Basilectal-Inspired Health Questions Expose Robustness Gaps in Small Medical QA Models

Saiankit Shankar[1] **BHVSP Subrahmanyam**[2] **Neeraj Choudhary**[1]

[1]Department of Computer Science and Engineering, Mahindra University, India
[2]MU-VT Interdisciplinary Advanced Research Centre for Transformative Technologies, Mahindra University, India

Abstract

Small local language models are attractive for medical question answering in settings where connectivity, cost, and privacy limit cloud-based use. Yet users in these settings may express health questions in forms that move away from benchmark-standard English, including compressed, phonetic, and basilectal-inspired phrasing. We present a controlled robustness study using 102 TREC LiveQA Medical questions, paired L0–L3 synthetic basilectal-inspired variants, six small instruction-tuned models, and 7,416 free-form generations. Across the baseline setting, answer quality is stable under mild variation but degrades sharply in the L3 basilectal-inspired condition: BERTScore-style F1 falls from 0.707 at L0 to 0.614 at L3, ROUGE-L from 0.136 to 0.095, and Medical Concept Overlap recall from 0.129 to 0.046. The dominant failure mode is fluent incompleteness: models often continue producing plausible medical text while preserving fewer medically relevant concepts from trusted reference answers. We also examine prompt intervention, quantization, feature ablations, behavioral proxies, and claim-level support. The study is intentionally bounded: it is a controlled robustness benchmark, not a clinical safety certification, not a naturally collected dialect corpus, and not a claim about any specific speech community. Code is available at https://github.com/ankitsblade/Basilectal-Robustness-in-SLMs

1 Introduction

The use of smaller language models as local health aides is growing more realistic. It is becoming possible to execute tiny instruction-tuned models on smartphones and inexpensive edge computing devices, and that has significance in areas where connectivity is unstable, cellular data is costly, and sensitive healthcare inquiries cannot affordably be processed by cloud-based systems. That makes local models compelling for routine health guidance, but it means that risks become more immediate for people lacking access to medical care or reliable information networks.(Garg et al., 2025; Magnini et al., 2025).

One key but understudied risk is language mismatch. Potential users of English-based health tools rarely write in standard American or British English. They may pose questions using local varieties of English, nonstandard spellings, lack of functional words, contact-language syntax, or condensed telephonic speech. The field of World Englishes research demonstrates that linguistic diversity is structured and socially informed, rather than chaotic mistakes (Kachru, 1990; Mesthrie and Bhatt, 2008). Meanwhile, most assessments of medical question answering systems rely on standard benchmark inputs, despite evidence that language technology applications are unevenly deployed among languages and speaker groups (Blasi et al., 2022), and recent benchmark studies of dialects suggest that such evaluations overlook these inequalities (Faisal et al., 2024; Lin et al., 2025).

Such an ambiguity is of importance to the reliability of medical question answering systems. A user inquiring about fever, pregnancy, dosages, side effects, or preventive actions can speak nonstandard English without failing to convey health-related intentions. When the locally trained model fails to understand the query, misses out medical terminology, and generates a non-specific response by pretending fluency, the mistake transcends language issues, resulting in an access issue. Currently, there is no study concerning the performance of small deployable models when handling health questions that approach more phonetic, compressed, and basilectalized queries, particularly in free-form generation tasks similar to conversational chatbots.

Herein, *basilectal-inspired* has been defined in

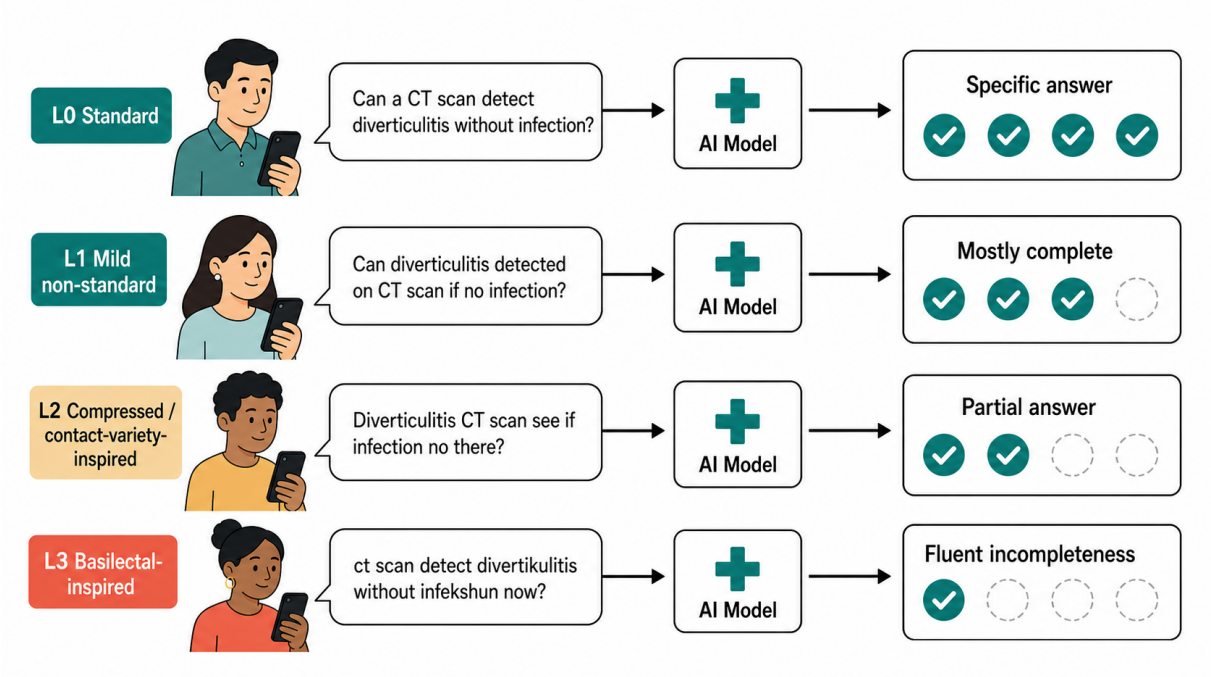

Figure 1: Basilectal-inspired L0–L3 input continuum. The medical intent is fixed, while the input moves farther from benchmark-standard English and model outputs become progressively less complete, ending in *fluent incompleteness*.

the context of testing robustness near intelligibility boundaries. Our study does not aim to model a particular basilect, national variety, or speech community. Rather, the notion refers to health query variants that become increasingly distant from standard English through various features like function-word omission, non-canonical syntax, phonetic spelling, simplified tenses/aspect, and telegraphic compression.

The investigation of this issue is carried out within the framework of the controlled robustness benchmarking task. Based on the patient's queries about their health issues with reliable references to the expected answer, we generate L0–L3 basilectal-inspired medical query variations preserving the health context but altering the input format. Figure 1 illustrates the resulting L0–L3 continuum: the health intent remains fixed, while the input form moves progressively farther from benchmark-standard English toward a severe basilectal-inspired condition. The performance of six tiny instruction-trained language models is assessed in the free-form generation setting using the clear automatic measures instead of an LLM-based evaluator, such as contextual relevance, lexical overlap, and medical concepts recall.

Our contributions are:

- We introduce a controlled paired benchmark derived from 102 TREC LiveQA Medical questions, with L0–L3 basilectal-inspired health-query variants that keep the medical intent and reference answer fixed while varying input form.
- We evaluate six small instruction-tuned models over 7,416 free-form generations, using transparent non-LLM-judge metrics for contextual similarity, lexical overlap, medical concept recall, and behavioral proxies.
- We find that all 54 baseline model/category/metric L0-to-L3 comparisons show degradation, with the largest aggregate loss in Medical Concept Overlap recall.
- We report secondary analyses of literacy-adaptive prompting, quantization, and feature ablations, while separating supported robustness claims from unsupported clinical, sociolinguistic, or speech-community claims.

2 Related Work

Prior work spans three relevant areas. First, research on World Englishes shows that localized English varieties are socially structured and systematic, not merely erroneous versions of British or American English (Kachru, 1990; Schneider,

2007; Mesthrie and Bhatt, 2008). Studies of Indian English article use and Nigerian English tense-aspect marking further show systematic variation at grammatical and discourse levels (Sharma, 2005; Werner and Fuchs, 2017). This motivates evaluation beyond standard English, while requiring caution: our synthetic variants are robustness probes, not claims of sociolinguistic authenticity.

Second, NLP work has documented dialect underrepresentation and dialect-linked robustness failures. DIALECTBENCH highlights gaps in dialect coverage in standard NLP evaluation (Faisal et al., 2024); dialect robustness studies show degradation on parallel dialectal inputs (Lin et al., 2025); and reward-model analyses suggest that alignment can encode anti-dialect preferences (Mire et al., 2025). However, this literature largely focuses on general NLP, reasoning, or preference modeling rather than consumer medical QA.

Third, medical QA benchmarks such as TREC LiveQA Medical, MedQuAD, MedMCQA, MultiMedQA, and Med-PaLM have advanced evaluation for consumer and biomedical question answering (Ben Abacha et al., 2017; Ben Abacha and Demner-Fushman, 2019; Pal et al., 2022; Singhal et al., 2023). Consumer-health QA research further shows that patient questions differ from professional medical tasks in vocabulary, form, and intent (Welivita and Pu, 2023). Yet these evaluations generally assume standard English input. In parallel, small medical language models are increasingly relevant for private, low-cost, resource-constrained deployment (Garg et al., 2025; Magnini et al., 2025). Reference-based metrics such as BERTScore, ROUGE, and biomedical entity overlap enable auditable free-form answer evaluation without an LLM judge, although they do not certify clinical correctness (Zhang et al., 2020; Lin, 2004; Neumann et al., 2019; Zhao et al., 2026).

Together, prior work leaves a specific gap: small deployable medical QA models have not been systematically tested under graded basilectal-inspired health-query shifts. We address this gap with a controlled paired benchmark, transparent automatic metrics, and analyses of concept omission, confidence behavior, prompting, and quantization.

3 Dataset and Methodology

3.1 Data

We use the medical subset of the TREC 2017 LiveQA task as our source corpus (Ben Abacha et al., 2017). LiveQA is well suited to our setting because its questions are consumer-facing rather than exam-style, and its reference answers are drawn from trusted medical information sources. We use the 102-question evaluation set for systems that answered all questions. Each item consists of an original question q_i, a trusted reference answer r_i, and a coarse analysis label c_i.

We group questions into three deterministic categories: drug and dosage questions ($n = 30$), general medical questions ($n = 47$), and symptom or prevention questions ($n = 25$). These categories are used only for robustness analysis; they are not intended as a clinician-reviewed diagnostic taxonomy.

3.2 Paired Input Construction

Our benchmark is built as a paired input-form shift. For each source item (q_i, r_i, c_i), we construct four variants

$$x_{i,0}, x_{i,1}, x_{i,2}, x_{i,3},$$

where all variants preserve the same medical intent and reference answer r_i, but differ in surface form. This gives a main benchmark of

$$102 \times 4 = 408$$

model-input instances.

The four levels define a synthetic continuum away from benchmark-standard English:

- **L0: Standard benchmark English.** The original normalized LiveQA question.
- **L1: Mild non-standard English.** Morphosyntactic variation with conventional spelling, including article deletion, tense errors, bare verb forms, and agreement errors.
- **L2: Compressed/contact-variety-inspired English.** Structural compression, topic-comment ordering, dropped pronouns, and non-canonical word order.
- **L3: Basilectal-inspired compressed English.** The L1–L2 changes combined with phonetic spelling, function-word deletion, reduplication, and short telegraphic phrasing.

This design is motivated by World Englishes and contact-variety research, which treats localized English forms as patterned rather than random error (Kachru, 1990; Sharma, 2005; Werner and Fuchs,

2017). Mean question length decreases from 29.2 words at L0 to 27.2 at L1, 13.4 at L2, and 7.8 at L3.

Benchmark scope. The resulting data are a controlled robustness benchmark, not a naturally collected corpus of patient writing. We do not claim that the generated variants represent any particular basilect, national variety, or speech community. The benchmark instead operationalizes a graded input-form shift: the medical intent and reference answer remain fixed, while the question moves progressively farther from benchmark-standard English. This design supports claims about content preservation under controlled basilectal-inspired input shift, not claims about sociolinguistic authenticity or clinical safety.

3.3 Models and Generation

Figure 2 summarizes the evaluation pipeline. We evaluate six small instruction-tuned models: Qwen2.5-0.5B-Instruct, Gemma-1B-Instruct, Llama-3B-Instruct, Qwen2.5-3B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen2.5-7B-Instruct. Two originally planned gated checkpoints were replaced with public alternatives; all claims therefore refer to the resolved public model IDs.

All generations use deterministic decoding with sampling set to False and maximum tokens to 200. For each run, we store the prompt, response, generated token IDs, output length, first-five generated-token log probabilities, mean first-five log probability, model ID, quantization setting, question ID, category, level, and run metadata.

3.4 Evaluation Metrics

We evaluate free-form answers against the original reference answer r_i using transparent automatic metrics. We avoid multiple-choice evaluation because answer options can mask comprehension failures. We also avoid LLM-as-judge scoring because judge models may encode standard-language or dialect-related preferences, which would be problematic for a study of basilectal-inspired input shift. The metrics below are therefore used as reproducible stress-test indicators, not as clinical safety measures.

For a model answer $\hat{y}_{i,\ell}^m$ generated by model m for input level ℓ, we compute three primary metrics:

- **Contextual similarity.** A BERTScore-style F1 score against r_i (Zhang et al., 2020). We use `distilbert-base-uncased` embeddings and therefore describe the measure as BERTScore-style rather than the exact `bert-score` package.
- **Lexical overlap.** ROUGE-L F1 against r_i, capturing sequence-level lexical recall (Lin, 2004).
- **Medical Concept Overlap recall.** Biomedical entities are extracted from the reference and model answer using scispaCy (Neumann et al., 2019). If $E(r_i)$ is the set of extracted reference concepts and $E(\hat{y}_{i,\ell}^m)$ is the set of extracted answer concepts, then

$$\mathrm{MCO}(\hat{y}_{i,\ell}^m, r_i) = \frac{|E(\hat{y}_{i,\ell}^m) \cap E(r_i)|}{|E(r_i)|}.$$

We also compute behavioral proxies to interpret the primary metrics: generic-response rate, abstention or confusion patterns, semantic drift, unsupported medical concept rate, response length, and mean first-five-token log probability as an internal confidence proxy. These measures do not establish clinical correctness; they indicate whether basilectal-inspired inputs produce shorter answers, generic disclaimers, omitted concepts, or confidence-quality mismatch.

3.5 Analysis

Results are aggregated by model, level, category, condition, and metric. The main robustness quantity is the paired drop from standard input to a shifted level:

$$\Delta_\ell^{m,k} = S_0^{m,k} - S_\ell^{m,k},$$

where $S_\ell^{m,k}$ is the mean score for model m on metric k at level ℓ. Positive values indicate degradation relative to L0.

For intervention experiments, recovery is computed at the same model, category, level, and metric:

$$R_\ell^{m,k} = S_{\ell,\text{intervention}}^{m,k} - S_{\ell,\text{baseline}}^{m,k}.$$

For quantization, we compare the L0-to-L3 drop under INT8 or INT4 against the corresponding FP16 drop for the same role slot. For feature ablations, each isolated perturbation is compared against the clean L0 condition on the balanced ablation subset.

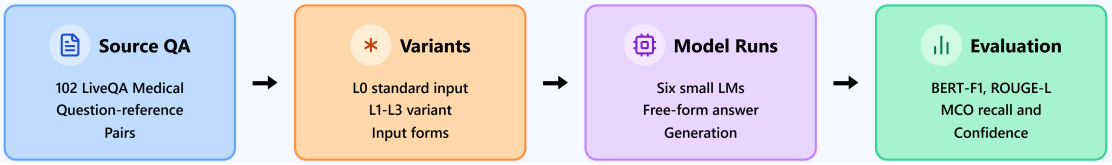

Figure 2: Evaluation pipeline. TREC LiveQA Medical questions are expanded into L0–L3 variants, answered by small models under baseline, quantized, and adaptive-prompt conditions, and scored with answer-quality, medical-concept, confidence, and behavioral metrics.

Level	BERT-F1	ROUGE-L	MCO	Logprob
L0	0.707	0.136	0.129	-0.316
L1	0.708	0.136	0.135	-0.304
L2	0.685	0.126	0.107	-0.329
L3	0.614	0.095	0.046	-0.392

Table 1: Aggregate baseline scores by input level. BERT-F1 denotes BERTScore-style contextual similarity; MCO denotes Medical Concept Overlap recall.

4 Results and Discussion

4.1 Main Result: Degradation Emerges at the Basilectal-Inspired End

The central finding is that small medical QA models tolerate mild non-standard variation but degrade sharply as questions move toward the basilectal-inspired end of the input continuum. In the baseline FP16 setting, all 54 model/category/metric L0-to-L3 comparisons show positive drops. Paired bootstrap intervals over questions support the same conclusion: every L0-minus-L3 mean drop has a positive 95% lower bound. Even the smallest observed drop, ROUGE-L in one drug/dosage slice, remains positive (0.019; 95% CI 0.008–0.032).

Table 1 shows a clear threshold pattern. L1 remains nearly indistinguishable from L0, suggesting that mild morphosyntactic deviations alone do not meaningfully reduce answer quality. The decline begins at L2, where compression and non-canonical ordering are introduced, and becomes substantial at L3. From L0 to L3, BERTScore-style F1 falls from 0.707 to 0.614, ROUGE-L from 0.136 to 0.095, and Medical Concept Overlap recall from 0.129 to 0.046.

The MCO drop is the most important result for our setting because it measures preservation of extracted medical concepts from the trusted reference answer. The models are therefore not merely producing lexically different answers; they are preserving less medical content. The strongest supported claim is narrow but consequential: under controlled basilectal-inspired input shift, small medical QA models become less reference-aligned and less complete in medically relevant information.

4.2 The Pattern Holds Across Models and Question Types

The degradation is not an artifact of one model family, size range, or question category. At L3, BERTScore-style drops range from 0.058 to 0.106, ROUGE-L drops from 0.030 to 0.054, and MCO drops from 0.062 to 0.095. Larger models reduce the effect but do not remove it: Qwen2.5-7B still loses 0.080 BERTScore-style F1 and 0.080 MCO recall at L3. Mistral-7B has the smallest BERTScore-style drop, but also the highest L3 generic-response rate at 14.7%, indicating that contextual similarity alone can hide important incompleteness.

The same broad pattern appears across question types. At L3, drug/dosage questions have the lowest ROUGE-L and MCO values, reaching 0.086 and 0.041, respectively. Symptom/prevention questions reach 0.109 ROUGE-L and 0.048 MCO, while general medical questions reach 0.604 BERTScore-style F1 and 0.047 MCO. We therefore do not claim that one category is uniformly most fragile. The more robust conclusion is that basilectal-inspired input shift reduces medical concept preservation across categories.

4.3 The Failure Mode Is Fluent Incompleteness

The degradation is not simply refusal or abstention. L3 generic-response rates range from 0.0% to 14.7%, so some models do fall back to generic medical advice. However, metric drops also occur when generic responses are uncommon. Models often continue answering, but the answers become less specific and preserve fewer reference-aligned medical concepts.

We call this failure mode *fluent incompleteness*: the model produces plausible medical text while losing content needed for a complete answer. This matters because a response may appear coopera-

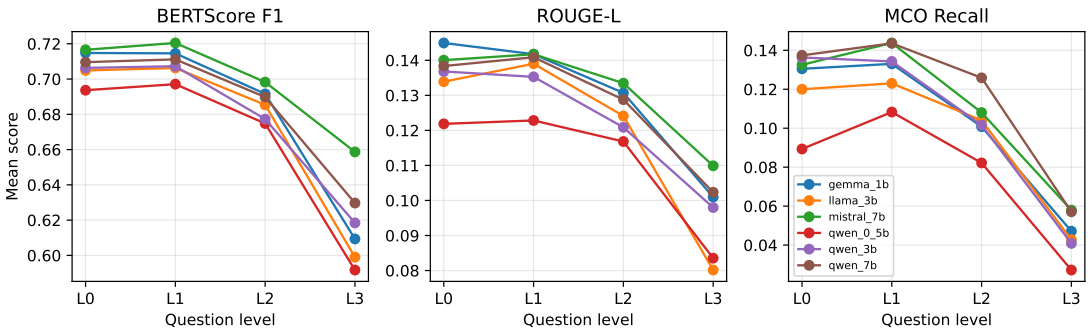

Figure 3: Score degradation across input levels. Mild morphosyntactic variation is largely tolerated, whereas structural compression and the L3 basilectal-inspired condition produce clear losses across all primary metrics.

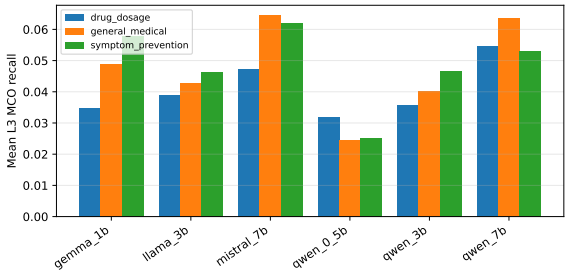

Figure 4: L3 Medical Concept Overlap by category and model. Concept preservation is reduced across question types under the basilectal-inspired condition.

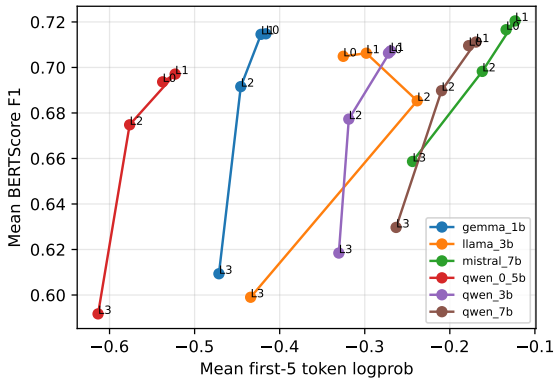

Figure 5: Internal confidence proxy versus contextual similarity by input level and model.

tive and medically relevant even when it omits key information from the trusted reference.

The confidence proxy supports the same interpretation. As inputs move from L0 to L3, contextual similarity decreases while models still produce fluent outputs. The risk is therefore not only visible failure, but quiet degradation: an answer can sound adequate while becoming medically thinner.

4.4 Structural Perturbations Explain More Than Spelling Alone

The feature-ablation results help explain why degradation becomes severe at L2 and L3. On the balanced drug/symptom subset, topic-comment and word-order perturbation has the largest mean MCO impact (0.056). In contrast, bare verb forms, article deletion, phonetic medical spelling, and zero copula have near-zero aggregate effects.

This suggests that the L3 failure is not primarily a spelling-normalization problem. In this benchmark, models lose more medical content when the question's information structure is reorganized than when isolated surface features are perturbed. This also explains why models can handle mild non-standard grammar at L1 but degrade under the compressed and reordered forms introduced at L2 and L3.

4.5 Prompting Helps Little, and Quantization Is Not the Main Driver

If the main failure is loss of medical content under reordered and compressed input, then a simple instruction to interpret non-standard English should only partly help. That is what we observe. The literacy-adaptive prompt improves L3 BERTScore-style F1 by 0.020 on average, but ROUGE-L improves by only 0.003 and MCO recall by 0.001. At L1 and L2, average recovery is slightly negative across the three primary metrics. Prompting therefore improves surface-level contextual similarity slightly at the most severe level, but it does not restore medical concept preservation.

Quantization also does not explain the main result. INT8 interactions are near zero across BERTScore-style F1, ROUGE-L, and MCO recall. INT4 effects are metric-dependent, with interactions of 0.016 for BERTScore-style F1, 0.002 for ROUGE-L, and -0.006 for MCO recall. Low-bit

deployment was tested because local models are often compressed, but the dominant degradation is linguistic: as the same health intent moves farther from benchmark-standard English, models preserve less reference-aligned medical content.

Taken together, these findings narrow the mitigation target. The problem is unlikely to be solved by a charitable prompt alone, and it is not simply an artifact of low-bit inference. Stronger mitigation likely requires medical term normalization, query rewriting, retrieval from trusted sources, uncertainty-aware abstention, and human escalation for high-risk cases.

5 Conclusion

We introduced a controlled benchmark for evaluating small medical QA models under graded basilectal-inspired health-query shifts. Across six small-model role slots, three question categories, and 7,416 free-form generations, answer quality declines as inputs move away from benchmark-standard English. The degradation is most pronounced at L3, where BERTScore-style F1, ROUGE-L, and Medical Concept Overlap all drop substantially. The most important failure is not simple refusal, but *fluent incompleteness*: models often continue producing plausible medical text while preserving fewer medically relevant concepts from trusted reference answers.

The findings suggest that basilectal-inspired input shift should be treated as a core evaluation condition for local medical AI. Standard medical QA evaluation usually measures performance on clean benchmark English, yet the settings that motivate small local models—limited connectivity, privacy constraints, high data costs, and weaker access to professional care—may also include users whose written English differs substantially from benchmark-standard form. If systems degrade for these users, access-oriented deployment may reproduce the very gaps it aims to reduce.

The claim is deliberately bounded. The benchmark is synthetic: it enables controlled paired comparisons, but it does not replace naturally collected patient-writing datasets or natural basilectal corpora. The first-pass L3 validation rate shows that severe basilectal-inspired rewriting is difficult and required regeneration before inclusion. The category labels are deterministic rather than clinician-reviewed. Automatic metrics are transparent and scalable, but they do not certify clinical correctness or clinical safety. The model set is limited to the resolved public small-model role slots tested here.

The supported conclusion is therefore precise: within this controlled benchmark, basilectal-inspired input shift reduces reference alignment and medical concept preservation in small local medical QA models. Future work should extend this evaluation with naturally occurring patient questions, broader English varieties, clinician review of safety-critical omissions, and deployment-oriented mitigation pipelines that reduce fluent incompleteness without making systems less useful for the users they are intended to serve.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at TREC 2017 LiveQA. In *Proceedings of the Twenty-Sixth Text REtrieval Conference*. National Institute of Standards and Technology.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: A NLP benchmark for dialects, varieties, and closely-related languages. *Preprint*, arXiv:2403.11009.

Muskan Garg, Shaina Raza, Shebuti Rayana, Xingyi Liu, and Sunghwan Sohn. 2025. The rise of small language models in healthcare: A comprehensive survey. *Preprint*, arXiv:2504.17119.

Braj B. Kachru. 1990. World englishes and applied linguistics. *World Englishes*, 9(1):3–20.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342,

Vienna, Austria. Association for Computational Linguistics.

Matteo Magnini, Gianluca Aguzzi, and Sara Montagna. 2025. Open-source small language models for personal medical assistant chatbots. *Intelligence-Based Medicine*, 11:100197.

Rajend Mesthrie and Rakesh M. Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge University Press, Cambridge.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected dialects: Biases against African American Language in reward models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7483–7502, Albuquerque, New Mexico. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Edgar W. Schneider. 2007. *Postcolonial English: Varieties Around the World*. Cambridge University Press, Cambridge.

Devyani Sharma. 2005. Language transfer and discourse universals in Indian English article use. *Studies in Second Language Acquisition*, 27(4):535–566.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Malitha Seneviratne, Paul Gamble, Chris Kelly, Abdelkareem Babiker, Nathanael Scharli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Anuradha Welivita and Pearl Pu. 2023. A survey of consumer health question answering systems. *AI Magazine*, 44(4):482–507.

Valentin Werner and Robert Fuchs. 2017. The present perfect in Nigerian English. *English Language and Linguistics*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations*.

Min Zhao, Inez Y. Oh, Aditi Gupta, Sally Cohen-Cutler, Kathryn M. Harmoney, Albert M. Lai, and Bryan A. Sisk. 2026. Automating Evaluation of LLM-generated Responses to Patient Questions About Rare Diseases. *JAMIA Open*, 9(2):ooag054.