.56)	0.521
	Medium	53.19 (-3.87)	46.87 (+29.36)	0.665
	Large-v3-turbo	32.65 (-2.63)	59.30 (+28.06)	0.491
	Large-v3	34.16 (-2.45)	55.32 (+27.06)	0.738
Salamandra-7B	Tiny	93.76 (-3.56)	34.93 (+28.26)	0.478
	Base	87.32 (-3.67)	41.89 (+29.35)	0.496
	Small	67.47 (-3.77)	46.27 (+30.85)	0.540
	Medium	53.05 (-4.01)	47.76 (+30.25)	0.686
	Large-v3-turbo	32.44 (-2.84)	60.60 (+29.36)	0.509
	Large-v3	33.95 (-2.66)	56.02 (+27.76)	0.757