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

Ronghao Pan, José Antonio García-Díaz, Rafael Valencia-García*
Departamento de Informática y Sistemas, Universidad de Murcia,
Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan, joseantonio.garcia8, valencia}@um.es

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 the method on a controlled synthetic benchmark containing 1,005 utterances per language in English, Spanish, Catalan, and Basque, with one target medical term per utterance and full lexicon coverage. Across six Whisper sizes, the selected configurations improve term accuracy by 16.12–35.32 percentage points and reduce WER by 0.88–4.01 points relative to WhisperX. LLM filtering reduces WER relative to the phonetic stage in all evaluated configurations, although it reduces term accuracy in three Basque configurations. Perplexity results provide additional evidence of improved linguistic plausibility. The selected configurations require approximately 1.26x–2.38x the baseline inference time. Our code and data are publicly available at https://github.com/NLP-UMUTeam/WhisperX-LLM-TermFixer.

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 and weakly supervised training on large volumes of data enable robust recognition in general-domain scenarios without the need for explicit adaptation.

Despite its strong overall performance, Whisper exhibits limitations in specialized applications. In particular, it can struggle to recognize rare words, including domain-specific terminology (Jogi et al., 2024). Clinical applications also require attention to speech details that conventional transcription may fail to preserve. Hwang et al. (2025) identify omissions of clinically relevant filler words in psychiatric interviews, while Lin et al. (2024) highlight that Whisper's context-dependent output correction can obscure pronunciation details needed for child speech assessment. These findings motivate targeted approaches to terminology recognition and task-specific evaluation in sensitive clinical settings.

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

* Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3019–3042
November 6-10, 2026 ©2026 Association for Computational Linguistics

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 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 adaptation strategies discussed above, including fine-tuning, adapter-based methods, and external language model fusion or re-scoring, differ in their requirements for training data and access to the decoding process. Fine-tuning and adapter-based methods require additional optimization, whereas external language model fusion or re-scoring can operate without retraining the acoustic model. Our approach assumes access to a domain-specific lexicon and pretrained language models, without additional in-domain training.

We propose **WhisperX-LLM-TermFixer**, a modular framework 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 using a diverse set of LLMs, including Gemma-3 (1B and 4B) (Gemma Team et al., 2025), Gemma-2 (9B) (Gemma Team 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. **WhisperX-LLM-TermFixer**, a modular framework for improving domain-specific term recognition in Whisper-based ASR systems without requiring model retraining or architectural modifications.

2. A hybrid correction strategy that combines N-best hypothesis re-ranking, phonetic and textual similarity matching, and LLM-based contextual re-scoring to improve term recovery while accounting for sentence-level plausibility.

3. 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 adaptation through fine-tuning. First, WhisperX segments the audio using Voice Activity Detection (VAD). Each segment is decoded with beam search to obtain an N-best list and decoding scores. Lexicon-guided selection prioritizes lexical matches, with the Whisper score used to break ties. Next, a phonetic and textual n-gram module proposes term replacements. Finally, an LLM scores the resulting sentence variants and selects the lowest-loss valid candidate among those evaluated, provided that the improvement exceeds a threshold. Constraints restrict partial replacements and duplicate insertions. This sequential search does not guarantee a global optimum over all possible combinations of edits.

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

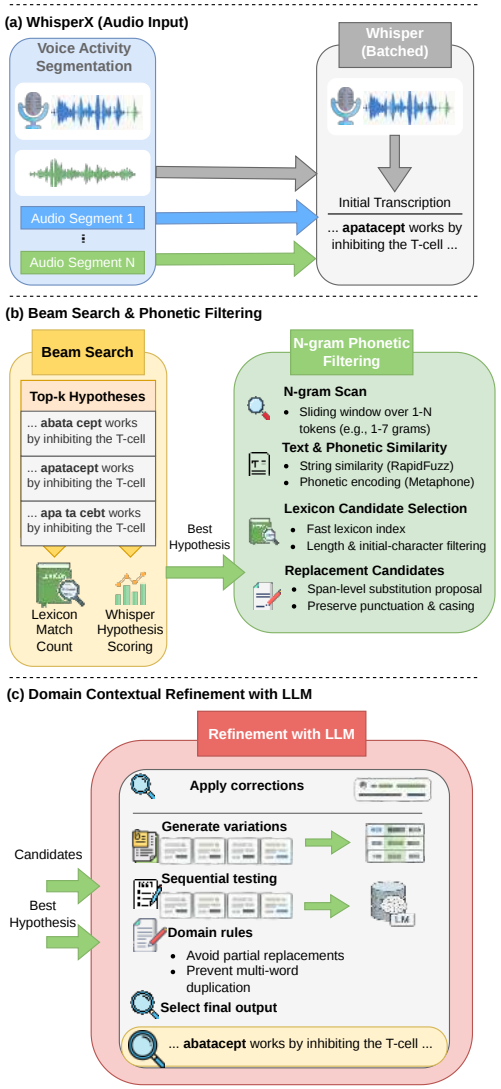


Figure 1: System architecture. WhisperX segments the input audio; beam decoding returns multiple hypotheses, followed by lexicon-guided selection and phonetic candidate generation. An LLM scores replacement variants to select a valid lower-loss output. Beam search operates during acoustic decoding.

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

We extend the WhisperX decoding interface to return multiple hypotheses per segment, together with their decoding scores, and use this N-best list for lexicon-guided selection. The extension concerns hypothesis access and selection in the pipeline, rather than introducing beam search as a new capability of Faster-Whisper. Alternative hypotheses can contain useful term variants that are absent from the highest-scoring transcription.

2.3.1 Lexicon-guided hypothesis selection

We select a hypothesis from the N-best list by prioritizing matches with the domain-specific lexicon. When multiple hypotheses contain the same number of matches, the Whisper decoding score is used to break ties.

Before lexical matching, both lexicon entries and decoded words are normalized using the same procedure. Strings are converted to lowercase, decomposed using NFD normalization, and stripped of combining marks. Consequently, matching is case-insensitive and accent-insensitive; for example, *radiografía* and *radiografia* are treated as equivalent. Words are extracted using a Unicode-aware regular expression. The normalized forms are used only for lexical comparison, while the canonical lexicon forms are retained for subsequent correction. The complete normalization, indexing, and output-reconstruction rules are summarized in Appendix B.

Formally, let $\{h_k\}_{k=1}^{K}$ denote the 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:

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

where $\mathrm{words}(\cdot)$ denotes word-level tokenization and $\mathbf{I}$ 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 the 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 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 reduces 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 an 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 is intended to reject replacements that lower sentence likelihood. It does not guarantee semantic correctness, and its effect on transcription quality is assessed using 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., replacing a prefix with "agni casti fructus" while retaining the following token "fructus", 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 the loss reduction is at least τ.

3 Experimental Setup

3.1 Dataset

To evaluate medical terminology recognition, we construct a synthetic speech dataset comprising 1,005 domain-specific terms per language, covering English, Spanish, 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 control over lexical coverage, sentence content, and speaker generation settings, enabling a focused analysis of term-detection errors independent 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. For Spanish, Catalan, and Basque, speech was generated using the Microsoft Text-to-Speech API, which provides neural voices for the three languages. English speech was generated using `Coqui-TTS`[2] with multiple speakers.

Table 1 summarizes the synthetic benchmark. Each language contains 1,005 utterances constructed around one target term from the same terminology inventory. The correction lexicon contains all target terms, making this a closed-set, full-coverage evaluation. This design isolates recovery of known terminology rather than open-domain term discovery. Sentence length and duration differ across languages; these differences reflect the generated material and synthesis settings as well as linguistic variation. Because English uses a different TTS system, cross-language differences cannot be attributed to language alone.

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

Language	Total words	Words / Utterance	Mean duration (s)
English	16,190	16.11	8.10
Spanish	19,579	19.48	7.75
Catalan	19,323	19.23	9.47
Basque	13,123	13.06	9.36

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

3.2 Parameter Settings

All experiments are conducted using WhisperX with the Faster-Whisper backend. Audio is segmented using VAD (`vad_onset=0.15`, `vad_offset=0.20`) and decoded using the corresponding language setting for each evaluation language, 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 Whisper's 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 an LLM; a replacement is accepted only if it improves sentence likelihood by at least `lm_threshold=0.01`, favoring conservative corrections with lower LLM loss. 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.

Appendix C reports utterance-level bootstrap 95% confidence intervals for each corrected system's WER and term accuracy, based on 5,000 resamples. These marginal intervals quantify uncertainty in the individual metrics; they are not confidence intervals for paired differences. We therefore do not infer the statistical significance of system differences from these intervals alone.

4 Experimental Results

4.1 Main Results

Table 2 reports, for each language and Whisper size, the selected LLM variant and its performance within each model family. Selection is based on the highest observed term accuracy, with lower WER used as a tie-breaker. If both metrics are identical at the reported precision, the smaller LLM is selected; if model size is also tied, the base non-merged variant is retained. Table 3 applies the same selection rule across families.

These summaries describe the best observed configurations on this benchmark and should not be interpreted as a model-selection procedure validated on independent data. The grid comprises 11 LLMs, six Whisper sizes, and four languages (264 configurations). All evaluated families improve the reported metrics relative to WhisperX. Small differences between LLMs do not establish a statistically reliable family ranking. Detailed results and marginal confidence intervals appear in Appendix C.

We additionally compare against a lexicon-based hotword baseline. A first Whisper pass produces a transcript; our phonetic candidate-generation module retrieves up to five lexicon terms per utterance, which are injected as hotwords in a second decoding pass. Thus, this implementation requires two Whisper decoding passes. The selected full-system configurations achieve lower WER and higher term accuracy than this baseline in English, Spanish, and Catalan. In Basque, hotwords achieve higher accuracy for Small, Medium, and Large-v3, while the full system retains lower WER. The full pipeline is faster in the reported implementation, but this does not imply an intrinsic speed advantage over other hotword implementations.

Table 17 in Appendix D reports a staged ablation. Lexicon-guided N-best selection improves term recovery over WhisperX in every configuration. Adding phonetic correction increases term accuracy further, although WER can increase. The

Whisper	Gemma			BioMistral			LLaMA			Salamandra		
	Model	WER	Acc	Model	WER	Acc	Model	WER	Acc	Model	WER	Acc
English												
Tiny	2-9B	16.85	35.52	DARE	16.80	36.12	3.2-3B	16.81	36.12	7B	**16.79**	**36.22**
Base	2-9B	11.08	49.85	DARE	**11.02**	**49.95**	3.2-1B	11.05	49.95	7B	11.04	49.95
Small	2-9B	7.22	62.29	7B	7.21	62.09	3.2-1B	7.23	62.19	7B	**7.17**	**62.59**
Medium	2-9B	4.97	70.55	DARE	4.97	70.55	3.1-8B	4.99	70.45	7B	**4.93**	**70.95**
Large-t	2-9B	4.61	72.74	7B	4.61	72.74	3.2-1B	4.55	73.73	7B	**4.54**	**73.83**
Large	2-9B	3.61	78.81	DARE	3.59	79.00	3.2-3B	3.61	79.00	7B	**3.55**	**79.20**
Spanish												
Tiny	3-4B	14.57	67.06	SLERP	14.54	67.06	3.1-8B	14.58	66.87	2B	**14.63**	**67.16**
Base	2-9B	8.77	77.81	SLERP	**8.72**	**77.91**	3.1-8B	8.75	77.61	7B	8.73	77.81
Small	3-4B	**4.46**	**85.97**	SLERP	4.52	85.47	3.1-8B	4.46	85.87	7B	4.51	85.67
Medium	3-4B	**2.50**	**90.95**	7B	2.50	90.95	3.2-3B	2.50	90.85	7B	2.51	90.95
Large-t	2-9B	**1.85**	**93.63**	7B	1.86	93.23	3.1-8B	1.87	93.33	7B	1.87	93.23
Large	3-4B	1.81	93.73	7B	1.79	93.73	3.1-8B	**1.81**	**93.83**	7B	**1.81**	**93.83**
Catalan												
Tiny	2-9B	41.78	30.75	SLERP	41.79	30.75	3.1-8B	41.75	31.24	7B	**41.71**	**31.34**
Base	3-4B	27.91	39.30	SLERP	28.01	38.91	3.1-8B	27.94	39.20	7B	**27.78**	**39.80**
Small	2-9B	15.66	50.05	DARE	15.62	49.85	3.1-8B	15.63	50.45	7B	**15.51**	**50.65**
Medium	2-9B	9.55	60.00	DARE	9.59	59.30	3.1-8B	**9.54**	**60.20**	7B	9.50	60.10
Large-t	2-9B	6.85	64.58	7B	6.95	63.78	3.1-8B	6.83	65.07	7B	**6.80**	**65.17**
Large	2-9B	6.28	70.05	DARE	6.30	70.05	3.1-8B	6.30	70.15	7B	**6.22**	**70.35**
Basque												
Tiny	2-9B	94.18	31.64	DARE	94.38	33.03	3.2-1B	94.15	33.33	7B	**93.76**	**34.93**
Base	2-9B	87.72	36.62	DARE	87.80	39.50	3.2-1B	87.63	40.00	7B	**87.32**	**41.89**
Small	2-9B	67.92	42.69	DARE	68.04	43.88	3.2-1B	67.59	45.87	7B	**67.47**	**46.27**
Medium	2-9B	53.44	45.57	DARE	53.82	45.97	3.2-3B	53.21	46.87	7B	**53.05**	**47.76**
Large-t	2-9B	32.59	59.40	DARE	33.01	58.81	3.1-8B	32.62	58.81	7B	**32.44**	**60.60**
Large	2-9B	34.21	54.53	DARE	34.47	55.82	3.1-8B	34.12	55.52	7B	**33.95**	**56.02**

Table 2: Best-performing LLM within each model family for each language and Whisper model size. The *Model* column identifies the selected LLM variant. Selection is based on the highest observed term accuracy, with lower WER used as a tie-breaker. If both metrics are identical at the reported precision, the smaller LLM is selected; if model size is also tied, the base non-merged variant is retained. Boldface marks the WER and accuracy of the best-performing family in each row under the same criterion. Thus, bold WER values are not necessarily the lowest in the row. WER and term accuracy are reported as percentages. *Large-t* denotes *large-v3-turbo*, and *Large* denotes *large-v3*.

largest accuracy increment comes from the phonetic stage in 14 of 24 configurations and from N-best selection in the remaining ten. The full system has the lowest WER among the stages in all 24 configurations and the highest term accuracy in English, Spanish, and Catalan. In Basque, its term accuracy is lower than that of the phonetic stage for Small, Medium, and Large-v3. These results show a language-dependent trade-off between term recovery and overall transcription quality.

4.2 Multilingual Analysis

Table 3 compares baseline WhisperX with the full pipeline across four languages and six Whisper model sizes. For each configuration, the LLM is selected according to the highest observed term accuracy, with WER used as a tie-breaker. All selected configurations improve both metrics relative to the baseline. Term-accuracy gains range from 16.12 to 35.32 percentage points, while WER decreases by 0.88–4.01 points.

English improves by 21.99–30.05 points in term accuracy, accompanied by WER reductions of 1.69–2.64 points. Whisper Base exhibits the largest accuracy gain, increasing from 19.90% to 49.95%, while its WER decreases from 13.66% to 11.02%. Improvements remain substantial for larger models: Whisper Large increases term accuracy from 55.62% to 79.20% and reduces WER from 5.24% to 3.55%. This combination shows that a comparatively low overall WER can coexist with considerable difficulty recognizing the target terminology.

Spanish achieves the highest final term accuracy among the evaluated languages, reaching 93.83%

with Whisper Large. The largest absolute gains occur for Tiny and Base, whose term accuracies increase by 35.22 and 35.32 points, respectively. For Tiny, accuracy rises from 31.94% to 67.16%, while WER decreases from 17.61% to 14.63%. As baseline accuracy increases with model size, the remaining improvement becomes smaller: Large-v3-turbo and Large gain 17.21 and 16.12 points. Nevertheless, their WER also decreases, from 2.85% to 1.85% and from 2.69% to 1.81%, respectively. The smaller absolute gains for these configurations are consistent with their higher starting accuracy and reduced headroom for improvement.

Catalan shows term-accuracy gains of 24.77–34.13 points and WER reductions of 2.20–2.76 points. Unlike Spanish, substantial accuracy gains persist for the larger Whisper models. Medium improves from 26.07% to 60.20%, corresponding to the largest gain for Catalan, while Large improves from 37.01% to 70.35%. These results indicate that increasing ASR model size alone does not eliminate terminology errors in this benchmark. At the same time, absolute output quality remains strongly dependent on the underlying model: the full pipeline yields a WER of 41.71% for Tiny, compared with 6.22% for Large. Term correction therefore complements a stronger ASR backbone but does not compensate fully for a weak one.

Basque obtains consistent term-accuracy gains of 27.76–30.85 points and the largest absolute WER reductions in the comparison, ranging from 2.66 to 4.01 points. However, these improvements must be interpreted alongside the high residual WER. For Small, term accuracy increases from 15.42% to 46.27%, while WER decreases from 71.24% to 67.47%. Medium achieves the largest WER reduction, from 57.06% to 53.05%, but its final term accuracy remains 47.76%. Large-v3-turbo provides the strongest selected Basque result on both metrics, with 60.60% term accuracy and 32.44% WER, compared with 56.02% and 33.95% for Large-v3. Thus, performance does not improve monotonically with model size. More broadly, the Basque results demonstrate recovery of known terminology under weak baseline ASR conditions; they do not establish clinically reliable transcription or a general solution to low-resource ASR.

The relationship between baseline quality and improvement is not uniform across languages. Spanish exhibits its largest term-accuracy gains for smaller models, whereas Catalan retains gains above 32 points for Medium, Large-v3-turbo, and Large. Basque shows comparatively similar accuracy gains across sizes despite large differences in absolute WER. These patterns support evaluating both term accuracy and WER: recovering the target expression and producing an accurate transcript are related but distinct outcomes. In particular, substantial improvements in the former can coexist with a high error rate elsewhere in the transcription.

Whisper	**Baseline**		**Ours**		**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

Table 3: Comparison between baseline WhisperX and WhisperX-LLM-TermFixer (Ours) for each language and Whisper model size. For Ours, the LLM is selected across all evaluated models using the highest observed term accuracy, with lower WER, smaller model size, and preference for the base non-merged variant used as successive tie-breakers. WER and term accuracy are reported as percentages. Δ denotes Ours minus Baseline in percentage points; negative WER differences and positive accuracy differences indicate improvements. *Large-t* denotes *large-v3-turbo*, and *Large* denotes *large-v3*.

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

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.

Table 4: These examples illustrate how LLM-based filtering can reject overly broad replacements and preserve relevant surrounding tokens. In the middle column, green highlights the span requiring correction, while red marks tokens that should be retained.

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 spans 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.

A staged component ablation is provided in Appendix D. The results quantify the contribution of lexicon-guided N-best selection, phonetic correction, and LLM-based contextual filtering across all languages and Whisper model sizes.

Table 4 illustrates representative examples of this 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 selects outputs with greater sentence-level plausibility.

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.

Table 5 shows lower perplexity for the full system than for the phonetic configuration at every Whisper size in English, Spanish, and Catalan. In Basque, the full system improves on the phonetic configuration for Base and Large-v3, but has slightly higher perplexity for the other four sizes. All corrected configurations have lower perplexity than the baseline. These scores provide evidence about linguistic plausibility under the evaluation models; they do not establish semantic correctness.

Perplexity depends on the evaluation model, tokenization, and aggregation procedure. Comparisons should therefore be made within each language and evaluation setup rather than treating values across languages as directly comparable.

The Basque results indicate that lower WER and lower language-model perplexity need not select the same correction strategy. The current experiment does not isolate the effects of morphology, training coverage, and candidate generation, so we do not assign a causal explanation to these differences.

4.4 Efficiency Analysis

Figure 2 reports full-pipeline latency relative to WhisperX for the configurations selected in Table 3. Ratios calculated from the published average segment times range from approximately 1.26× to 2.38×. Across all 264 LLM configurations in Appendix C, the range is approximately 1.05×–2.48×.

The full-to-baseline ratio includes changes in hypothesis generation, phonetic correction, and contextual filtering; it does not isolate LLM cost. Table 17 reports the intermediate-stage times needed to inspect these contributions. Ratios vary with language, Whisper size, and LLM choice. These measurements characterize the tested 1,005-term lexicon and implementation, without establishing scaling behavior for larger lexicons or other hardware.

Appendix E compares the full pipeline on clean audio and audio mixed with hospital ambient noise at 10 dB SNR, using Salamandra-7B throughout. The largest WER increase is 1.66 points and the largest accuracy decrease is 2.09 points, both for Basque Large-v3-turbo. This experiment describes sensitivity to one noise condition. Because it does not report a matched noisy WhisperX baseline, it

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**

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 greater likelihood under the corresponding evaluation model, rather than guaranteed semantic correctness.

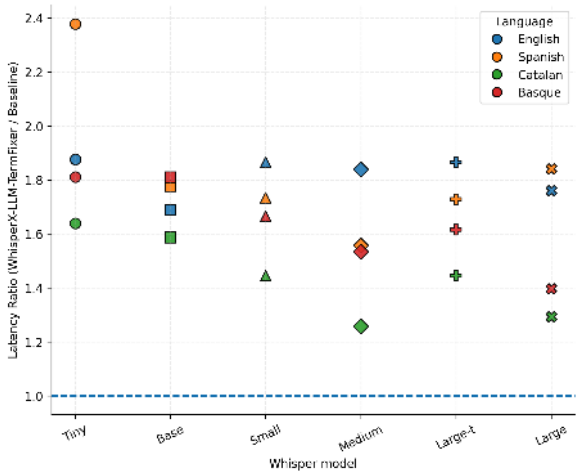


Figure 2: Full-pipeline latency divided by baseline WhisperX latency for each language and Whisper size, using the configurations selected in Table 3. The ratio includes hypothesis generation, phonetic correction, and contextual filtering, not only LLM scoring.

does not establish the size of the method's advantage under noise.

5 Conclusions

We presented *WhisperX-LLM-TermFixer*, a modular framework for improving domain-specific terminology recognition without retraining the acoustic model. The approach combines lexicon-guided N-best hypothesis selection, phonetic and textual n-gram correction, and LLM-based contextual filtering. Together, these components recover candidate terms while using sentence-level likelihood to guide replacement decisions.

Evaluation on a controlled synthetic medical benchmark covering English, Spanish, Catalan, and Basque shows improvements across all six Whisper sizes. The selected configurations increase term accuracy by 16.12–35.32 percentage points and reduce WER by 0.88–4.01 points relative to baseline WhisperX. Gains remain substantial for several larger models, indicating that stronger general transcription performance does not eliminate errors on specialized terminology.

The staged ablation clarifies the contribution of each component. Lexicon-guided selection improves initial term recovery, and phonetic correction further increases term accuracy, although it can introduce additional transcription errors. Adding LLM-based filtering produces the lowest WER among the evaluated stages in every configuration. Its effect on term accuracy is language-dependent: the full pipeline achieves the highest accuracy in English, Spanish, and Catalan, but loses some recovered terms in three Basque configurations. This illustrates the need to evaluate terminology recovery alongside overall transcription quality.

Perplexity results provide complementary evidence of improved linguistic plausibility, although they do not establish semantic or clinical correctness. The selected configurations require approximately 1.26×–2.38× the baseline inference time, reflecting the combined cost of hypothesis generation, phonetic correction, and contextual filtering. The practical value of this additional computation therefore depends on the application's accuracy and latency requirements.

Overall, the findings support post-decoding correction for recovering known terminology under full lexicon coverage. Their scope remains limited by synthetic speech and the single-target-term design; in particular, high residual Basque WER prevents interpreting term recovery as clinically reliable transcription.

Limitations

First, the benchmark uses automatically generated sentences and TTS audio, without human validation of term pronunciation. The different TTS sys-

tems used for English and the other languages confound comparisons across languages. Adding hospital noise at a single 10 dB SNR does not validate pronunciations or substitute for real, spontaneous clinical speech. Human-verified and real-speech evaluation remains necessary.

Second, the lexicon contains all 1,005 target terms, with one target per utterance. This is a favorable closed-set setting for term recovery. The evaluation does not measure false-positive insertions on general or out-of-lexicon speech, partial coverage, held-out terminology, or multi-term utterances. Future work should assess these settings using both term-recovery and false-positive metrics.

Third, we do not vary lexicon size. Length and initial-character indexing restrict candidate retrieval, but larger lexicons may increase ambiguity and processing cost. The reported latency ratios should not be generalized beyond the tested lexicon and implementation. A scaling study should jointly measure latency, WER, term accuracy, and spurious insertions.

Fourth, the best-performing LLM summaries describe observed benchmark results and do not demonstrate selection on independent data. The reported marginal confidence intervals do not establish significance of paired system differences. The hotword comparison covers one two-pass implementation; no comparison against an in-domain n-gram LM is provided.

Finally, perplexity measures model-dependent linguistic plausibility, not clinical or semantic correctness. Basque WER remains high despite improved term recovery, so those results should be interpreted as a stress test under weak baseline ASR conditions. The method targets domain terminology and does not solve low-resource ASR more generally.

Ethical Considerations

This work aims 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

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00), EmpaTIC-ES (PID2025-169617OB-I00) and MedicaLLM (CPP2024-011574) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU) — a way of making Europe.

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 Journal of Selected Topics in Signal Processing*, 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 192 others. 2025. Gemma 3 Technical Report. *ArXiv*, abs/2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 177 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *ArXiv*, abs/2408.00118.

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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 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 Transactions on Audio, Speech, and Language Processing*, 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.

Yash Jogi, Vaibhav Aggarwal, Shabari S Nair, Yash Verma, and Aayush Kubba. 2024. Improving Rare-Word Recognition of Whisper in Zero-Shot Settings. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 216–223.

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 Transactions on Audio, Speech, and Language Processing*, 32:325–351.

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.

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 Transactions on Audio, Speech, and Language Processing*, 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, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2023. Gemini: a family of highly capable multimodal models.

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 illustrates selected medical terms, their cross-language variants, and the generated sentence contexts used in the synthetic benchmark.

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

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. Table 8 details the normalization, candidate-matching, and output-preservation rules.

C Detailed Results

This appendix reports detailed results for all evaluated LLMs, Whisper sizes, and languages, including baseline WhisperX and hotword biasing. Each table reports WER, term accuracy, and average processing time per segment. For the LLM configurations, brackets contain utterance-level bootstrap 95% confidence intervals for WER and term accuracy, based on 5,000 resamples.

C.1 English

Tables 9 and 10 report the English results. The first table presents the shared WhisperX and WhisperX with hotword biasing baselines together with the Gemma and BioMistral variants, while the second continues the comparison with the LLaMA and Salamandra variants.

C.2 Spanish

Tables 11 and 12 report the Spanish results. The first table presents the shared WhisperX and WhisperX with hotword biasing baselines together with the Gemma and BioMistral variants, while the second continues the comparison with the LLaMA and Salamandra variants.

C.3 Catalan

Tables 13 and 14 report the Catalan results. The first table presents the shared WhisperX and WhisperX with hotword biasing baselines together with the Gemma and BioMistral variants, while the second continues the comparison with the LLaMA and Salamandra variants.

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 antitrypsin 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.

Table 6: Examples of medical terms with their cross-lingual variants and contextualized sentences.

C.4 Basque

Tables 15 and 16 report the Basque results. The first table presents the shared WhisperX and WhisperX with hotword biasing baselines together with the Gemma and BioMistral variants, while the second continues the comparison with the LLaMA and Salamandra variants.

D Component Ablation

We conduct a staged ablation to quantify the contribution of the three main components of the proposed pipeline: (i) lexicon-guided N-best hypothesis selection (*N-best*); (ii) N-best selection followed by phonetic and textual n-gram correction (*N-best + Phonetic*); and (iii) the complete pipeline with LLM-based contextual filtering (*Full*). For each language and Whisper size, the full configuration uses the LLM selected based on the highest term accuracy, with WER as a tie-breaker. Table 17 reports WER, term-level accuracy, and average processing time per segment.

E Noise Robustness Analysis

To evaluate robustness beyond clean synthetic speech, we mixed the evaluation audio with hospital ambient noise at 10 dB SNR. The background recordings were obtained from the Hospital Ambient Noise Dataset, which contains equipment sounds, alarms, and background activity representative of clinical environments. Table 18 compares WER and term accuracy under clean conditions and hospital ambient noise at 10 dB SNR, using Salamandra-7B for every language and Whisper size.

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`.

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

Operation	Implementation
Case normalization	Lexicon entries and decoded text are converted to lowercase before matching.
Diacritic handling	Unicode NFD decomposition is applied and combining marks (`Mn`) are removed, enabling accent-insensitive matching.
Word extraction	Words are extracted with the Unicode-aware regular expression `\w+`; punctuation is ignored during lexical match counting.
Canonical form	A mapping from each normalized entry to its canonical lexicon form is retained for accepted replacements.
Phonetic input	Normalized candidate spans and lexicon terms are compacted by removing whitespace before Metaphone encoding.
Similarity score	RapidFuzz textual similarity and Metaphone similarity are combined as $0.6S_{\text{text}} + 0.4S_{\text{phon}}$.
Candidate indexing	Lexicon entries are indexed by the first character and compact string length. Only entries within ± 2 characters of the candidate length are considered.
Minimum length	Candidate spans shorter than four characters are discarded.
Exact-match filtering	Phonetic replacement is skipped when the normalized candidate already matches a lexicon entry.
Punctuation preservation	Leading punctuation from the first replaced token and trailing punctuation from the last replaced token are retained.
Capitalization	Uppercase, title-case, or lowercase patterns are transferred to the canonical replacement when applicable.

Table 8: Lexicon normalization, candidate retrieval, and output reconstruction rules used in the implementation.

		Baselines		WhisperX-LLM-TermFixer					
Whisper	**Metric**	**WhisperX**	**WhisperX + Hotword**	**Gemma 3-1B**	**Gemma 3-4B**	**Gemma 2-9B**	**BioMistral 7B**	**BioMistral 7B-dare**	**BioMistral 7B-slerp**
Tiny	WER ↓	18.91	20.70	16.84 [16.04–17.69]	16.82 [16.02–17.68]	16.85 [16.05–17.70]	16.85 [16.04–17.69]	**16.80** [16.00–17.65]	16.83 [16.03–17.67]
	Accuracy ↑	14.23	17.71	35.22 [32.34–38.21]	34.93 [32.04–37.91]	35.52 [32.54–38.51]	35.82 [32.94–38.81]	**36.12** [33.13–39.10]	35.82 [32.94–38.81]
	Time (s)	**0.234**	0.742	0.371	0.398	0.390	0.372	0.381	0.382
Base	WER ↓	13.66	14.97	11.08 [10.43–11.72]	11.12 [10.48–11.77]	11.08 [10.43–11.74]	11.09 [10.43–11.75]	**11.02** [10.37–11.68]	11.09 [10.44–11.73]
	Accuracy ↑	19.90	32.24	48.96 [45.87–52.14]	48.46 [45.37–51.54]	49.85 [46.77–52.94]	**49.95** [46.87–53.03]	**49.95** [46.87–53.13]	49.25 [46.17–52.34]
	Time (s)	**0.235**	0.740	0.392	0.424	0.410	0.397	0.397	0.400
Small	WER ↓	9.41	8.68	7.24 [6.74–7.74]	7.26 [6.76–7.77]	7.22 [6.72–7.73]	**7.21** [6.70–7.72]	7.22 [6.73–7.73]	**7.21** [6.71–7.71]
	Accuracy ↑	33.33	54.33	61.79 [58.71–64.78]	61.59 [58.51–64.58]	**62.29** [59.20–65.27]	62.09 [59.00–65.07]	62.09 [59.00–65.07]	62.09 [59.00–65.07]
	Time (s)	**0.269**	0.775	0.432	0.460	0.544	0.434	0.438	0.439
Medium	WER ↓	6.88	6.73	5.02 [4.63–5.41]	5.06 [4.67–5.46]	**4.97** [4.57–5.37]	4.98 [4.59–5.37]	**4.97** [4.58–5.36]	5.00 [4.60–5.39]
	Accuracy ↑	43.28	64.88	69.45 [66.57–72.24]	68.96 [66.07–71.74]	**70.55** [67.76–73.33]	70.45 [67.56–73.23]	**70.55** [67.76–73.33]	70.15 [67.26–72.94]
	Time (s)	**0.366**	0.884	0.537	0.568	0.567	0.545	0.540	0.538
Large-v3 turbo	WER ↓	6.50	6.03	4.62 [4.25–5.01]	4.66 [4.28–5.04]	**4.61** [4.24–4.99]	**4.61** [4.23–5.00]	4.62 [4.24–5.00]	4.64 [4.27–5.02]
	Accuracy ↑	44.88	60.40	72.54 [69.75–75.32]	72.14 [69.35–74.93]	**72.74** [69.95–75.42]	**72.74** [69.95–75.52]	72.54 [69.75–75.22]	72.44 [69.65–75.12]
	Time (s)	**0.282**	0.789	0.435	0.462	0.643	0.442	0.444	0.438
Large-v3	WER ↓	5.24	4.75	3.64 [3.33–3.98]	3.63 [3.32–3.96]	3.61 [3.30–3.93]	**3.58** [3.26–3.91]	3.59 [3.28–3.93]	3.61 [3.29–3.94]
	Accuracy ↑	55.62	71.64	78.21 [75.62–80.70]	78.41 [75.82–81.00]	78.81 [76.32–81.29]	78.71 [76.22–81.19]	**79.00** [76.42–81.59]	78.51 [75.92–81.09]
	Time (s)	**0.426**	0.984	0.604	0.642	0.818	0.608	0.606	0.601

Table 9: WER, term-level accuracy, and average processing time per segment for WhisperX, WhisperX with hotword biasing, and WhisperX-LLM-TermFixer using Gemma and BioMistral variants in English. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples. The WhisperX and WhisperX with hotword biasing results serve as shared baselines for all **English** TermFixer comparisons.

		WhisperX-LLM-TermFixer (cont.)				
Whisper	**Metric**	**LLaMA 3.2-1B**	**LLaMA 3.2-3B**	**LLaMA 3.1-8B**	**Salamandra 2B**	**Salamandra 7B**
Tiny	WER ↓	16.85 [16.04–17.70]	16.81 [16.00–17.66]	16.89 [16.08–17.74]	16.83 [16.02–17.67]	**16.79** [15.98–17.63]
	Accuracy ↑	35.92 [32.94–38.91]	36.12 [33.23–39.10]	35.52 [32.64–38.41]	35.62 [32.64–38.61]	**36.22** [33.23–39.20]
	Time (s)	**0.354**	0.366	0.386	0.427	0.439
Base	WER ↓	11.05 [10.39–11.70]	11.15 [10.49–11.80]	11.11 [10.45–11.75]	11.12 [10.46–11.78]	**11.04** [10.38–11.69]
	Accuracy ↑	**49.95** [46.87–53.13]	49.45 [46.37–52.64]	49.65 [46.57–52.84]	48.96 [45.87–52.04]	**49.95** [46.87–53.13]
	Time (s)	**0.370**	0.383	0.403	0.441	0.457
Small	WER ↓	7.23 [6.73–7.73]	7.24 [6.74–7.76]	7.28 [6.78–7.80]	7.27 [6.77–7.78]	**7.17** [6.68–7.69]
	Accuracy ↑	62.19 [59.10–65.17]	61.89 [58.81–64.88]	61.39 [58.31–64.38]	61.29 [58.21–64.28]	**62.59** [59.50–65.57]
	Time (s)	**0.412**	0.422	0.443	0.488	0.502
Medium	WER ↓	4.99 [4.59–5.38]	5.00 [4.60–5.39]	4.99 [4.60–5.39]	5.03 [4.63–5.43]	**4.93** [4.53–5.32]
	Accuracy ↑	70.35 [67.46–73.23]	70.15 [67.26–73.03]	70.45 [67.66–73.33]	69.45 [66.57–72.34]	**70.95** [68.06–73.83]
	Time (s)	**0.517**	0.535	0.536	0.654	0.673
Large-v3 turbo	WER ↓	4.55 [4.18–4.94]	4.59 [4.22–4.98]	4.62 [4.24–5.01]	4.63 [4.25–5.02]	**4.54** [4.17–4.93]
	Accuracy ↑	73.73 [70.95–76.42]	73.13 [70.35–75.82]	73.13 [70.35–75.82]	72.24 [69.35–75.02]	**73.83** [71.14–76.52]
	Time (s)	**0.418**	0.424	0.448	0.513	0.526
Large-v3	WER ↓	3.63 [3.31–3.96]	3.61 [3.30–3.95]	3.64 [3.33–3.97]	3.65 [3.34–3.99]	**3.55** [3.23–3.88]
	Accuracy ↑	78.61 [76.02–81.09]	79.00 [76.42–81.59]	78.91 [76.32–81.39]	78.21 [75.62–80.70]	**79.20** [76.62–81.69]
	Time (s)	**0.577**	0.592	0.601	0.734	0.750

Table 10: WER, term-level accuracy, and average processing time per segment for WhisperX-LLM-TermFixer using LLaMA and Salamandra variants in English. The shared WhisperX and WhisperX with hotword biasing baselines are reported in Table 9. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples.

		Baselines		WhisperX-LLM-TermFixer					
Whisper	**Metric**	**WhisperX**	**WhisperX + Hotword**	**Gemma 3-1B**	**Gemma 3-4B**	**Gemma 2-9B**	**BioMistral 7B**	**BioMistral 7B-dare**	**BioMistral 7B-slerp**
Tiny	WER ↓	17.61	18.16	14.60 [13.88–15.37]	14.57 [13.84–15.32]	14.60 [13.87–15.36]	14.57 [13.84–15.32]	14.60 [13.86–15.35]	**14.54** [13.80–15.30]
	Accuracy ↑	31.94	56.02	66.67 [63.78–69.45]	**67.06** [64.08–69.95]	66.77 [63.88–69.65]	66.77 [63.88–69.55]	66.47 [63.58–69.35]	**67.06** [64.18–69.95]
	Time (s)	**0.246**	0.918	0.433	0.464	0.470	0.448	0.450	0.442
Base	WER ↓	11.53	10.79	8.86 [8.28–9.45]	8.81 [8.23–9.41]	8.77 [8.19–9.37]	8.83 [8.26–9.42]	**8.72** [8.14–9.31]	**8.72** [8.14–9.30]
	Accuracy ↑	42.59	68.66	76.32 [73.63–79.00]	77.61 [75.02–80.20]	77.81 [75.22–80.40]	76.52 [73.83–79.10]	77.81 [75.32–80.40]	**77.91** [75.32–80.50]
	Time (s)	**0.252**	0.894	0.449	0.477	0.479	0.458	0.461	0.448
Small	WER ↓	6.54	5.54	4.52 [4.13–4.93]	**4.46** [4.08–4.88]	4.48 [4.10–4.89]	4.55 [4.17–4.96]	4.52 [4.14–4.93]	4.52 [4.14–4.93]
	Accuracy ↑	56.82	78.41	84.98 [82.69–87.06]	**85.97** [83.78–88.06]	**85.97** [83.78–88.06]	84.78 [82.49–86.97]	85.17 [82.89–87.26]	85.47 [83.28–87.66]
	Time (s)	**0.298**	0.958	0.494	0.517	0.516	0.501	0.502	0.483
Medium	WER ↓	3.74	4.33	2.51 [2.22–2.82]	**2.50** [2.21–2.80]	2.52 [2.23–2.82]	**2.50** [2.21–2.80]	2.51 [2.23–2.81]	2.54 [2.25–2.84]
	Accuracy ↑	70.85	85.97	90.35 [88.56–92.14]	**90.95** [89.15–92.74]	**90.95** [89.15–92.74]	**90.95** [89.15–92.64]	90.75 [88.96–92.44]	90.55 [88.76–92.34]
	Time (s)	**0.400**	1.108	0.614	0.623	0.624	0.610	0.611	0.612
Large-v3 turbo	WER ↓	2.85	2.07	1.91 [1.66–2.16]	1.87 [1.63–2.12]	**1.85** [1.61–2.11]	1.86 [1.62–2.11]	1.86 [1.62–2.11]	1.86 [1.62–2.11]
	Accuracy ↑	76.42	88.46	92.84 [91.24–94.33]	93.13 [91.54–94.63]	**93.63** [92.14–95.02]	93.23 [91.64–94.73]	93.03 [91.44–94.63]	92.94 [91.34–94.43]
	Time (s)	**0.283**	0.953	0.465	0.484	0.489	0.478	0.475	0.475
Large-v3	WER ↓	2.69	2.02	1.84 [1.60–2.08]	1.81 [1.57–2.05]	1.81 [1.57–2.06]	**1.79** [1.56–2.03]	1.82 [1.59–2.07]	1.82 [1.59–2.07]
	Accuracy ↑	77.71	89.35	93.33 [91.84–94.83]	**93.73** [92.24–95.22]	**93.73** [92.24–95.22]	**93.73** [92.24–95.22]	93.43 [91.84–94.93]	93.33 [91.74–94.83]
	Time (s)	**0.478**	1.243	0.687	0.689	0.708	0.677	0.691	0.680

Table 11: WER, term-level accuracy, and average processing time per segment for WhisperX, WhisperX with hotword biasing, and WhisperX-LLM-TermFixer using Gemma and BioMistral variants in Spanish. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples. The WhisperX and WhisperX with hotword biasing results serve as shared baselines for all **Spanish** TermFixer comparisons.

		WhisperX-LLM-TermFixer (cont.)				
Whisper	**Metric**	**LLaMA 3.2-1B**	**LLaMA 3.2-3B**	**LLaMA 3.1-8B**	**Salamandra 2B**	**Salamandra 7B**
Tiny	WER ↓	14.67 [13.95–15.42]	14.64 [13.92–15.40]	14.58 [13.85–15.35]	14.63 [13.90–15.39]	**14.52** [13.79–15.28]
	Accuracy ↑	65.97 [62.99–68.86]	66.47 [63.58–69.25]	66.87 [63.98–69.75]	**67.16** [64.28–70.05]	66.77 [63.88–69.55]
	Time (s)	**0.421**	0.437	0.452	0.585	0.611
Base	WER ↓	8.91 [8.33–9.50]	8.79 [8.22–9.38]	8.75 [8.17–9.33]	8.80 [8.23–9.40]	**8.73** [8.16–9.32]
	Accuracy ↑	76.02 [73.33–78.61]	77.41 [74.83–80.00]	77.61 [74.93–80.20]	77.61 [75.02–80.20]	**77.81** [75.22–80.40]
	Time (s)	**0.434**	0.450	0.463	0.595	0.613
Small	WER ↓	4.65 [4.27–5.06]	4.55 [4.17–4.96]	**4.46** [4.08–4.88]	4.50 [4.12–4.91]	4.51 [4.12–4.92]
	Accuracy ↑	84.38 [82.09–86.57]	84.98 [82.69–87.16]	**85.87** [83.68–87.96]	85.57 [83.38–87.66]	85.67 [83.48–87.76]
	Time (s)	**0.481**	0.496	0.503	0.663	0.676
Medium	WER ↓	2.56 [2.26–2.86]	**2.50** [2.22–2.80]	**2.50** [2.21–2.80]	**2.50** [2.21–2.80]	2.51 [2.22–2.80]
	Accuracy ↑	90.75 [88.96–92.54]	90.85 [89.05–92.54]	90.85 [89.05–92.54]	90.85 [89.05–92.64]	**90.95** [89.15–92.64]
	Time (s)	**0.592**	0.600	0.618	0.774	0.786
Large-v3 turbo	WER ↓	1.92 [1.67–2.17]	1.90 [1.66–2.16]	1.87 [1.63–2.13]	**1.85** [1.60–2.10]	1.87 [1.62–2.12]
	Accuracy ↑	92.64 [91.04–94.23]	92.74 [91.14–94.33]	**93.33** [91.84–94.83]	93.13 [91.54–94.63]	93.23 [91.64–94.73]
	Time (s)	**0.454**	0.470	0.482	0.572	0.580
Large-v3	WER ↓	1.84 [1.61–2.09]	1.83 [1.59–2.08]	**1.81** [1.58–2.06]	1.82 [1.58–2.06]	**1.81** [1.57–2.06]
	Accuracy ↑	93.53 [92.04–95.02]	93.43 [91.94–94.93]	**93.83** [92.34–95.32]	93.63 [92.14–95.12]	**93.83** [92.34–95.32]
	Time (s)	**0.662**	0.679	0.684	0.867	0.880

Table 12: WER, term-level accuracy, and average processing time per segment for WhisperX-LLM-TermFixer using LLaMA and Salamandra variants in Spanish. The shared WhisperX and WhisperX with hotword biasing baselines are reported in Table 11. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples.

		Baselines		WhisperX-LLM-TermFixer					
Whisper	**Metric**	**WhisperX**	**WhisperX + Hotword**	**Gemma 3-1B**	**Gemma 3-4B**	**Gemma 2-9B**	**BioMistral 7B**	**BioMistral 7B-dare**	**BioMistral 7B-slerp**
Tiny	WER ↓	43.91	48.45	41.88 [40.09–43.99]	41.80 [40.02–43.95]	**41.78** [40.00–43.92]	41.84 [40.06–44.00]	41.80 [40.01–43.93]	41.79 [39.99–43.93]
	Accuracy ↑	6.57	18.11	30.25 [27.46–33.13]	**30.75** [27.96–33.63]	**30.75** [27.86–33.63]	30.35 [27.46–33.23]	30.45 [27.66–33.33]	**30.75** [27.96–33.63]
	Time (s)	**0.288**	0.972	0.461	0.485	0.489	0.475	0.478	0.472
Base	WER ↓	30.54	33.61	28.02 [26.82–29.43]	**27.91** [26.72–29.31]	28.01 [26.81–29.42]	28.01 [26.81–29.42]	27.93 [26.75–29.35]	28.01 [26.81–29.42]
	Accuracy ↑	11.64	28.86	38.61 [35.62–41.69]	**39.30** [36.32–42.39]	38.21 [35.22–41.29]	38.31 [35.32–41.39]	38.51 [35.52–41.59]	38.91 [35.92–41.99]
	Time (s)	**0.307**	0.946	0.473	0.497	0.496	0.488	0.488	0.488
Small	WER ↓	18.02	17.49	15.72 [15.02–16.43]	**15.61** [14.91–16.31]	15.66 [14.97–16.36]	15.76 [15.06–16.47]	15.62 [14.94–16.33]	15.68 [14.99–16.40]
	Accuracy ↑	20.20	44.38	49.05 [45.97–52.14]	49.85 [46.77–53.03]	**50.05** [46.97–53.13]	48.96 [45.97–52.04]	49.85 [46.77–52.94]	49.55 [46.47–52.64]
	Time (s)	**0.377**	0.960	0.527	0.552	0.662	0.544	0.539	0.541
Medium	WER ↓	11.91	10.41	9.67 [9.12–10.20]	**9.54** [9.00–10.07]	9.55 [9.00–10.07]	9.63 [9.10–10.17]	9.59 [9.05–10.12]	9.59 [9.04–10.12]
	Accuracy ↑	26.07	52.64	58.51 [55.42–61.49]	59.80 [56.82–62.89]	**60.00** [56.92–62.99]	58.81 [55.72–61.79]	59.30 [56.32–62.29]	59.20 [56.22–62.19]
	Time (s)	**0.535**	1.071	0.655	0.677	0.678	0.670	0.665	0.666
Large-v3 turbo	WER ↓	9.03	7.63	6.99 [6.54–7.43]	6.91 [6.47–7.35]	**6.85** [6.41–7.29]	6.95 [6.51–7.39]	6.95 [6.51–7.40]	6.97 [6.53–7.40]
	Accuracy ↑	32.84	56.42	63.48 [60.50–66.47]	63.78 [60.80–66.77]	**64.58** [61.59–67.56]	63.78 [60.80–66.67]	63.48 [60.50–66.37]	63.18 [60.20–66.07]
	Time (s)	**0.382**	0.938	0.499	0.518	0.523	0.517	0.512	0.512
Large-v3	WER ↓	8.53	6.90	6.39 [5.95–6.84]	6.33 [5.88–6.77]	**6.28** [5.84–6.71]	6.34 [5.90–6.78]	6.30 [5.85–6.73]	6.31 [5.86–6.74]
	Accuracy ↑	37.01	62.99	69.35 [66.57–72.34]	69.25 [66.47–72.14]	**70.05** [67.26–72.94]	69.55 [66.67–72.34]	**70.05** [67.26–72.84]	69.75 [66.97–72.54]
	Time (s)	**0.691**	1.217	0.733	0.754	0.752	0.743	0.742	0.736

Table 13: WER, term-level accuracy, and average processing time per segment for WhisperX, WhisperX with hotword biasing, and WhisperX-LLM-TermFixer using Gemma and BioMistral variants in Catalan. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples. The WhisperX and WhisperX with hotword biasing results serve as shared baselines for all **Catalan** TermFixer comparisons.

		WhisperX-LLM-TermFixer (cont.)				
Whisper	**Metric**	**LLaMA 3.2-1B**	**LLaMA 3.2-3B**	**LLaMA 3.1-8B**	**Salamandra 2B**	**Salamandra 7B**
Tiny	WER ↓	41.88 [40.10–44.02]	41.79 [40.00–43.94]	41.75 [39.97–43.90]	**41.71** [39.93–43.88]	**41.71** [39.94–43.88]
	Accuracy ↑	30.25 [27.46–33.03]	30.65 [27.76–33.53]	31.24 [28.36–34.03]	30.55 [27.76–33.43]	**31.34** [28.46–34.23]
	Time (s)	**0.452**	0.467	0.476	0.456	0.472
Base	WER ↓	28.03 [26.83–29.43]	27.99 [26.79–29.40]	27.94 [26.74–29.34]	27.90 [26.71–29.31]	**27.78** [26.59–29.19]
	Accuracy ↑	38.01 [35.02–41.09]	39.00 [36.02–42.09]	39.20 [36.22–42.29]	39.40 [36.42–42.39]	**39.80** [36.82–42.89]
	Time (s)	**0.465**	0.476	0.493	0.469	0.487
Small	WER ↓	15.70 [15.01–16.41]	15.69 [15.00–16.41]	15.63 [14.93–16.34]	15.56 [14.87–16.28]	**15.51** [14.83–16.23]
	Accuracy ↑	49.65 [46.67–52.74]	49.55 [46.37–52.74]	50.45 [47.36–53.53]	50.45 [47.36–53.53]	**50.65** [47.66–53.73]
	Time (s)	**0.517**	0.529	0.545	0.523	0.545
Medium	WER ↓	9.61 [9.06–10.15]	9.54 [9.00–10.06]	9.54 [8.99–10.06]	9.54 [8.99–10.06]	**9.50** [8.96–10.03]
	Accuracy ↑	59.30 [56.22–62.29]	60.00 [57.01–62.99]	**60.20** [57.11–63.18]	59.90 [56.72–62.89]	60.10 [57.01–63.08]
	Time (s)	**0.625**	0.659	0.673	0.647	0.658
Large-v3 turbo	WER ↓	7.06 [6.61–7.51]	6.92 [6.48–7.36]	6.83 [6.39–7.27]	6.93 [6.50–7.37]	**6.80** [6.37–7.23]
	Accuracy ↑	63.38 [60.40–66.37]	64.68 [61.79–67.56]	65.07 [62.19–68.06]	63.68 [60.70–66.67]	**65.17** [62.29–68.16]
	Time (s)	0.493	0.498	0.520	**0.492**	0.552
Large-v3	WER ↓	6.40 [5.95–6.84]	6.35 [5.91–6.79]	6.30 [5.85–6.73]	6.29 [5.84–6.73]	**6.22** [5.78–6.65]
	Accuracy ↑	69.55 [66.77–72.34]	69.95 [67.16–72.84]	70.15 [67.36–73.03]	70.15 [67.36–73.03]	**70.35** [67.56–73.23]
	Time (s)	0.731	0.731	0.745	**0.726**	0.894

Table 14: WER, term-level accuracy, and average processing time per segment for WhisperX-LLM-TermFixer using LLaMA and Salamandra variants in Catalan. The shared WhisperX and WhisperX with hotword biasing baselines are reported in Table 13. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples.

		Baselines		WhisperX-LLM-TermFixer					
Whisper	**Metric**	**WhisperX**	**WhisperX + Hotword**	**Gemma 3-1B**	**Gemma 3-4B**	**Gemma 2-9B**	**BioMistral 7B**	**BioMistral 7B-dare**	**BioMistral 7B-slerp**
Tiny	WER ↓	97.32	104.16	94.36 [92.76–96.36]	94.30 [92.71–96.27]	**94.18** [92.59–96.18]	94.56 [92.96–96.56]	94.38 [92.79–96.37]	94.40 [92.81–96.42]
	Accuracy ↑	6.67	26.87	29.85 [27.06–32.74]	30.65 [27.86–33.53]	31.64 [28.76–34.53]	32.14 [29.25–35.12]	**33.03** [30.05–35.92]	32.54 [29.65–35.52]
	Time (s)	**0.264**	0.683	0.462	0.488	0.492	0.477	0.475	0.477
Base	WER ↓	90.99	94.23	87.91 [86.88–88.96]	87.81 [86.78–88.87]	87.72 [86.69–88.77]	87.95 [86.93–89.01]	**87.80** [86.77–88.85]	87.92 [86.90–88.99]
	Accuracy ↑	12.54	28.66	35.82 [32.74–38.71]	36.52 [33.53–39.40]	36.62 [33.63–39.50]	38.41 [35.32–41.39]	**39.50** [36.52–42.49]	38.41 [35.42–41.39]
	Time (s)	**0.274**	0.638	0.481	0.507	0.513	0.492	0.492	0.494
Small	WER ↓	71.24	72.27	68.02 [66.95–69.09]	**67.86** [66.81–68.93]	67.92 [66.86–68.97]	68.10 [67.04–69.18]	68.04 [66.98–69.11]	68.10 [67.04–69.15]
	Accuracy ↑	15.42	**46.57**	40.80 [37.71–43.78]	41.39 [38.31–44.38]	42.69 [39.60–45.67]	43.58 [40.60–46.57]	**43.88** [40.80–46.97]	43.28 [40.20–46.37]
	Time (s)	**0.324**	0.660	0.525	0.553	0.556	0.540	0.541	0.539
Medium	WER ↓	57.06	57.15	53.74 [52.63–54.90]	53.50 [52.38–54.65]	**53.44** [52.31–54.60]	53.90 [52.79–55.06]	53.82 [52.70–54.97]	53.87 [52.75–55.04]
	Accuracy ↑	17.51	**52.64**	43.48 [40.40–46.57]	44.38 [41.29–47.36]	45.57 [42.49–48.66]	44.68 [41.69–47.76]	**45.97** [42.89–49.05]	44.88 [41.89–47.96]
	Time (s)	**0.447**	0.839	0.664	0.693	0.697	0.687	0.684	0.686
Large-v3 turbo	WER ↓	35.28	33.59	32.93 [31.95–33.93]	32.74 [31.76–33.76]	**32.59** [31.60–33.60]	33.06 [32.09–34.08]	33.01 [32.03–34.02]	33.05 [32.06–34.06]
	Accuracy ↑	31.24	57.61	56.82 [53.73–59.80]	57.61 [54.53–60.60]	**59.40** [56.32–62.29]	58.11 [55.02–61.09]	58.81 [55.72–61.79]	58.41 [55.32–61.39]
	Time (s)	**0.315**	0.643	0.493	0.516	0.520	0.507	0.506	0.508
Large-v3	WER ↓	36.61	34.61	34.53 [33.58–35.54]	34.30 [33.35–35.29]	**34.21** [33.25–35.22]	34.67 [33.74–35.64]	34.47 [33.54–35.48]	34.68 [33.72–35.66]
	Accuracy ↑	28.26	**59.20**	52.64 [49.45–55.72]	52.94 [49.85–56.02]	54.53 [51.34–57.51]	54.13 [50.95–57.11]	**55.82** [52.64–58.81]	54.63 [51.44–57.61]
	Time (s)	**0.542**	0.950	0.739	0.768	0.771	0.756	0.757	0.756

Table 15: WER, term-level accuracy, and average processing time per segment for WhisperX, WhisperX with hotword biasing, and WhisperX-LLM-TermFixer using Gemma and BioMistral variants in Basque. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples. The WhisperX and WhisperX with hotword biasing results serve as shared baselines for all **Basque** TermFixer comparisons.

		WhisperX-LLM-TermFixer (cont.)				
Whisper	**Metric**	**LLaMA 3.2-1B**	**LLaMA 3.2-3B**	**LLaMA 3.1-8B**	**Salamandra 2B**	**Salamandra 7B**
Tiny	WER ↓	94.15 [92.56–96.14]	94.11 [92.52–96.11]	94.09 [92.49–96.09]	93.89 [92.31–95.91]	**93.76** [92.15–95.76]
	Accuracy ↑	33.33 [30.45–36.22]	32.14 [29.15–35.12]	32.44 [29.45–35.42]	33.83 [30.95–36.82]	**34.93** [31.94–37.91]
	Time (s)	**0.447**	0.464	0.481	0.455	0.478
Base	WER ↓	87.63 [86.60–88.70]	87.63 [86.62–88.69]	87.54 [86.52–88.61]	87.42 [86.40–88.48]	**87.32** [86.30–88.36]
	Accuracy ↑	40.00 [37.01–42.99]	39.30 [36.32–42.29]	39.30 [36.22–42.19]	40.60 [37.51–43.58]	**41.89** [38.91–44.88]
	Time (s)	**0.465**	0.480	0.500	0.472	0.496
Small	WER ↓	67.59 [66.53–68.64]	67.50 [66.44–68.54]	67.53 [66.49–68.59]	67.64 [66.56–68.70]	**67.47** [66.42–68.54]
	Accuracy ↑	45.87 [42.89–48.86]	45.37 [42.29–48.36]	45.27 [42.19–48.26]	44.98 [41.89–47.96]	**46.27** [43.28–49.35]
	Time (s)	**0.513**	0.528	0.544	0.521	0.540
Medium	WER ↓	53.53 [52.41–54.70]	53.21 [52.10–54.35]	53.20 [52.07–54.36]	53.19 [52.07–54.35]	**53.05** [51.94–54.21]
	Accuracy ↑	46.27 [43.18–49.25]	46.87 [43.68–49.95]	46.77 [43.68–49.85]	46.87 [43.68–49.95]	**47.76** [44.58–50.85]
	Time (s)	**0.654**	0.673	0.689	0.665	0.686
Large-v3 turbo	WER ↓	32.80 [31.81–33.81]	32.67 [31.68–33.68]	32.62 [31.63–33.63]	32.65 [31.67–33.66]	**32.44** [31.46–33.46]
	Accuracy ↑	57.91 [54.83–61.00]	58.81 [55.72–61.79]	58.81 [55.62–61.79]	59.30 [56.22–62.39]	**60.60** [57.61–63.58]
	Time (s)	**0.483**	0.497	0.510	0.491	0.509
Large-v3	WER ↓	34.41 [33.45–35.41]	34.30 [33.35–35.29]	34.12 [33.18–35.10]	34.16 [33.21–35.16]	**33.95** [33.01–34.95]
	Accuracy ↑	54.83 [51.64–57.81]	54.23 [51.14–57.31]	55.52 [52.34–58.51]	55.32 [52.24–58.31]	**56.02** [52.94–59.00]
	Time (s)	**0.729**	0.746	0.758	0.738	0.757

Table 16: WER, term-level accuracy, and average processing time per segment for WhisperX-LLM-TermFixer using LLaMA and Salamandra variants in Basque. The shared WhisperX and WhisperX with hotword biasing baselines are reported in Table 15. Point estimates for WER and accuracy are accompanied by utterance-level bootstrap 95% confidence intervals based on 5,000 resamples.

		N-best			N-best + Phonetic			Full		
Language	**Whisper**	**WER**	**Acc.**	**Time**	**WER**	**Acc.**	**Time**	**WER**	**Acc.**	**Time**
English	Tiny	18.77	19.30	**0.341**	19.83	31.14	0.350	**16.79**	**36.22**	0.439
	Base	13.48	27.66	**0.344**	14.24	40.20	0.360	**11.02**	**49.95**	0.397
	Small	8.87	44.08	**0.380**	9.25	54.63	0.399	**7.17**	**62.59**	0.502
	Medium	6.26	55.12	**0.448**	6.76	64.28	0.468	**4.93**	**70.95**	0.673
	Large-v3-turbo	5.86	58.71	**0.387**	5.76	68.16	0.388	**4.54**	**73.83**	0.526
	Large-v3	4.61	67.76	**0.521**	5.00	71.64	0.550	**3.55**	**79.20**	0.750
Spanish	Tiny	17.07	39.90	**0.384**	18.80	61.29	0.392	**14.63**	**67.16**	0.585
	Base	10.82	54.33	**0.398**	11.13	70.35	0.405	**8.72**	**77.91**	0.448
	Small	5.63	70.75	**0.448**	6.00	78.71	0.457	**4.46**	**85.97**	0.517
	Medium	3.06	82.09	**0.550**	3.61	86.27	0.556	**2.50**	**90.95**	0.623
	Large-v3-turbo	2.18	87.86	**0.431**	2.76	88.66	0.438	**1.85**	**93.63**	0.489
	Large-v3	2.11	87.66	**0.633**	2.68	89.45	0.639	**1.81**	**93.83**	0.880
Catalan	Tiny	44.13	8.86	**0.426**	48.49	28.56	0.435	**41.71**	**31.34**	0.472
	Base	30.47	15.92	**0.440**	33.02	35.52	0.443	**27.78**	**39.80**	0.487
	Small	17.83	26.77	**0.483**	18.09	45.27	0.497	**15.51**	**50.65**	0.545
	Medium	11.27	37.61	**0.575**	10.72	52.94	0.606	**9.54**	**60.20**	0.673
	Large-v3-turbo	8.17	46.87	**0.464**	7.64	58.71	0.471	**6.80**	**65.17**	0.552
	Large-v3	7.10	54.23	**0.681**	7.17	61.79	0.697	**6.22**	**70.35**	0.894
Basque	Tiny	98.29	7.46	**0.419**	100.85	33.43	0.422	**93.76**	**34.93**	0.478
	Base	91.47	12.74	**0.436**	91.59	39.90	0.441	**87.32**	**41.89**	0.496
	Small	70.76	17.91	**0.479**	72.99	**46.87**	0.515	**67.47**	46.27	0.540
	Medium	55.99	20.10	**0.587**	57.87	**48.26**	0.621	**53.05**	47.76	0.686
	Large-v3-turbo	36.13	36.62	**0.458**	35.31	59.90	0.485	**32.44**	**60.60**	0.509
	Large-v3	35.90	33.33	**0.683**	35.38	**58.71**	0.721	**33.95**	56.02	0.757

Table 17: Staged ablation of the proposed pipeline across languages and Whisper model sizes. WER, term-level accuracy (%), and average processing time per segment (s) are reported separately for lexicon-guided N-best hypothesis selection (*N-best*), N-best selection followed by phonetic and textual n-gram correction (*N-best + Phonetic*), and the complete pipeline with LLM-based contextual filtering (*Full*). The lowest WER, highest term accuracy, and lowest processing time in each row are shown in bold.

		WER ↓			Accuracy (%) ↑		
Language	**Whisper**	**Clean**	**10 dB**	**Δ**	**Clean**	**10 dB**	**Δ**
English	Tiny	16.79	17.11	+0.32	36.22	36.52	+0.30
	Base	11.04	11.72	+0.68	49.95	48.06	-1.89
	Small	7.17	7.52	+0.35	62.59	61.00	-1.59
	Medium	4.93	5.22	+0.29	70.95	69.65	-1.30
	Large-v3-turbo	4.54	4.77	+0.23	73.83	74.53	+0.70
	Large-v3	3.55	3.84	+0.29	79.20	78.61	-0.59
Spanish	Tiny	14.52	14.87	+0.35	66.77	65.17	-1.60
	Base	8.73	9.05	+0.32	77.81	77.31	-0.50
	Small	4.51	4.59	+0.08	85.67	85.87	+0.20
	Medium	2.51	2.56	+0.05	90.95	90.85	-0.10
	Large-v3-turbo	1.87	1.90	+0.03	93.23	93.03	-0.20
	Large-v3	1.81	1.83	+0.02	93.83	93.33	-0.50
Catalan	Tiny	41.71	42.10	+0.39	31.34	32.04	+0.70
	Base	27.78	28.35	+0.57	39.80	40.10	+0.30
	Small	15.51	15.82	+0.31	50.65	50.55	-0.10
	Medium	9.50	9.51	+0.01	60.10	60.50	+0.40
	Large-v3-turbo	6.80	6.87	+0.07	65.17	64.68	-0.49
	Large-v3	6.22	6.01	-0.21	70.35	69.75	-0.60
Basque	Tiny	93.76	95.30	+1.54	34.93	35.02	+0.09
	Base	87.32	88.52	+1.20	41.89	40.80	-1.09
	Small	67.47	67.92	+0.45	46.27	46.37	+0.10
	Medium	53.05	52.95	-0.10	47.76	47.06	-0.70
	Large-v3-turbo	32.44	34.10	+1.66	60.60	58.51	-2.09
	Large-v3	33.95	33.79	-0.16	56.02	55.72	-0.30

Table 18: Noise-robustness results for WhisperX-LLM-TermFixer using Salamandra-7B. Hospital ambient noise is mixed with the evaluation audio at 10 dB SNR. For WER and term-level accuracy, clean and noisy results are reported together with Δ, computed as the 10 dB result minus the clean result. Positive WER deltas and negative accuracy deltas indicate degradation under noise.