A Dataset for Evaluating ASR on Specialized Vocabulary

Emily Haubert Klering[1], Eduardo Gabriel Cortes[1,2], Tatjana Chernenko[3], Mariana Vargas Trarbach[1], Gabriel de Oliveira Ramos[1], Sandro José Rigo[1], Maitê Dupont[4], Ana Luiza Treichel Vianna[2], Gabriela Krause dos Santos[1], Vinicius Meirelles Pereira[2], Denis Andrei de Araujo[1], and Rafael Kunst[1,2]

[1]UNISINOS, Brazil
[2]SAP SE, Brazil
[3]SAP SE, Germany
[4]SAP SE, Ireland
{emilyklering, egcortes, gdoramos, rigo, krause, denandrei, rafaelkunst}@unisinos.br
{tatjana.chernenko, maite.dupont, ana.treichel.vianna, vinicius.meirelles}@sap.com
trarbachmv@gmail.com

Abstract

Evaluating the ability of Automatic Speech Recognition (ASR) models to transcribe specialized vocabulary remains a persistent challenge, as standard datasets predominantly feature common words and thus obscure weaknesses on rare or out-of-vocabulary (OOV) terms. To address this limitation, we introduce a linguistically curated bilingual dataset (English and Portuguese) comprising 13,846 utterances (18.7 hours) distributed across synthetic and literature-derived subsets, with OOV rates reaching up to 100%. We further propose a diagnostic evaluation framework that partitions recognition performance into Biased Word Error Rate (B-WER), which targets domain-specific jargon, and Unbiased Word Error Rate (U-WER), which focuses on general vocabulary. Baseline evaluations with Whisper models (`medium`, `large-v3`, and `large-v3-turbo`) confirm the need for this framework. On the most challenging datasets, B-WER ranges from 0.88 to 0.90, whereas U-WER remains low, ranging from 0.06 to 0.19, demonstrating that conventional WER masks critical failure modes in jargon recognition. Additionally, an oracle upper-bound experiment shows that providing correct jargon via prompting reduces B-WER by 0.50–0.70 absolute, quantifying the considerable potential for contextual biasing. We release the datasets and evaluation scripts as a reproducible benchmark to foster research on domain-aware contextual biasing and OOV handling in ASR systems.

Keywords: Automatic Speech Recognition, Jargon Recognition, Contextual Biasing

1. Introduction

Although modern Automatic Speech Recognition (ASR) systems achieve remarkable accuracy on general-purpose benchmarks, their performance often degrades when deployed in specialized domains such as finance, medicine, or manufacturing (Li et al., 2024b). This degradation is particularly evident when transcribing domain-specific jargon, product names, or other terms that are rare or absent from the training data (Li et al., 2024a; Jogi et al., 2025). Inability to accurately recognize this specialized vocabulary can result in critical misunderstandings, compromising aspects ranging from patient safety in clinical documentation to the reliability of financial reports (Afonja et al., 2024; Ng et al., 2025).

Conventional evaluation practices mostly hide this performance gap. Standard ASR datasets such as LibriSpeech (Panayotov et al., 2015) and Common Voice (Ardila et al., 2020), while invaluable for measuring overall Word Error Rate (WER), are misaligned with the challenge of recognizing rare words. Their vocabularies are dominated by high-frequency, general-domain terms, reflecting a distribution where a small number of words account for the vast majority of content (Vietti et al., 2024; Solberg et al., 2023; Zuluaga-Gomez et al., 2023). Consequently, these benchmarks rarely expose ASR systems to the domain-specific jargon that can break transcription pipelines in real-world deployments, leading to erroneous transcriptions and jeopardizing communication in business and formal settings. This evaluation paradigm underestimates the task's difficulty and, consequently, masks the true impact of innovations for domain adaptation and OOV handling (Vuong et al., 2023; Hwang et al., 2022).

To provide concrete evidence of this masked weakness, we evaluated a state-of-the-art Whisper model on several proposed datasets that contain a high density of specialized terms. As illustrated in Table 1, the results reveal a massive performance difference: while the Unbiased WER (U-WER) on common words remains low, the Biased WER (B-WER) on jargon terms is exceptionally high. This occurs because the model, without prior exposure, cannot rely on learned statistical patterns and must rely on phonetic evidence alone, often failing. The reduction in B-WER when a prompt containing bi-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 470–480
11–16 May 2026.

Table 1: Performance disparity on common vs. specialized vocabulary and the impact of an oracle prompt. The table groups results by vocabulary type to directly compare performance under standard and prompted conditions. Error rates are averages across several datasets (the lower, the better).

Vocabulary Type	Example Terms	Experimental Condition	Error Rate (↓)
Common Terms	*"ice cream"*, *"person"*	Standard (No prompt)	0.182 (U-WER)
		Jargon-Informed Prompt	0.122 (U-WER)
Jargon (Biased) Terms	*"Kalousis"*, *"Chaloem Phra Kiat"*	Standard (No prompt)	0.752 (B-WER)
		Jargon-Informed Prompt	0.397 (B-WER)

ased words is provided underscores the potential of context-aware ASR systems, yet also highlights the need for benchmarks to drive their development.

Addressing this challenge requires extensive and representative data (Fazel et al., 2021). However, collecting naturally occurring speech is both time-consuming and costly (Dhamyal et al., 2024). Additionally, organizations may be reluctant to share internal or sensitive information, and privacy laws hinder the collection and labeling of human audio data (Banerjee et al., 2024), further complicating efforts to obtain authentic and domain-specific datasets. A promising solution lies in the development of synthetic datasets (Fazel et al., 2021; Zheng et al., 2021; Banerjee et al., 2024; Melechovsky et al., 2024), which enable the controlled injection of jargon at scale and are augmented with challenging examples curated from the existing literature.

This paper directly addresses this evaluation challenge. We introduce a suite of linguistically curated datasets in English and Portuguese designed specifically to benchmark ASR performance on specialized vocabulary[1]. We develop novel datasets where OOV jargon is systematically introduced and present a methodology for curating hard-to-transcribe examples from existing corpora. Using these resources, we establish a comprehensive baseline that quantifies the limitations of current models and demonstrates significant potential for improvement.

The main contributions of this work are:

- A collection of novel, synthetic datasets in English and Portuguese built using a dual-source pipeline: (1) the generation of entirely novel, 100% OOV technical terms, and (2) the extraction of challenging, foreign-origin proper nouns from Wikidata.
- A semi-automated curation methodology for extracting challenging, naturally-occurring jargon from public corpora. This method uses a baseline model's transcription failures to identify jargon terms, which human linguists then validate.
- A diagnostic evaluation framework and a comprehensive baseline evaluation using different Whisper models. This framework uses a B-WER / U-WER split to assess performance on jargon versus general vocabulary. Our baselines use this framework to quantify a large performance gap and demonstrate a failure mode masked by standard WER.

[1]The datasets and evaluation scripts are available for reproducibility at: https://github.com/eduardogc8/ASR-Jargon-Dataset-Code

2. Related Work

The difficulty of recognizing OOV and low-frequency terms is a well-documented challenge in ASR, with model performance often correlating directly with word frequency (Vietti et al., 2024; Solberg et al., 2023). Research to address this gap has evolved along two primary directions: data-centric augmentation and model-centric adaptation.

A foundational data-centric strategy is the use of synthetic data, leveraging a pipeline of LLMs and Text-to-Speech (TTS) systems to create large-scale, in-domain corpora (Fazel et al., 2021; Banerjee et al., 2024; Nguyen et al., 2024; Zheng et al., 2021). This approach can be used for adapting pre-trained models to new domains, such as sports or music, without requiring real data, as demonstrated by frameworks like DAS (Tran et al., 2025), and for creating complex, multi-speaker conversational data that is otherwise difficult to collect (Cornell et al., 2024). However, the field is rapidly moving beyond simple data generation towards more intelligent synthesis. For instance, the HardSynth framework (Yu et al., 2024) focuses on creating "hard samples" by using an LLM to rewrite text and to clone the acoustic styles of utterances that a model misrecognizes, thereby addressing the model's specific weaknesses in a data-efficient manner.

In parallel, model-centric approaches adapt the ASR model's architecture or inference process. A key technique is contextual biasing, which dynamically increases the probability of a predefined list of relevant terms during decoding (Sheikh et al.,

2022). Recent work has improved this by addressing the recognition of multi-word entities, proposing phrase biasing to ensure that contextual phrases are treated as integrated units rather than disconnected subwords (Lin et al., 2025). To move beyond static bias lists, researchers are also employing Retrieval-Augmented Generation, which allows the ASR system to automatically discover and use relevant context from the ongoing discourse without requiring a predefined vocabulary (Mathur et al., 2024).

An innovative trend is the convergence of these two paradigms: using synthetic data specifically to train model-centric components. Researchers now generate customized audio containing rare words to train contextual biasing modules. Nevertheless, introduces the risk of the model overfitting to acoustic artifacts in the synthetic TTS audio. To mitigate this, specialized solutions, such as keyword-aware loss functions, are being developed to force the model to focus on the phonetic characteristics of the target words, rather than the synthetic voice (Kwok et al., 2025).

This progression from simple data augmentation to intelligent synthesis and highly specialized model adapters highlights a limitation. The efficacy of these jargon-focused methods cannot be reliably measured using standard benchmarks like LibriSpeech, where high-frequency general vocabulary dominates and masks the specific failure modes these methods aim to address. Consequently, it remains difficult to compare the true impact of different domain-adaptation and contextual-biasing strategies. Our work directly confronts this evaluation bottleneck. We propose a suite of specialized datasets, not as a training corpus, but as a new resource for reproducibly quantifying the performance of these solutions on their primary target: specialized and out-of-vocabulary terminology.

3. Methodology

To construct new datasets for evaluating ASR on specialized vocabulary containing jargon-like terms, we developed two distinct classes of datasets in both English and Portuguese: synthetic corpora designed for controlled testing, and curated subsets derived from existing literature to capture naturally occurring challenges. This dual approach allows for a comprehensive assessment of model performance on both entirely novel terms and difficult real-world jargon.

3.1. Synthetic Datasets

Our synthetic datasets were created through a three-stage pipeline: (1) jargon term acquisition, (2) contextual sentence generation, and (3) TTS audio synthesis. We employed two strategies in the first stage, yielding two distinct sets of synthetic corpora: *Synthetic Terms* and *Wikidata*.

The **Synthetic Terms** datasets were designed to test ASR models on completely OOV terms that they could not have encountered during pretraining. To create these terms, linguists first analyzed domain-specific corpora to identify the morphological and phonological patterns of existing jargon. Based on this analysis, we engineered specialized prompts for an LLM (GPT-4o) to generate a list of entirely new but plausible-sounding, technical terms (e.g., product, process, and material names).

The **Wikidata** datasets were created to address the common ASR challenge of transcribing foreign terms, particularly proper nouns with phonetic structures that are foreign to Portuguese and English. We sourced these terms from Wikidata (Vrandečić and Krötzsch, 2014), a multilingual knowledge base that offers greater linguistic diversity than more Anglocentric resources (Piscopo et al., 2017). Our selection process targeted named entities (e.g., people and places) that included an International Phonetic Alphabet (IPA) transcription, which was crucial for ensuring accurate pronunciation during audio synthesis.

Once the jargon lists for both dataset types were complete, we automatically embedded the terms into context-specific sentences that simulate business meeting scenarios. These sentences were then synthesized into audio using KokoroTTS (hexgrad, 2025), with variations in speaking rate, voice, and accent to increase diversity. For the English datasets, KokoroTT used the provided IPA strings to render Wikidata terms accurately within English sentences.

Since KokoroTTS does not yet support IPA for Portuguese, we developed a process called **Phonetic Approximation in Brazilian Portuguese (PABP)**. Linguists created an initial set of orthographic adaptations to guide the TTS engine toward the correct Brazilian Portuguese phonemes for foreign-sounding terms, grounding these changes in the IPA transcriptions. For example, the term 'Imgur' (IPA: /mdr/) was transformed into 'ímdjãr' through PABP. This process was then scaled by providing GPT-4o with the linguistic guidelines and examples in a prompt, which generated PABP versions for the entire list of Portuguese Wikidata terms.

Finally, to ensure the quality of the datasets, all synthesized audio files were reviewed by linguists, who discarded any instances with omissions, additions, or significant pronunciation errors. Pronunciation errors were detected by comparing the audio with the original IPA transcription. Cases of phonological mismatch were discarded. This rigorous

curation step ensures that transcription errors during evaluation can be attributed to the ASR model's limitations rather than to flaws in the audio data.

3.2. Literature-Derived Datasets

The second class of proposed datasets consists of challenging subsets curated from existing, well-known corpora such as Earnings-22 (Rio et al., 2022), SPGISpeech (O'Neill et al., 2021), FLEURS (Conneau et al., 2022), CHiME-6 (Barker et al., 2018; Watanabe et al., 2020), CORAA (Junior et al., 2021) and LapsBM[2]. These datasets provide naturally occurring speech that contains the type of domain-specific jargon our task targets.

To isolate the most difficult samples, we developed a semi-automated curation methodology. First, we processed the full datasets with a state-of-the-art ASR model, Whisper `large-v3` (Radford et al., 2023). We then filtered the output to select only utterances with errors in the model's transcription. A team of linguistic experts subsequently reviewed these failed transcriptions to identify cases where the errors were directly related to the misrecognition of jargon terms (defined as products, processes, materials, or proper nouns). The experts then annotated the correct jargon in a new column, as this information is useful for metrics focused on jargon transcription during ASR model evaluations. The result is a collection of challenging test cases with an expert-annotated column identifying the correct jargon terms that specifically target the weaknesses of a strong baseline model in recognizing domain-specific vocabulary.

3.3. Dataset Statistics

The datasets' statistics are present in Table 2. The datasets are grouped by language and type (derived from existing literature or synthetically generated). The statistics highlight the diversity of the corpora in terms of size, jargon density, and OOV rate.

In the context of this work, a jargon term is considered OOV if at least one of its constituent words is not present in the vocabulary of the Whisper `large-v3` tokenizer for the corresponding language. This definition is important because it directly measures whether a model recognizes the fundamental components of a specialized term as uncommon during its pre-training. Analysis of OOV rates helps understand the lexical challenges a corpus presents to ASR models (Tachbelie et al., 2020). A high OOV rate thus represents a challenge for transcription, as the model cannot rely on its existing vocabulary and must infer the word from phonetic evidence alone.

[2]https://github.com/falabrasil/gitlab-resources

An analysis of Table 2 reveals the datasets' challenging nature across several dimensions. First, most datasets exhibit a high OOV rate, frequently exceeding 60%, confirming their utility for testing on unseen terminology. This is particularly true for our synthetic datasets, with the `Synthetic Terms` corpora reaching a 100% OOV rate and the `Wikidata` corpora over 87%. Second, the datasets present structural complexity. The `Wikidata (en)` and `FLEURS (pt)` datasets, for example, contain long multi-word entities with maximum lengths of 15 and 9 words, respectively. This challenges a model's ability to handle complex named entities requiring segmentation. Third, our synthetic datasets feature the highest jargon density, with an average of over 1.5 specialized terms per utterance, ensuring that evaluation is consistently focused on the target issue. Finally, some datasets pose unique challenges; for instance, while `CHiME-6` has a low OOV rate (10.0%), it is characterized by considerable background noise, which tests the model's robustness in a different way. Finally, it is worth noting that the `CHiME-6 (en)` and `LapsBM (pt)` subsets are considerably smaller. While they contain valuable and challenging examples, results derived from them should be considered exploratory and illustrative of model behavior rather than statistically conclusive.

To provide a more concrete understanding of the types of jargon present in our datasets, Table 3 lists representative examples from each dataset. These examples span a range of categories, including financial terminology, proper nouns, technical specifications, and novel words.

4. Experiments

To establish a strong baseline performance on our proposed datasets, we evaluated different versions of the Whisper model: `medium`, `large-v3`, and `large-v3-turbo`. These models are known for their high transcription accuracy and serve as a starting point for our analysis.

We conducted experiments under two conditions. The first is a standard transcription setup, in which the models process the audio without any additional context. The second condition uses a prompting mechanism, in which the jargon terms in each utterance are provided to the decoder as an initial prompt. This prompting approach is not intended to simulate a realistic application, as it requires prior knowledge of the utterance's content. Instead, it serves as an "oracle" baseline, demonstrating the upper bound of performance achievable if an ASR system had perfect, in-context knowledge of the specialized vocabulary. This helps quantify the potential gains from techniques that aim to identify and supply relevant jargon during transcription, such as

Table 2: Key statistics of the datasets. Utts.: number of utterances; Dur(h): total duration in hours; J/U: average jargon terms per utterance; Max W. and Avg W.: maximum and average jargon length in words, respectively; OOV(%): percentage of jargon terms that are OOV for the Whisper `large-v3` tokenizer.

				Jargon Terms			Jargon Length		
Dataset	**Type**	**Utts.**	**Dur(h)**	**Total**	**Uniq.**	**J/U**	**Max W.**	**Avg W.**	**OOV(%)**
English Datasets									
CHiME-6 (en)	Lit.	38	0.03	43	40	1.13	2	1.33	10.0
Earnings-22 (en)	Lit.	515	1.33	629	595	1.22	4	1.42	53.1
FLEURS (en)	Lit.	285	0.92	366	351	1.28	4	1.09	71.2
GLOBE (en)	Lit.	4,297	4.25	4,551	3,924	1.06	5	1.33	60.6
SPGISpeech (en)	Lit.	376	1.02	394	368	1.05	4	1.12	70.7
Synth. Terms (en)	Synth.	520	0.71	859	246	1.65	2	1.00	100.0
Wikidata Synth. (en)	Synth.	5,235	6.67	9,165	2,620	1.75	15	1.49	88.8
Portuguese Datasets									
CORAA (pt)	Lit.	505	0.69	549	513	1.09	6	1.29	87.9
FLEURS (pt)	Lit.	237	0.99	327	319	1.38	9	1.47	84.3
LapsBM (pt)	Lit.	22	0.03	22	22	1.00	4	1.32	86.4
Synth. Terms (pt)	Synth.	519	0.66	787	244	1.52	2	1.00	100.0
Wikidata Synth. (pt)	Synth.	1,207	1.39	1,854	659	1.54	6	1.44	87.9

Table 3: Examples of specialized jargon terms from the English and Portuguese datasets. These terms include proper nouns, technical language, and algorithmically generated words.

Dataset	**Example Jargon Terms**
English Datasets	
CHiME-6 (en)	Metrotown, science beaker
Earnings-22 (en)	npls, gross margin
FLEURS (en)	cyanuric, KwaZulu-Natal
SPGISpeech (en)	quickvue, akamai
Synth. Terms (en)	Xenon Switch, IOT Bridge
Wikidata (Synth.) (en)	Niemarzyn, Athée-sur-Cher
GLOBE (en)	Pontypridd, Salishan
Portuguese Datasets	
CORAA (pt)	agrônomo (ENG: agronomist), magistração (ENG: magistracy)
FLEURS (pt)	Ilhas Cook (ENG: Cook Islands), Kalousis
LapsBM (pt)	Abernéssia (location), Latasa (company name)
Synth. Terms (pt)	PharmaPhenom, LinkLattice
Wikidata (Synth.) (pt)	Chaloem Phra Kiat, Penzing

retrieval-augmented methods proposed by Mathur et al. (2024) or keyword spotting methods proposed by Shamsian et al. (2024); Navon et al. (2024) and Li et al. (2024b).

We evaluated all models across all datasets using the following metrics:

- **Word Error Rate (WER):** The standard metric for ASR performance, calculated as the sum of substitutions (S), deletions (D), and insertions (I) divided by the total number of words in the reference transcript: $(S + D + I)/N$.
- **Biased WER (B-WER):** The WER calculated exclusively on the predefined jargon terms. This metric directly measures the model's ability to recognize specialized vocabulary.
- **Unbiased WER (U-WER):** The WER calculated on all non-jargon words. This metric measures the model's performance on the general vocabulary within the utterance, helping to isolate errors related to jargon.
- **Match Error Rate (MER):** An alternative to WER that is less sensitive to the length of the reference text. It is calculated as the total number of errors divided by the sum of matches, substitutions, and deletions (Morris et al., 2004).
- **Word Information Lost (WIL):** A metric that approximates the information loss from a human perspective, balancing word-level precision and recall to provide a more detailed evaluation of transcription quality (Morris et al., 2004).

5. Results and Analysis

The results of our experiments are presented in Table 4 for the standard transcription condition and Table 5 for the condition with oracle prompting. These tables show the performance of the most pertinent Whisper models. Although experiments were also conducted with the `tiny`, `base`, and `small` versions, only the results from the stronger models are shown for summarization purposes.

An analysis of these results reveals several key insights into the performance of current ASR models on specialized vocabulary.

First, a consistent pattern in the standard condition (Table 4) is the high B-WER compared to the

Table 4: Performance of Whisper models under the **standard condition** (no prompting). The best result for each metric within a dataset group is shown in bold. Lower values indicate better performance (↓).

Language	Dataset	Model	WER ↓	B-WER ↓	U-WER ↓	MER ↓	WIL ↓
English	CHiME-6	medium	0.382	0.811	0.416	0.367	0.330
		large-v3	**0.362**	**0.785**	**0.402**	**0.347**	0.308
		large-v3-turbo	0.377	0.868	0.425	0.354	**0.303**
	Earnings-22	medium	0.292	0.786	0.278	0.264	0.217
		large-v3	0.267	**0.750**	0.248	**0.243**	**0.199**
		large-v3-turbo	**0.266**	0.836	**0.241**	0.251	0.204
	FLEURS	medium	0.116	0.663	0.085	0.111	0.100
		large-v3	**0.107**	**0.587**	0.078	**0.102**	**0.090**
		large-v3-turbo	0.112	0.760	**0.077**	0.108	0.098
	SPGISpeech	medium	0.099	0.745	0.084	0.078	0.066
		large-v3	**0.071**	**0.725**	**0.053**	**0.070**	**0.059**
		large-v3-turbo	0.081	0.903	0.057	0.079	0.068
	Synth. Terms	medium	0.215	0.704	0.140	0.197	0.169
		large-v3	**0.189**	**0.647**	**0.119**	**0.172**	**0.146**
		large-v3-turbo	0.194	0.662	0.122	0.177	0.150
	Wikidata (Synth.)	medium	0.265	0.879	0.186	0.244	0.218
		large-v3	**0.246**	**0.859**	0.167	**0.229**	**0.205**
		large-v3-turbo	0.247	0.869	**0.163**	0.230	0.206
	GLOBE	medium	0.182	0.648	0.142	0.167	0.156
		large-v3	0.178	0.859	0.110	0.171	0.162
		large-v3-turbo	**0.148**	**0.631**	**0.101**	**0.141**	**0.133**
Portuguese	CORAA	medium	0.483	0.741	0.488	0.404	0.362
		large-v3	**0.344**	**0.640**	**0.335**	**0.309**	**0.267**
		large-v3-turbo	0.400	0.835	0.356	0.358	0.313
	FLEURS	medium	0.145	0.634	0.108	0.122	0.110
		large-v3	**0.091**	**0.535**	**0.057**	**0.089**	**0.081**
		large-v3-turbo	0.105	0.649	0.068	0.101	0.092
	LapsBM	medium	0.150	0.693	0.103	0.145	0.131
		large-v3	**0.128**	**0.466**	**0.078**	**0.124**	**0.108**
		large-v3-turbo	0.143	0.852	0.102	0.141	0.127
	Synth. Terms	medium	0.216	0.832	0.132	0.197	0.166
		large-v3	0.197	0.773	0.117	0.179	0.148
		large-v3-turbo	**0.193**	**0.749**	**0.115**	**0.176**	**0.146**
	Wikidata (Synth.)	medium	0.183	0.777	0.091	0.178	0.163
		large-v3	**0.165**	**0.730**	0.071	**0.161**	**0.148**
		large-v3-turbo	0.165	0.750	**0.066**	0.162	0.150

U-WER across all datasets. This difference confirms that the Whisper models, despite their strong general transcription abilities, struggle significantly more with jargon than with common words. This finding validates the core challenge our datasets are designed to underscore: standard ASR systems are vulnerable when faced with OOV or domain-specific language. The datasets also vary in difficulty, with noisy corpora such as CHiME-6 and high OOV-rate corpora such as our synthetic sets posing the greatest challenges.

Second, by comparing the standard results with those in Table 5, it is clear that the prompting approach significantly reduces the B-WER, often by more than 50%, while having a minimal impact on the U-WER. This shows that when the model is given explicit knowledge of the target jargon, it can transcribe it with much higher accuracy without degrading its performance on the surrounding text. This result is expected, as we provide only the correct jargon terms in the prompt. In a real-world application, an imperfect list of potential jargon could introduce noise and negatively affect the U-WER. Nonetheless, this "oracle" condition clearly shows the potential for improvement, highlighting the effectiveness of providing jargon terms to enhance transcription accuracy.

Third, an analysis of the different model sizes yields interesting results. In the standard (non-prompted) setup shown in Table 4, `large-v3`

Table 5: Performance of Whisper models under the **prompted condition**. The models were provided with an oracle prompt containing the correct jargon terms. The best result for each metric within a dataset group is shown in bold. Lower values indicate better performance (↓).

Language	Dataset	Model	WER ↓	B-WER ↓	U-WER ↓	MER ↓	WIL ↓
English	CHiME-6	medium	**0.359**	**0.377**	0.428	**0.338**	**0.279**
		large-v3	0.373	0.746	0.437	0.355	0.306
		large-v3-turbo	0.409	0.522	**0.411**	0.340	0.280
	Earnings-22	medium	0.254	**0.417**	0.271	0.238	0.187
		large-v3	**0.251**	0.709	**0.242**	0.237	0.191
		large-v3-turbo	0.257	0.564	0.258	**0.230**	**0.182**
	FLEURS	medium	**0.083**	**0.250**	0.076	**0.080**	**0.068**
		large-v3	0.095	0.509	**0.073**	0.091	0.079
		large-v3-turbo	0.093	0.419	0.080	0.089	0.078
	SPGISpeech	medium	0.099	**0.248**	0.101	0.068	0.052
		large-v3	0.072	0.692	**0.056**	0.071	0.059
		large-v3-turbo	**0.068**	0.434	0.062	**0.064**	**0.051**
	Synth. Terms	medium	**0.072**	**0.196**	0.062	**0.071**	**0.053**
		large-v3	0.148	0.519	0.100	0.139	0.117
		large-v3-turbo	0.074	0.221	**0.057**	**0.071**	0.055
	Wikidata (Synth.)	medium	**0.141**	**0.386**	**0.133**	**0.132**	**0.105**
		large-v3	0.196	0.671	0.144	0.183	0.158
		large-v3-turbo	0.176	0.573	0.135	0.163	0.142
	GLOBE	medium	**0.101**	**0.208**	0.111	0.097	0.084
		large-v3	0.148	0.711	0.099	0.142	0.134
		large-v3-turbo	**0.101**	0.316	**0.095**	**0.092**	**0.083**
Portuguese	CORAA	medium	0.369	**0.351**	0.415	0.332	0.279
		large-v3	**0.320**	0.548	**0.321**	**0.291**	**0.247**
		large-v3-turbo	0.343	0.484	0.367	0.323	0.272
	FLEURS	medium	**0.081**	**0.206**	0.076	0.079	0.071
		large-v3	0.108	0.441	0.092	0.083	0.075
		large-v3-turbo	0.084	0.314	**0.069**	**0.078**	**0.069**
	LapsBM	medium	**0.069**	**0.045**	0.074	**0.067**	**0.055**
		large-v3	0.109	0.398	0.085	0.106	0.093
		large-v3-turbo	0.082	0.216	**0.070**	0.080	0.065
	Synth. Terms	medium	**0.071**	**0.136**	0.072	**0.069**	**0.046**
		large-v3	0.138	0.548	0.088	0.130	0.104
		large-v3-turbo	0.077	0.214	**0.060**	0.073	0.055
	Wikidata (Synth.)	medium	**0.068**	**0.216**	0.061	**0.066**	**0.054**
		large-v3	0.110	0.507	**0.052**	0.108	0.099
		large-v3-turbo	0.083	0.315	**0.052**	0.081	0.073

consistently performs the best, establishing it as the strongest baseline. However, in the prompted experiments (Table 5), the `medium` model often achieves the best B-WER. A particularly noteworthy finding is that the performance of `large-v3` improves only marginally with prompting. This suggests that the largest model may place less weight on the initial prompt, relying more on its internal representations learned during pre-training.

In summary, the results highlight a large performance gap between current ASR models and systems that can effectively leverage in-domain terminology. The substantial B-WER reductions under oracle prompting quantify a clear objective for future research: to develop systems that can bridge the performance gap between standard transcription and context-aware systems. This work validates the necessity for specialized evaluation datasets, demonstrating that our proposed framework can effectively expose this limitation and, consequently, guide the development of more domain-aware ASR systems.

6. Conclusion

In this paper, we introduced new datasets to evaluate how well ASR models handle specialized vo-

cabulary. We showed that standard models, such as Whisper, struggle to correctly transcribe jargon, proper names, and other terms that are rare in everyday language. This was confirmed by our experiments, which showed a much higher B-WER on jargon terms compared to the error rate on more common words across all datasets.

Our key finding is that providing the model with the correct jargon terms in an initial prompt significantly improves its accuracy on those specific terms. This "oracle" experiment highlights a notable performance gap and points to a clear direction for future work. The goal should be to develop methods that automatically identify and provide relevant domain-specific vocabulary to the ASR system during transcription, closing the gap between standard and prompted results, and minimizing understanding issues, especially in business and formal scenarios. Our datasets provide the resources for the research community to develop and test such solutions, ultimately leading to more reliable ASR systems for specialized, real-world applications.

7. Acknowledgements

This research was partially supported by Brazil's National Council for Scientific and Technological Development - CPNq (grant 313845/2023-9).

8. Ethics Statement and Limitations

8.1. Ethics Statement

The creation and distribution of this dataset follows ethical guidelines for creating academic resources.

- **Data Provenance and Licensing:** All data sources are clearly cited. Terms in our synthetic datasets were either generated by an LLM or taken from Wikidata, which is available under a CC0 1.0 Universal Public Domain license. The audio was synthesized using open-source TTS voices from KokoroTTS, which are available under the Apache 2.0 license, permitting the use and redistribution of the generated audio. For the literature-derived subsets, we do not share original audio or text data. Instead, we provide annotation files and a script that lets researchers rebuild the subsets on their own machines after downloading the original datasets and agreeing to their licenses.

- **Privacy:** No new personally identifiable information was collected. The literature-derived datasets are based on public corpora that were already checked for privacy issues by their creators. Our synthetic data contains no personal information.

- **Potential for Bias:** We acknowledge that the synthesized voices may not represent the full diversity of accents and demographics. Likewise, our jargon selection process may not cover the entire range of specialized vocabulary. These datasets are meant for evaluating technical ASR performance and should not be used to draw sociological conclusions.

8.2. Limitations

We identify limitations of our work, which suggest directions for future research.

- **Acoustic Realism:** While modern TTS systems create high-quality audio, synthetic speech can lack the natural rhythm, background noise, and audio variations found in real-world recordings. Our literature-derived sets help with this, but performance on the synthetic data may not directly translate to performance on real-world audio.

- **Jargon Representativeness:** The LLM-generated terms in the *Synthetic Terms* datasets are phonologically plausible but are not real words. They test how a model handles new phonetic sequences, not its knowledge of a specific domain.

- **Curation Bias:** Our literature-derived subsets were created by finding transcription errors made by the Whisper `large-v3` model. As a result, these datasets are specifically designed to be hard for Whisper-like models.

- **Phonetic Approximation in Portuguese:** The PABP method used for our Portuguese Wikidata set is a practical approach, not a perfect phonetic conversion. It guides the TTS engine toward a reasonable pronunciation but may create small sound artifacts compared to audio generated directly from IPA. To control for this, every audio sample generated via this method was manually reviewed and validated by linguistic experts.

9. Bibliographical References

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant ASR Models for Medical Entities in Accented Speech.

Sourav Banerjee, Ayushi Agarwal, and Promila Ghosh. 2024. High-precision medical speech recognition through synthetic data and semantic correction: United-medasr.

Samuele Cornell, Jordan Darefsky, Zhiyao Duan, and Shinji Watanabe. 2024. Generating data with text-to-speech and large-language models for conversational speech recognition.

Hira Dhamyal, Leda Sari, Vimal Manohar, Nayan Singhal, Chunyang Wu, Jay Mahadeokar, Matt Le, Apoorv Vyas, Bowen Shi, Wei-Ning Hsu, et al. 2024. Using voicebox-based synthetic speech for asr adaptation. In *Proc. SynData4GenAI 2024*, pages 36–40.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. Synthasr: Unlocking synthetic data for speech recognition. *arXiv preprint arXiv:2106.07803*.

hexgrad. 2025. Kokoro-tts. `https://huggingface.co/hexgrad/Kokoro-82M`. Hugging Face Model. DOI: 10.57967/hf/4329.

Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, and Yanzhang He. 2022. Large-scale asr domain adaptation using self- and semi-supervised learning. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6627–6631.

Yash Jogi, Vaibhav Aggarwal, Shabari S Nair, Yash Verma, and Aayush Kubba. 2025. Improving rare-word recognition of whisper in zero-shot settings.

Chin Yuen Kwok, Jia Qi Yip, and Eng Siong Chng. 2025. Improving synthetic data training for contextual biasing models with a keyword-aware cost function. *arXiv preprint arXiv:2509.09197*.

Siqi Li, Danni Liu, and Jan Niehues. 2024a. Optimizing rare word accuracy in direct speech translation with a retrieval-and-demonstration approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12703–12719, Miami, Florida, USA. Association for Computational Linguistics.

Yuang Li, Min Zhang, Chang Su, Yinglu Li, Xiaosong Qiao, Mengxin Ren, Miaomiao Ma, Daimeng Wei, Shimin Tao, and Hao Yang. 2024b. A multitask training approach to enhance whisper with contextual biasing and open-vocabulary keyword spotting.

Zhennan Lin, Kaixun Huang, Wei Ren, Linju Yang, and Lei Xie. 2025. Contextualized automatic speech recognition with dynamic vocabulary prediction and activation.

Puneet Mathur, Zhe Liu, Ke Li, Yingyi Ma, Gil Karen, Zeeshan Ahmed, Dinesh Manocha, and Xuedong Zhang. 2024. DOC-RAG: ASR language model personalization with domain-distributed co-occurrence retrieval augmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5132–5139, Torino, Italia. ELRA and ICCL.

Jan Melechovsky, Ambuj Mehrish, Berrak Sisman, and Dorien Herremans. 2024. Accent conversion in text-to-speech using multi-level vae and adversarial training. In *TENCON 2024-2024 IEEE Region 10 Conference (TENCON)*, pages 473–476. IEEE.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768.

Aviv Navon, Aviv Shamsian, Neta Glazer, Gill Hetz, and Joseph Keshet. 2024. Open-vocabulary keyword-spotting with adaptive instance normalization. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11656–11660.

Joel Jia Wei Ng, Eugene Wang, Xinyan Zhou, Kevin Xiang Zhou, Charlene Xing Le Goh, Gabriel Zheng Ning Sim, Hiang Khoon Tan, Serene Si Ning Goh, and Qin Xiang Ng. 2025. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. *BMC Medical Informatics and Decision Making*, 25(1):236.

Minh-Tien Nguyen, Dat Phuoc Nguyen, Tuan-Hai Luu, Xuan-Quang Nguyen, Tung-Duong Nguyen, and Jeff Yang. 2024. Improving speech recognition with jargon injection. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 490–499.

Alessandro Piscopo, Pavlos Vougiouklis, Lucie-Aimée Kaffee, Christopher Phethean, Jonathon Hare, and Elena Simperl. 2017. What do wikidata and wikipedia have in common? an analysis of their use of external references. In *Proceedings of the 13th International Symposium on Open Collaboration*, OpenSym '17, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th*

International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Aviv Shamsian, Aviv Navon, Neta Glazer, Gill Hetz, and Joseph Keshet. 2024. Keyword-guided adaptation of automatic speech recognition.

Imran Sheikh, Emmanuel Vincent, and Irina Illina. 2022. Transformer versus LSTM language models trained on uncertain ASR hypotheses in limited data scenarios. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 393–399, Marseille, France. European Language Resources Association.

Per Erik Solberg, Pablo Ortiz, Phoebe Parsons, Torbjørn Svendsen, and Giampiero Salvi. 2023. Improving generalization of Norwegian ASR with limited linguistic resources. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 508–517, Tórshavn, Faroe Islands. University of Tartu Library.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2020. Analysis of GlobalPhone and Ethiopian languages speech corpora for multilingual ASR. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4152–4156, Marseille, France. European Language Resources Association.

Minh Tran, Yutong Pang, Debjyoti Paul, Laxmi Pandey, Kevin Jiang, Jinxi Guo, Ke Li, Shun Zhang, Xuedong Zhang, and Xin Lei. 2025. A domain adaptation framework for speech recognition systems with only synthetic data. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alessandro Vietti, Domenico De Cristofaro, and Picciau Sara. 2024. Sensitivity of syllable-based ASR predictions to token frequency and lexical stress. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 983–989, Pisa, Italy. CEUR Workshop Proceedings.

Tyler Vuong, Karel Mundnich, Dhanush Bekal, Veera Elluru, Srikanth Ronanki, and Sravan Bodapati. 2023. AdaBERT-CTC: Leveraging BERT-CTC for text-only domain adaptation in ASR. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 364–371, Singapore. Association for Computational Linguistics.

Jiawei Yu, Yuang Li, Xiaosong Qiao, Huan Zhao, Xiaofeng Zhao, Wei Tang, Min Zhang, Hao Yang, and Jinsong Su. 2024. Hard-synth: Synthesizing diverse hard samples for asr using zero-shot tts and llm.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE.

Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Seyyed Saeed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan. 2023. How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 205–212.

10. Language Resource References

Ardila, R. and Branson, M. and Davis, K. and Henretty, M. and Kohler, M. and Meyer, J. and Morais, R. and Saunders, L. and Tyers, F. M. and Weber, G. 2020. *Common Voice: A Massively-Multilingual Speech Corpus*.

Jon Barker and Shinji Watanabe and Emmanuel Vincent and Jan Trmal. 2018. *The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines*.

Conneau, Alexis and Ma, Min and Khanuja, Simran and Zhang, Yu and Axelrod, Vera and Dalmia, Siddharth and Riesa, Jason and Rivera, Clara and Bapna, Ankur. 2022. *FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech*.

Arnaldo Candido Junior and Edresson Casanova and Anderson Soares and Frederico Santos de Oliveira and Lucas Oliveira and Ricardo Corso Fernandes Junior and Daniel Peixoto Pinto da Silva and Fernando Gorgulho Fayet and Bruno Baldissera Carlotto and Lucas Rafael Stefanel Gris and Sandra Maria Aluísio. 2021. *CORAA: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese*.

O'Neill, Patrick K. and Lavrukhin, Vitaly and Majumdar, Somshubra and Noroozi, Vahid and Zhang, Yuekai and Kuchaiev, Oleksii and Balam, Jagadeesh and Dovzhenko, Yuliya and Freyberg,

Keenan and Shulman, Michael D. and Ginsburg, Boris and Watanabe, Shinji and Kucsko, Georg. 2021. *SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition*.

Panayotov, Vassil and Chen, Guoguo and Povey, Daniel and Khudanpur, Sanjeev. 2015. *Librispeech: An ASR corpus based on public domain audio books*.

Miguel Del Rio and Peter Ha and Quinten McNamara and Corey Miller and Shipra Chandra. 2022. *"Earnings-22: A Practical Benchmark for Accents in the Wild"*.

Vrandečić, Denny and Krötzsch, Markus. 2014. *Wikidata: a free collaborative knowledgebase*. Association for Computing Machinery.

Watanabe, Shinji and Mandel, Michael and Barker, Jon and Vincent, Emmanuel and Arora, Ashish and Chang, Xuankai and Khudanpur, Sanjeev and Manohar, Vimal and Povey, Daniel and Raj, Desh and others. 2020. *CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings*.