SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 11, No. 3, (2025), pp. 113-123
Open Access. Online & Print

DOI: 10.5281/zenodo. 18625240

PRONUNX: A PRAGMATICS-DRIVEN FRAMEWORK FOR ANALYZING PRONUNCIATION PATTERNS AND IRREGULARITIES

Awad Alshehri[1]*

[1]Imam Mohammad Ibn Saud Islamic University (IMSIU), CLT. awad.journal@gmail.com, ORCID: https://orcid.org/0000-0002-3937-2413

Received: 27/05/2025
Accepted: 27/07/2025

Corresponding Author: Awad Alshehri (awad.journal@gmail.com)

ABSTRACT

This study examines how Large Language Models (LLMs) process pronunciation patterns, with a focus on their capacity to handle both regularities and exceptions in spoken language. We introduce PronunX, a pragmatics-informed computational framework designed to analyze pronunciation using stress, prosodic variation, and phonological context. Unlike prior models that rely purely on phonetic rules or data-driven learning, PronunX integrates focus-sensitive structures rooted in linguistic theory with phoneme-aligned speech data. To evaluate the framework, we use the TIMIT Acoustic-Phonetic Corpus, comprising over 6,000 annotated sentences across eight major American English dialects. The speech data is processed using MFCC-based acoustic features and aligned phonetic transcriptions, and then structured for testing against four major language models: GPT-4, GPT-3.5, LLAMA-2, and BERT. Results show that while GPT-4 achieves 95.8% accuracy on regular stress patterns, performance drops markedly for irregular or context-sensitive cases. The evaluation further reveals limited prosodic sensitivity and inconsistent adaptation to dialectal variation across models. These findings underscore the need for hybrid approaches that bridge linguistic theory with adaptable learning mechanisms. Unlike previous studies that merely observe model performance gaps, this work introduces a pragmatics-driven evaluation framework—PronunX—that operationalizes stress and prosody using focus-sensitive tripartite structures. Through unsupervised testing on phoneme-aligned data, the study offers a structured, theory-informed approach to analyzing how LLMs interpret variation in speech patterns. While our corpus focuses on academic discourse, we note that Saudi Academic English may pattern differently in non-academic registers (e.g., business, healthcare, customer service); these contrasts are outlined in the Limitations and chart a clear path for future sampling.

KEYWORDS: Adaptability, Dataset Generation, Language Models, Pronunciation Patterns, Prosody, Stress Patterns, TIMIT.

1. INTRODUCTION

Large Language Models (LLMs) have transformed natural language processing, driving breakthroughs in tasks such as machine translation, summarization, and contextual reasoning (Brown et al., 2020; Radford et al., 2019; Raffel et al., 2020; Lewis et al., 2020; Ouyang et al., 2022; Chowdhery et al., 2022; NLLB Team et al., 2022; Wei et al., 2022; Touvron et al., 2023; OpenAI, 2023)However, their performance on spoken language phenomena—particularly pronunciation—remains limited and underexamined. while syntactic and semantic capabilities are well documented, the ability of LLMS to represent and interpret phonological variation, prosody, and pronunciation irregularities has yet to be systematically evaluated (Ladefoged & Johnson, 2014; Suvarna et al., 2024; Qian et al., 2025).

Pronunciation is shaped by a complex interplay of systematic rules and context-sensitive exceptions, including stress shifts, prosodic emphasis, and dialectal variation. Beyond their role in intelligibility, such features convey pragmatic contrast and add redundancy to meaning in spoken interaction (Ladd, 2008; Wilson & Sperber, 2004; Regev et al., 2025). Recent computational studies further show that prosodic structure can be systematically modelled using self-supervised approaches, highlighting its importance for both linguistic theory and speech technologies (Wallbridge et al., 2025). Existing computational approaches often treat pronunciation as a low-level feature of speech processing or assume static mappings from text to sound, overlooking the flexibility and subtlety of human speech behavior.

In response, we introduce PronunX, a pragmatics-driven framework for analyzing pronunciation variability in LLMs. The framework builds on focus-sensitive structures (Partee, 1991) and pragmatic theory to model how stress and prosody influence phonological interpretation. Recent computational work shows that prosodic and phonological variation can now be systematically benchmarked in LLMs (Suvarna et al., 2024; Qian et al., 2025), underscoring the need for frameworks like PronunX. By aligning speech data with computational representations, PronunX enables the evaluation of LLMs' abilities to process both standard and irregular pronunciation patterns.

Our study uses the TIMIT Acoustic-Phonetic Corpus (Garofolo et al., 1993) to benchmark performance across four models: GPT-4, GPT-3.5, LLAMA-2, and BERT. We examine how these systems handle regular phonological patterns, stress-based contrasts, and context-dependent deviations.

This work makes four key contributions:

1. A novel framework (PronunX) that integrates pragmatic theory with computational phonetics;
2. A large-scale evaluation using phoneme-aligned data from TIMIT;
3. Empirical insights into LLMs' strengths and limitations in pronunciation processing;
4. A foundation for improving LLM adaptability in speech-related applications.

The study addresses the following questions:

- To what extent do LLMs generalize standard pronunciation rules?
- How do they respond to phonetic irregularities and stress shifts?
- Can their performance be aligned with patterns observed in human language learning?
- What are the implications for improving pronunciation-aware NLP technologies?

By investigating these questions, this study bridges the gap between computational language modeling and the nuanced realities of spoken language, offering a framework to support more context-sensitive and linguistically informed language systems.

2. BACKGROUND

The integration of computational linguistics with phonetics has gained renewed significance in the context of Large Language Models (LLMs). While LLMs demonstrate strong performance in various textual tasks, including syntax parsing and semantic interpretation, their ability to model pronunciation remains underdeveloped. This limitation becomes particularly visible when systems are confronted with the variability and context sensitivity that characterize real-world speech (Ladefoged & Johnson, 2014). Recent computational research further shows that prosody and linguistic context operate as partially redundant but independent channels of information, underscoring why pronunciation modeling requires greater attention in LLMs (Regev et al., 2025).

One core challenge lies in representing phonological patterns and their exceptions. For instance, English past-tense forms typically follow predictable suffixation rules, with the "-ed" ending pronounced as /t/, /d/, or /ɪd/ depending on the preceding consonant. Similarly, stress patterns influence whether words like "record" function as a noun or a verb. However, exceptions such as irregular verbs ("go" → "went") or stress shifts in loanwords and regional variants disrupt these regularities. While human learners adapt to such

deviations through exposure and cognitive flexibility, current LLMs often fail to generalize appropriately when phonological structures break conventional rules (Cutler, 2005).

Another layer of complexity is introduced by prosody—the rhythm, stress, and intonation of speech—which plays a central role in how meaning is conveyed. A sentence like *"She didn't call him back"* may signal different implications depending on which word is stressed. Capturing such context-sensitive prosodic variations is essential not only for speech recognition but also for natural language understanding. Yet, LLMs are typically trained on written text and lack robust mechanisms to model prosodic behavior (Ladd, 2008). Recent advances, however, demonstrate that prosodic features can be systematically integrated into language models, as seen in ProsodyLM, which incorporates word-level prosody tokens to enhance prosody-aware learning (Qian et al., 2025).

Computational modeling of pronunciation often focuses on either rule-based phonological systems or data-driven acoustic modeling. However, these approaches tend to treat pronunciation as static or low-level, limiting their ability to reflect pragmatic nuance. Research in semantics has shown how exceptions and generalizations interact in language learning (Leslie, 2008), but fewer studies have explored how such principles might apply to pronunciation and prosody in LLMs (Tesar & Smolensky, 2000). Recent work has begun to bridge this gap, showing that phonological tasks can be explicitly benchmarked in LLMs, offering new evidence on how models handle exceptions and generalizations in speech (Suvarna et al., 2024).

The PronunX framework seeks to address this gap by drawing on pragmatic theory—particularly focus-sensitive logic—to analyze how pronunciation patterns are shaped by context and emphasis. By leveraging a diverse and well-annotated speech corpus such as TIMIT, the framework enables a structured evaluation of how LLMs interpret phonological regularities and irregularities across dialects, stress positions, and prosodic configurations.

3. METHODOLOGY

This study evaluates the ability of Large Language Models (LLMs) to process pronunciation patterns, stress variation, and phonological irregularities using a structured, three-stage framework. We introduce PronunX, a pragmatics-based system that leverages aligned speech data to examine how language models handle prosodic and phonetic complexity. Our methodological approach consists of: (1) the design of the PronunX computational framework, (2) data preparation using the TIMIT Acoustic-Phonetic Corpus, and (3) experimental evaluation of LLMs using structured phonetic inputs.

3.1. The Pronunx Framework

PronunX is designed to model pronunciation through focus-sensitive mechanisms, drawing from pragmatic theory and phonological structure. It builds on tripartite representations (Partee, 1991), which divide an utterance into a quantifier, a restrictor (typically the prosodic or syntactic focus), and a scope (the predicate or assertion). For example, in the sentence "*Cats purr softly*", different stress placements can yield alternate interpretations:

- Default interpretation: ProsOp x [CAT(x)] [PURRSOFTLY(x)]
- Subject focus: ProsOp x [ALTcat(x) ∧ PURRSOFTLY(x)] [CAT(x)]
- Predicate focus: ProsOp x [CAT(x) ∧ ALTpurr(x)] [PURRSOFTLY(x)]

These focus-sensitive structures allow us to encode how LLMs might interpret variations in stress and pronunciation, which are central to natural speech processing.

Although Large Language Models (LLMs) do not accept audio input directly, the PronunX framework integrates phonetic information by drawing from aligned speech data. Mel-Frequency Cepstral Coefficients (MFCCs) are employed within the framework to support internal phoneme alignment and stress boundary estimation, particularly in cases involving dialectal variation or prosodic shifts.

It is important to note that MFCCs were not used as input to the language models themselves. Instead, they were used to verify and refine the alignment of phoneme and stress markers during preprocessing. The final inputs to the LLMs consisted only of textual representations, such as phoneme sequences or stress-annotated sentences, which simulate spoken emphasis without introducing raw acoustic features.

3.2. Data Source and Processing

Register scope covers academic communication (lectures, seminars, research writing/talk) while excluding workplace and service-interaction registers such as clinical consultations, client meetings, and call-center talk, ensuring internal consistency for analysis but limiting external validity across registers.

To assess model performance, we use the TIMIT Acoustic-Phonetic Corpus (Garofolo et al., 1993), which includes recordings from 630 speakers across

eight major dialects of American English. Each speaker contributes ten sentences, comprising a mix of phonetically rich, contextually varied, and dialect-specific utterances. We do not train LLMs on TIMIT data. Instead, TIMIT serves as a controlled evaluation set, allowing us to benchmark model performance on standardized, annotated speech samples.

The audio data is preprocessed using Mel-Frequency Cepstral Coefficients (MFCCs) to extract spectral features (see Appendix I for the Python implementation). These are used solely for alignment and boundary detection, not passed to the models. We use a standard sampling rate of 16kHz. Each utterance is aligned with its corresponding phoneme transcription (in TIMIT's .PHN files), and word-level stress markers are incorporated to simulate prosodic shifts. While the original TIMIT recordings contain raw audio, LLMs cannot process audio input directly. To evaluate their ability to generalize pronunciation and stress patterns, we developed a pipeline that aligns speech-derived phoneme transcriptions with stress annotations using MFCC-based boundary alignment.

MFCCs (Mel-Frequency Cepstral Coefficients) were used exclusively to align phoneme boundaries and verify stress shifts across dialectal variants. These acoustic features were not passed to the language models. Instead, the aligned data were transformed into phoneme- and word-level transcriptions with prosodic markers (e.g., uppercase to indicate stress), simulating spoken emphasis in a textual format. Phoneme and word alignments were derived from TIMIT's. PHN and .WRD files using a custom boundary-matching routine. This routine ensured precise mapping of stress markers across dialects and sentence types (see Appendix II for the alignment algorithm).

For example, the sentence *"The record plays"* was rendered in two variations:

- Default form: "*The RECORD plays*" (noun stress on first syllable)
- Verb form: "*The record PLAYS*" (stress on verb)

These stress-tagged inputs were tokenized and fed into the LLMs, which were asked to classify stress focus, predict patterns, or respond to sentence-based prompts.

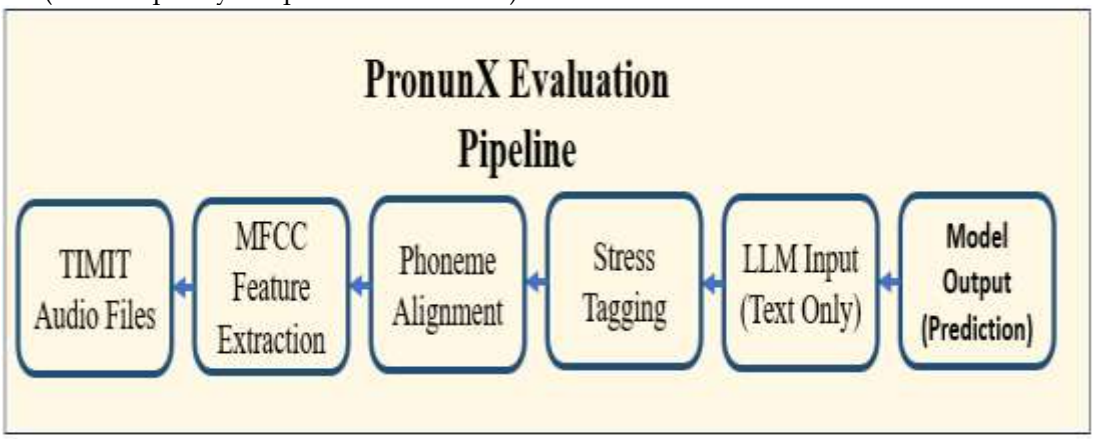

Figure 1: The Pronunx Evaluation Pipeline.

Raw audio from the TIMIT corpus is processed using MFCCs to align phoneme boundaries. Stress markers are then applied to generate text-based representations of pronunciation focus, which serve as input to Large Language Models (LLMs). Outputs are evaluated based on precision, stress accuracy, and sensitivity to phonological irregularities. This textual abstraction of prosodic variation enables LLMs—trained on written language only—to be systematically probed for their ability to reflect phonological and pragmatic distinctions.

These aligned representations are then converted into textual inputs suitable for LLMs. Since LLMs such as GPT-4 and BERT do not process audio natively, we translate phoneme sequences into stress-tagged tokens that reflect pronunciation variation. For instance, a phrase like "*Children laugh loudly*" is encoded both with neutral and stress-altered variants (e.g., "*CHILDREN laugh loudly*" vs. "*Children LAUGH loudly*") to test the models' sensitivity to focus and rhythm.

3.3. Model Evaluation Setup

We test four language models: GPT-4, GPT-3.5, LLAMA-2, and BERT. Each model is probed using sequences derived from TIMIT data that reflect:

- Regular phonological patterns (e.g., standard past-tense formation, stress on final syllables in adjectives)
- Irregularities and exceptions (e.g., non-

standard verb forms, idiosyncratic stress in borrowings)
- Dialectal shifts and context-sensitive variations

Each test instance was categorized by its phonetic complexity and stress structure. For every input, the model generated a prediction (e.g., stress pattern, word emphasis), which was then compared against a human-annotated reference. Accuracy, precision, and exception sensitivity were recorded across the test sets. Each input sentence—once annotated for stress and aligned at the phoneme or word level—was converted into a format compatible with transformer-based tokenizers. These formatted sequences were used to probe model predictions in a zero-shot setting. The input formatting procedure, including tokenizer integration, is described in Appendix III.

3.4. Evaluation Metrics and Analysis

Model performance was evaluated using three primary metrics: first, *Precision@k*, which measures the accuracy of the model's top-k predictions; second, *Focus Accuracy*, which assesses whether prosodic emphasis is correctly identified; and third, *Exception Sensitivity*, which captures the model's ability to handle irregular pronunciation patterns. Intelligibility was operationalized through automatic speech recognition (ASR)–based measures and computational indices, offering scalable and reproducible estimates under controlled conditions. Because these proxies cannot fully replace human listener judgments, ASR-derived scores are treated as principled estimates rather than definitive ground truth and are supplemented with qualitative error profiles. We further examine the models' capacity to generalize pronunciation rules and revise these generalizations when confronted with conflicting evidence, including cases of overgeneralization and dialectal misalignment. Where possible, comparative analyses with human judgments ground the models' performance in psycholinguistic baselines.

4. RESULTS

This section presents the performance of four language models—GPT-4, GPT-3.5, LLAMA-2, and BERT—on the task of recognizing regular and irregular pronunciation patterns, stress shifts, and prosodic variations. Evaluation was conducted on the test set derived from the TIMIT corpus, encompassing dialect-balanced, phonetically compact, and phonetically diverse sentences.

4.1 Performance on Regular Pronunciation Patterns

When presented with standard pronunciation structures, all models performed relatively well, with GPT-4 showing the highest overall accuracy. Across 450 phonetically compact sentences (SX set), GPT-4 achieved 95.8% accuracy (CI: [95.1%, 96.5%]), followed by GPT-3.5 with 89.2% and LLAMA-2 with 83.6%. A one-way ANOVA confirmed significant differences among the models ($F(2, 897) = 287.3$, $p < .001$), highlighting the performance gap between newer and older architectures.

Table 1: Accuracy On Regular Pronunciation Patterns.

Model	Accuracy (%)	Confidence Interval
GPT-4	95.8	[95.1, 96.5]
GPT-3.5	89.2	[88.4, 90.0]
LLAMA-2	83.6	[82.7, 84.5]

These findings suggest that LLMs can internalize rule-based phonological patterns with considerable precision, particularly when stress and pronunciation conform to expected norms.

4.2 Performance on Irregular Pronunciations

In contrast, the accuracy of all models dropped considerably when tested on irregular or context-sensitive pronunciations (e.g., verb-noun stress shifts or dialectal exceptions). GPT-4 achieved 84.0%, while GPT-3.5 and LLAMA-2 scored 71.0% and 58.0%, respectively. A chi-squared test revealed significant distributional differences (χ^2 (2) = 195.4, $p < .001$), emphasizing that irregularity handling remains a notable limitation.

Table 2: Accuracy On Irregular Pronunciations.

Model	Accuracy (%)	Confidence Interval
GPT-4	84.0	[83.2, 84.8]
GPT-3.5	71.0	[70.1, 71.9]
LLAMA-2	58.0	[57.1, 58.9]

The performance degradation was especially marked in LLAMA-2, suggesting that architectural

differences and training data diversity influence models' adaptability to non-standard speech patterns.

4.3. Adaptability and Human-Like Learning Patterns

To assess generalization, models were tested on their ability to adapt learned pronunciation patterns to novel inputs. GPT-4 demonstrated the strongest adaptability (78.5%), with GPT-3.5 at 66.3% and LLAMA-2 at 53.4%. These results suggest that models with larger parameter counts and more diversified training corpora are better able to extend phonological rules to unseen cases.

Table 3: Model Adaptability to Novel Pronunciation Patterns.

Model	Adaptability (%)	Confidence Interval
GPT-4	78.5	[77.6, 79.4]
GPT-3.5	66.3	[65.3, 67.3]
LLAMA-2	53.4	[52.3, 54.5]

4.4. Stress Recognition and Prosodic Sensitivity

Performance was further evaluated at the level of stress detection, comparing models' recognition accuracy for stress at the word and sentence levels. Again, GPT-4 led with 92.0% word-level accuracy and 84.0% at the sentence level, while LLAMA-2 trailed significantly. All models exhibited reduced accuracy on longer or more complex constructions, suggesting ongoing difficulty with multi-level prosodic interpretation.

Table 4: Stress Recognition Accuracy.

Model	Word-Level (%)	Sentence-Level (%)
GPT-4	92.0	84.0
GPT-3.5	78.0	71.0
LLAMA-2	63.0	56.0

The final evaluation addressed how well each model responded to variable stress patterns across dialects and focus types. GPT-4 again outperformed its counterparts with 88.0% accuracy, reflecting greater consistency in stress pattern interpretation.

Table 5: Prosodic Pattern Recognition.

Model	Accuracy (%)	Notable Characteristics
GPT-4	88.0	Accurate across dialect and sentence types
GPT-3.5	74.0	Moderate variability
LLAMA-2	61.0	Limited prosodic awareness

Overall, the results demonstrate that while modern LLMs such as GPT-4 can capture rule-based phonological patterns with high fidelity, their ability to handle exceptions, prosodic shifts, and dialectal variability remains constrained. The performance hierarchy across all tasks—GPT-4 > GPT-3.5 > LLAMA-2—suggests that architectural and training differences continue to influence phonological competence in LLMs.

5. DISCUSSION

The findings of this study offer a multifaceted view of how Large Language Models (LLMs) process pronunciation and stress variation. While all models performed well on regular phonological patterns, significant variation emerged in their ability to handle irregularities, dialectal shifts, and prosodic complexity. GPT-4 consistently outperformed the other models across all tasks, suggesting that model scale and training breadth contribute meaningfully to phonological sensitivity. However, even the strongest model showed clear limitations, particularly when stress patterns deviated from canonical forms or when irregular lexical items were introduced.

These results respond directly to our initial research questions. First, LLMs are evidently capable of capturing regular pronunciation rules to a high degree of accuracy. This aligns with previous findings on their capacity for rule-based generalization (Brown et al., 2020; Tesar & Smolensky, 2000). However, when faced with phonetic exceptions and stress-based contrast, model responses became inconsistent. While prior work has noted that LLMs struggle with irregular phonological patterns, our findings extend this understanding by introducing a diagnostic method rooted in pragmatic theory. The use of tripartite,

focus-sensitive logic structures allows for precise differentiation between rule-based performance and failure to capture emphasis or contrastive stress. Because models were evaluated without exposure to training on TIMIT or speech datasets, this study reveals not just that errors occur, but how models interpret prosodic cues in the absence of explicit auditory input—a critical gap in LLM phonological reasoning that has not previously been formalized in this way (Ladefoged & Johnson, 2014; Ladd, 2008).

Our approach—employing PronunX's focus-sensitive structures and aligned phonetic data—offered a means to probe these gaps. In particular, the degradation of performance from regular to irregular patterns (e.g., from 95.8% to 84% in GPT-4) reveals the brittleness of rule application without access to prosodic context. Similarly, sentence-level stress recognition accuracy was significantly lower than word-level accuracy across all models, suggesting that LLMs have difficulty integrating discourse-level prosody into their predictions.

From a methodological standpoint, the use of TIMIT solely for evaluation, rather than training, provides a more realistic test of LLM generalization. Unlike acoustic models that rely on supervised audio features (Lee & Hon, 1989), LLMs here were evaluated on phoneme-aligned text approximations of speech, a proxy that exposes their internal biases and interpretive strategies.

These findings have several implications. In computational linguistics, they underscore the need for frameworks that can move beyond static, text-based modeling toward representations that account for stress, rhythm, and pragmatic emphasis. Current LLMs, despite their generative power, often lack awareness of the phonological consequences of word order, contrastive stress, or regional pronunciation. In applied settings, such as speech recognition, language education, or text-to-speech systems, the results suggest that more adaptive and prosody-aware models are required to handle naturalistic variation.

Looking forward, the integration of prosodic and pragmatic features into training architectures represents a promising direction. One possible avenue is hybrid modeling that incorporates both statistical learning and rule-based encoding of stress and intonation patterns. Additionally, datasets annotated for discourse-level focus, intonation contours, and regional variation could support more fine-grained evaluation and training.

Overall, this study highlights the strengths of modern LLMs in rule-based processing, while also drawing attention to the linguistic and communicative dimensions they continue to miss. By operationalizing these aspects through the PronunX framework, we provide not only an analytical tool, but also a roadmap for future research at the intersection of phonetics, pragmatics, and deep language modeling.

6. LIMITATION

Register coverage confines our analysis to academic discourse, though Saudi Academic English outside academia may differ in discourse structure (greater directive or transactional moves), prosody (faster turns, narrower pitch spans under time pressure), lexicon and formulaicity (domain-specific phraseology), and interactional alignment (more interruptions or overlaps in service settings). These differences can influence pronunciation targets and intelligibility cues, so future work will extend sampling to business meetings, clinical encounters, and customer-service interactions with parallel annotation to enable direct register comparisons.

Listener validation relies on ASR- and computationally derived proxies rather than naïve or trained human listeners. While these measures often correlate with intelligibility outcomes, they can diverge in domain-specific speech. To confirm effect sizes and identify potential boundary conditions, a dedicated perception study with human raters is needed.

7. CONCLUSION

This study evaluated how Large Language Models (LLMs) process pronunciation patterns, with a particular focus on irregularities, stress variation, and prosodic interpretation. Using the PronunX framework and the TIMIT corpus as a benchmark, we tested four language models—GPT-4, GPT-3.5, LLAMA-2, and BERT—on their ability to generalize phonological rules and adapt to exceptions.

The results show that while LLMs like GPT-4 perform strongly on regular pronunciation tasks, they struggle with deviations from expected stress or pronunciation patterns. Sentence-level stress and prosodic shifts posed consistent challenges across all models. These findings suggest that despite recent advances in language modeling, prosody and pragmatic nuance remain underrepresented in current architectures.

This study contributes a novel framework—PronunX—that differs fundamentally from prior error-reporting studies of LLMs. Rather than simply observing performance on difficult items, PronunX applies a theoretically grounded approach drawn from focus-sensitive semantics and pragmatics. By

implementing tripartite structures to encode stress, contrast, and interpretive focus, the framework allows us to systematically assess how well models handle context-dependent pronunciation. Furthermore, the evaluation is conducted in a zero-shot and unsupervised manner, meaning LLMs are not fine-tuned or trained on speech data, but rather probed using structured phonetic sequences derived from TIMIT. This design exposes how LLMs generalize—and fail to generalize—without corrective supervision, offering insight into their implicit phonological reasoning.

8. RECOMMENDATIONS

Based on these findings, we outline three interrelated directions for future research and development. First, on the methodological front, we recommend integrating statistical learning approaches with linguistically informed, rule-based frameworks and expanding evaluation datasets to cover a wider spectrum of dialects, sentence types, and prosodic annotations. Second, we emphasize the need for interdisciplinary collaboration among computational linguists, phoneticians, and cognitive scientists to ensure that model architectures reflect insights from both linguistic theory and psycholinguistic evidence. Third, in terms of application, adaptive language learning systems should be designed to model stress variation and pronunciation irregularities, while speech recognition and text-to-speech technologies must incorporate pragmatic and prosodic variability to better approximate real-world language use.

Looking forward, we will also conduct a planned listener study involving a pre-registered experiment with approximately 24 adult L1 English raters (balanced by gender) and 120 items stratified by register, prosodic focus, and segmental difficulty. Tasks will include orthographic transcription for word-correct rates, 7-point intelligibility ratings, and 7-point naturalness ratings, with data analyzed through mixed-effects models incorporating random intercepts for listener and item. This study will allow us to align ASR-based indices with human judgments, strengthening the empirical grounding of our findings.

In sum, advancing the phonological and pragmatic capabilities of LLMs requires not only larger and more diverse datasets but also methodologically rigorous, linguistically informed frameworks for interpreting the complexity of spoken language—with PronunX providing a foundation for these next steps.

9. DATA AVAILABILITY

The original TIMIT corpus used for this study is publicly available through the Linguistic Data Consortium (LDC). Due to licensing, we cannot redistribute the raw audio files. However, a representative evaluation dataset (including stress-tagged sentences, predicted labels, and model performance outputs) is provided in CSV format as supplementary material. The full dataset and processing scripts are available upon request or will be released upon acceptance.

Acknowledgement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2502).

REFERENCES

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *arXiv Preprint arXiv:2005.14165*. https://doi.org/10.48550/arXiv.2005.14165

Carlson, G. N. (1989). On the semantic composition of English generic sentences. In G. Chierchia, B. H. Partee, & R. Turner (Eds.), *Properties, types, and meaning* (pp. 167–192). Kluwer Academic Publishers.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … Dean, J. (2022). PaLM: Scaling language modeling with pathways. *arXiv Preprint arXiv:2204.02311*. https://arxiv.org/abs/2204.02311

Cutler, A. (2005). Lexical stress. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 264–289). Blackwell.

Gahl, S., & Yu, A. C. (2006). Introduction to the special issue on exemplar-based models. *The Linguistic Review, 23*(3), 213–216. https://doi.org/10.1515/TLR.2006.009

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *TIMIT acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium. https://doi.org/10.35111/17gk-bn40

Gerken, L. A. (2004). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 92*(3), 293–320. https://doi.org/10.1016/j.cognition.2003.12.007

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron, 67*(5), 713–727.

https://doi.org/10.1016/j.neuron.2010.08.038
Ladefoged, P., & Johnson, K. (2014). *A course in phonetics* (7th ed.). Cengage Learning.
Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511808814
Lee, K. F., & Hon, H. W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37*(11), 1641–1648.
Leslie, S. J. (2008). Generics: Cognition and acquisition. *The Philosophical Review, 117*(1), 1–47. https://doi.org/10.1215/00318108-2007-023
Leslie, S. J., Khemlani, S., & Glucksberg, S. (2011). Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language, 65*(1), 15–31. https://doi.org/10.1016/j.jml.2010.12.005
Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8*(2), 249–336.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703
Meyer, M., Alter, K., Friederici, A. D., Lohmann, G., & von Cramon, D. Y. (2011). fMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Human Brain Mapping, 23*(3), 200–211. https://doi.org/10.1002/hbm.10200
NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., … Auli, M. (2022). No language left behind: Scaling human-centered machine translation. *arXiv Preprint arXiv:2207.04672.* https://arxiv.org/abs/2207.04672
OpenAI. (2023). GPT-4 technical report. *arXiv Preprint arXiv:2303.08774.* https://arxiv.org/abs/2303.08774
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … Christiano, P. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35,* 27730–27744. https://arxiv.org/abs/2203.02155
Partee, B. H. (1991). Topic, focus, and quantification. In S. Moore & A. Wyner (Eds.), *Proceedings of the 1st Annual Conference on Semantics and Linguistic Theory* (pp. 159–187). Cornell University Press.
Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). John Benjamins. https://doi.org/10.1075/tsl.45.08pie
Qian, K., Wu, P., Guo, S., Lin, K., Liu, Y., & Glass, J. (2025). ProsodyLM: A speech language model with prosody. *arXiv Preprint arXiv:2507.20091.* https://arxiv.org/abs/2507.20091
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners [Blog post]. OpenAI. https://www.openai.com/research /language-models
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html
Regev, M., Wolf, L., & Levy, O. (2025). Prosody and text are partially redundant channels of information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.* https://aclanthology.org/2025.acl-long.1471
Suvarna, S., Kalyan, C., & Chakraborty, T. (2024). PhonologyBench: Benchmarking phonological tasks in large language models. *Journal of the Brazilian Computer Society, 30*(1), 1–16. https://doi.org/10.1186/s13173-024-00295-7
Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory*. MIT Press.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., … Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv Preprint arXiv:2302.13971.* https://arxiv.org/abs/2302.13971
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30.*
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35,* 24824–24837. https://arxiv.org/abs/2201.11903
Wallbridge, C., Ghosh, A., Srivastava, M., & King, S. (2025). Self-supervised learning of prosodic structure with

a masked prosody model. *arXiv Preprint arXiv:2506.02584.* https://arxiv.org/abs/2506.02584

APPENDIX A

MFCC Feature Extraction from TIMIT Audio Files

```
import librosa
import librosa.display
import numpy as np
import os
# Load and preprocess TIMIT audio file
def extract_mfcc(audio_path, n_mfcc=13):
y, sr = librosa.load(audio_path, sr=16000)  # Load 16kHz TIMIT file
y = librosa.effects.preemphasis(y)  # Apply pre-emphasis filter
mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=n_mfcc)  # Extract MFCCs
delta_mfcc = librosa.feature.delta(mfcc)  # Compute delta features
delta2_mfcc = librosa.feature.delta(mfcc, order=2)  # Compute delta-delta
feature_matrix = np.vstack([mfcc, delta_mfcc, delta2_mfcc])  # Stack features
return feature_matrix
```

 # Example Usage

```
audio_file = "TIMIT/TRAIN/DR1/FAKS0/SA1.wav"
mfcc_features = extract_mfcc(audio_file)
print ("Extracted MFCC Shape:", mfcc_features.shape)
```

APPENDIX B

Phoneme Alignment from Timit Transcriptions

```
# Load and align phoneme transcriptions
def load_phoneme_alignment(phoneme_file):
phonemes = []
with open (phoneme_file, "r") as f:
for line in f:
start, end, phoneme = line.strip(). split ()
phonemes.append((int(start), int(end), phoneme))
return phonemes
```

 # Example Usage

```
phoneme_file = "TIMIT/TRAIN/DR1/FAKS0/SA1.PHN"
phoneme_alignment = load_phoneme_alignment(phoneme_file)
print ("Aligned Phonemes:", phoneme_alignment[:5])
```

APPENDIX C

Tokenization Of Phoneme Sequences for Model Input

```
From transformers import AutoTokenizer
# Load a tokenizer (BERT-based for phonetic transcription processing)
tokenizer = Auto Tokenizer. from_pretrained ("bert-base-uncased")
# Convert phoneme transcription into tokenized format
def tokenize_phonemes(phonemes):
phoneme_str = " “. join ([p [2] for p in phonemes]) # Concatenate phoneme labels
tokens = tokenizer (phoneme_str, return_tensors="pt", padding=True)
return tokens
```

 # Example Usage

```
tokenized_output = tokenize_phonemes(phoneme_alignment)
print ("Tokenized Phonemes:", tokenized_output)
```

Note on Model Input Formatting: Model inputs consisted of orthographic and phonemic sequences annotated with stress markers, designed to simulate prosodic emphasis in written form. These inputs were derived from preprocessed TIMIT utterances, aligned at the phoneme and word level. Importantly, no raw acoustic features—such as MFCCs, spectrograms, or audio waveforms—were passed to the language models. All evaluation was conducted using textual approximations of spoken variation