The Text Aphasia Battery (TAB): A Clinically-Grounded Benchmark for Aphasia-Like Deficits in Language Models

Nathan Roll[1], Jill Kries[1], Flora Jin[2], Catherine Wang[5,6], Ann Marie Finley[3], Meghan Sumner[1], Cory Shain[1], Laura Gwilliams[1]

[1]Stanford University, [2]University of California, San Francisco, [3]Temple University, [5]San Diego State University, [6]University of California, San Diego

Abstract

Large language models (LLMs) have emerged as a candidate 'model organism' for human language, offering an unprecedented opportunity to study the computational basis of linguistic disorders like aphasia. However, traditional clinical assessments are ill-suited for LLMs, as they presuppose human-like pragmatic pressures and probe cognitive processes not inherent to artificial architectures. We introduce the **Text Aphasia Battery (TAB)**, a text-only benchmark adapted from the Quick Aphasia Battery (QAB) to assess aphasic-like deficits in LLMs. The TAB comprises four subtests: Connected Text, Word Comprehension, Sentence Comprehension, and Repetition. This paper focuses on the design and validation of TAB as an assessment instrument, rather than on drawing conclusions from lesioned language models. Downstream work applies TAB to artificially ablated models to study how language behavior changes under parameter perturbation (Roll et al., 2026). To facilitate large-scale use, we validate an automated evaluation protocol using Gemini 2.5 Flash, which achieves reliability comparable to expert human raters (prevalence-weighted Cohen's $\kappa = 0.255$ for model–consensus agreement vs. 0.286 for human–human agreement). We release TAB as a clinically-grounded, scalable framework for analyzing language deficits in artificial systems.

1 Introduction

The clinical framework for assessing aphasia, a language impairment resulting from brain damage, has long relied on batteries such as the Western Aphasia Battery–Revised (WAB-R) (Kertesz, 2022; Clark et al., 2020), the Boston Diagnostic Aphasia Examination (BDAE), the Comprehensive Aphasia Test (CAT) (Bruce and Edmundson, 2010; Springer and Mantey, 2010), and the Quick Aphasia Battery (QAB) (Wilson et al., 2018). Recently, researchers have begun exploring connections between computational language models and aphasia, for instance using language model surprisals in tandem with clinical features to predict aphasia subtypes (Cong et al., 2024). However, directly evaluating language models using traditional aphasia batteries faces a fundamental challenge: these batteries use multimodal examination to probe deficits arising from lesions to specific brain structures. Their questions and scoring criteria make implicit assumptions about patterns of behavioral deficits that have no analog in text-only computational models.

To address this challenge, we introduce the **Text Aphasia Battery (TAB)**. Developed with speech-language pathologists, the TAB is a benchmark that adapts core components of the clinically-validated Quick Aphasia Battery (QAB) (Wilson et al., 2018) for text-constrained environments. We posit that while aphasia is a multimodal disorder, a significant component of its behavioral signature is identifiable and recognizable in transcribed text alone. Leveraging the availability of large-scale transcribed datasets like AphasiaBank, we can isolate these text-based linguistic patterns to study language breakdown at scale. The TAB reframes aphasia assessment from a neuropsychological tool to a *behavioral benchmark*, identifying patterns of language degradation in LLMs that are This distinction is central to our contribution. We do not claim that LLMs can be clinically aphasic, nor that parameter ablations are etiologically equivalent to stroke, tumor, or neurodegeneration. Rather, TAB operationalizes a narrower behavioral question: when only text is available, can clinically grounded language features be scored reliably in human and model outputs?

We intentionally separate instrument validation from lesion-based model analysis. A companion line of work uses TAB to evaluate artificially lesioned models (Roll et al., 2026); the present paper establishes the benchmark and scoring protocol

Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 340–354

needed for such studies. We detail its design principles, four subtests, and clinical grounding, offering it as a resource for standardized, text-only evaluation of LLMs. To facilitate large-scale adoption, we also present and validate an automated evaluation protocol that achieves high human-expert rater reliability, ensuring the benchmark is both robust and scalable. We release the TAB to provide the field with a new tool for assessing linguistic breakdown in artificial systems.

2 Background: Traditional Aphasia Batteries

Aphasia batteries are standardized assessments that evaluate language abilities in individuals with aphasia, defined as receptive and/or expressive language deficits due to brain damage resulting from stroke, traumatic brain injury, brain tumor, or neurodegenerative processes. These tests are essential in clinical practice, enabling speech-language pathologists (SLPs) to identify the presence, type, and severity of aphasia, inform treatment, and monitor recovery.

2.1 Major Aphasia Assessment Tools

Boston Diagnostic Aphasia Examination (BDAE) The BDAE is a comprehensive battery designed to diagnose aphasia and related disorders. It evaluates a broad range of language skills, including conversation, narrative speech, auditory comprehension, oral expression, repetition, reading, and writing, allowing for the classification of aphasia syndromes based on symptom patterns.

Western Aphasia Battery–Revised (WAB-R) One of the most widely used aphasia assessments, the WAB-R evaluates language and related cognitive functions through 8 subtests comprising 32 short tasks (Kertesz, 2022). It assesses spontaneous speech, auditory comprehension, repetition, naming, reading, and writing, along with nonlinguistic skills like apraxia and calculation. The test produces an Aphasia Quotient (AQ) to quantify severity and classify aphasia types such as Broca's or Wernicke's (Clark et al., 2020).

Comprehensive Aphasia Test (CAT) The CAT evaluates the recognition, comprehension, and production of spoken and written language through three components: cognitive screening, language assessment, and a disability questionnaire, with the latter assessing functional communication and psychosocial factors (Bruce and Edmundson, 2010; Springer and Mantey, 2010). The CAT is a reliable and validated tool for assessing underlying language impairments in adults with aphasia (Halai et al., 2022).

Quick Aphasia Battery (QAB) The QAB is a rapid, multidimensional assessment that provides a reliable language evaluation in 15 to 20 minutes (Wilson et al., 2018). It evaluates major language domains through eight subtests and uses a graded scoring system sensitive to changes in language function over time.

2.2 Common Assessment Domains

Major aphasia batteries share several core evaluation targets that are central to language function (Salter et al., 2006; Wilson et al., 2018). These include auditory comprehension of words and sentences, repetition of words, phrases, and sentences, naming based on auditory or written description, naming a visually presented object, reading comprehension, and writing production. Many batteries also assess adjunct domains such as praxis (motor programming), arithmetic, and visuoconstruction to distinguish aphasia from other cognitive disorders (Kertesz, 2022).

3 Limitations of Traditional Batteries for LLM Evaluation

Aphasia batteries have proven invaluable for diagnosing and managing aphasia in clinical settings. However, their design and underlying assumptions make them fundamentally ill-suited for evaluating LLMs. These limitations arise from multimodal dependencies, functional communication, and the inclusion of psychosocial measures.

3.1 Non-Linguistic and Contextual Dependencies

Clinical assessments require integration across sensory modalities (visual, auditory, gestural) and rely on human-specific psychosocial context. For example, picture naming tasks and verbal repetition with prosodic cues are common in traditional batteries (Wilson et al., 2018). Furthermore, these assessments often incorporate metacognitive measures probing self-awareness and motivation (Bruce and Edmundson, 2010; Springer and Mantey, 2010).

Many text-only LLMs, however, lack auditory perception, sensory grounding, and the pragmatic intent or self-awareness required for these tasks. They cannot perceive images, produce gestures, or

exhibit speech-motor impairments (e.g., apraxia). Therefore, it is impossible to apply these assessments directly. This mismatch highlights the need for benchmarks that isolate linguistic competence from these non-linguistic dependencies.

3.2 Subjectivity and Granularity of Measurement

Traditional batteries often rely on subjective human rating, which may not capture the full granularity of linguistic impairment or the complete range of disease presentation. Clinical ratings can be variable and may miss subtle distributional patterns in speech. Shifting to text-only evaluation enables the objective, scalable analysis of specific linguistic dimensions—such as lexical retrieval, syntactic complexity, and discourse coherence—that are robustly identifiable in transcripts. This allows for a more granular characterization of linguistic breakdown than is typically possible with composite clinical scores.

3.3 Functional Communication vs. Structural Language

Aphasia tests measure deficits in *functional communication*, the ability to use language flexibly for communicative purpose and meaning (Wilson et al., 2018). This includes pragmatic skills, contextual understanding, and the ability to convey meaning effectively. In contrast, to the extent that LLMs have social goals and intents, these are at best quite different from those of humans, and therefore the notion of functional communication does not necessarily apply to LLMs in the ways assumed by existing aphasia batteries. Nonetheless, beyond functional communication, aphasias also plausibly implicate aspects of structural language competence that appear to be present (and impairable) in LLMs (Mahowald et al., 2024), including grammatical accuracy, textual coherence, and fluency. To use LLMs to study such dimensions in aphasia, LLM-appropriate assessment strategies are needed.

4 Materials: The Text Aphasia Battery (TAB)

To address the limitations of traditional aphasia batteries while preserving their clinical insights, we developed the TAB, a text-only benchmark that adapts core QAB components for modality-constrained evaluation. The TAB focuses on *linguistic competence* rather than neuropsychological deficit mapping, which makes it suitable for large-scale LLM assessment.

This shift to text is justified because the core linguistic dimensions affected in aphasia—lexical retrieval (anomia), syntactic structure (agrammatism, paragrammatism), discourse coherence (empty speech), and repetition fidelity—are directly identifiable in transcribed text. By isolating these features, the TAB provides a principled lens for characterizing language breakdown in computational systems without requiring multimodal or psychosocial constructs. We developed the TAB in close collaboration with speech-language pathologists and aphasiologists to ensure that its adapted tasks maintain clinical relevance while being suitable for automated computational evaluation.

4.1 Design Principles

The TAB adheres to three core design principles. The first is **modality constraint**: all inputs and outputs are text-based, eliminating dependencies on auditory perception, visual naming, or motor production. The second is **computational interpretability**: scoring focuses on observable linguistic patterns (e.g., morphosyntactic errors, semantic paraphasias) that can be systematically identified through automated analysis. The third principle is **clinical grounding**: subtests and evaluation criteria are derived from established clinical aphasia research. This includes the QAB framework (Wilson et al., 2018) and the APROCSA (Auditory-Perceptual Rating of Connected Speech in Aphasia) system (Casilio et al., 2019), with additional insights from clinical discourse analysis methods (MacWhinney et al., 2011; Forbes et al., 2012). We validated the adaptation process with clinical experts.

4.2 TAB Subtests

The TAB consists of four subtests that evaluate complementary aspects of linguistic function (see Table 3 in the Appendix for a full overview). The complete battery includes approximately 104 unique prompts: 48 word comprehension prompts, 12 connected text prompts, 36 repetition prompts, and 8 picture description prompts. Representative examples from each subtest are provided below:

- **Connected Text**, which evaluates fluency, grammaticality, and discourse coherence using 5 open-ended prompts. Responses are scored for 19 linguistic features based on the

APROCSA framework (Casilio et al., 2019).

- **Word Comprehension**, which assesses lexical-semantic processing with 5 forced-choice questions, requiring selection from six competing alternatives.
- **Sentence Comprehension**, which tests syntactic processing through 5 Yes/No questions involving passive voice, negation, and conditional reasoning.
- **Repetition**, which measures morphosyntactic integrity and attentional stability by requiring exact reproduction of 5 items of increasing length and complexity.

4.2.1 Connected Text

Objective: Evaluate fluency, grammaticality, and discourse coherence.

Instructions: The LLM system is prompted to respond to five open-ended prompts in 3 to 5 full sentences. To elicit naturalistic responses and avoid meta-commentary (e.g., "As an AI, I don't have personal experiences"), we employ system prompts that establish a conversational context without explicitly requesting the model to assume a persona. The prompts are: "Tell me about the best trip you ever took," "Describe a happy childhood memory," "Tell me about your first job," "What do you like about where you live?," and "Describe the steps to make a simple meal."

Evaluation: Responses are analyzed for 19 aphasic features adapted from the APROCSA (Auditory-Perceptual Rating of Connected Speech in Aphasia) framework (Casilio et al., 2019). These features include anomia, paraphasias (semantic and phonemic), agrammatism (omission of bound morphemes or function words), paragrammatism, empty speech, perseverations, neologisms, and overall communication impairment. Each feature is scored as present (1) or absent (0).

4.2.2 Word Comprehension

Objective: Evaluate lexical-semantic processing and selection among competing alternatives.

Instructions: The system responds verbatim to five forced-choice items, with six options provided for each. The foils were selected to be phonologically and semantically varied. The items are: "Which one is an animal with a mane? (lion, drum, violin, giraffe, boot, boat)," "Which object is typically used to make music? (violin, giraffe, lion, door, boot, boat)," "Which item is usually worn on the feet? (boot, boat, lion, drum, violin, giraffe)," "Which object is used for cutting? (knife, kite, lion, drum, violin, giraffe)," and "Which one is a large mammal with a long neck? (giraffe, horse, lion, drum, violin, boot)."

Evaluation: Binary correct/incorrect scoring. Errors may indicate deficits in semantic processing or susceptibility to phonological similarity effects. The six-option format increases task difficulty and reduces chance performance.

4.2.3 Sentence Comprehension

Objective: Evaluate syntactic processing, comprehension of passive structures, and logical reasoning.

Instructions: The system responds verbatim (Yes/No) to five items: "Are babies watched by babysitters?" (Expected: Yes), "Do you cut the grass with an axe?" (Expected: No), "If you're about to leave, have you left yet?" (Expected: No), "Are witnesses questioned by police?" (Expected: Yes), and "If I was at the park when you arrived, did I get there first?" (Expected: Yes).

Evaluation: Binary correct/incorrect scoring. Errors may indicate difficulties with passive voice, negation, or temporal/conditional reasoning.

4.2.4 Repetition

Objective: Evaluate morphosyntactic integrity and exact reproduction.

Instructions: The system repeats five items exactly: "house," "breakfast," "catastrophe," "The sun rises in the East," and "The ambitious journalist discovered where we'd be going."

Evaluation: Exact match required. Errors (substitutions, deletions, insertions) indicate morphosyntactic processing deficits or instability in representation. For an LLM, this task is not a probe of the auditory-verbal loop. It is a test of attentional stability and the ability to resist semantic drift or elaboration. Failure on this verbatim copy task can reveal subtle deficits in sequence-to-sequence transduction. The model might paraphrase, overgeneralize, or otherwise deviate from the source text. This provides a window into its capacity for precise information transfer.

5 Method: Automated Evaluation

A key design goal of the TAB is scalability through automated evaluation. While Subtests 2–4 can be scored algorithmically (exact string matching or

binary classification), Subtest 1 (Connected Text) requires nuanced linguistic analysis. We employ in-context learning with Gemini 2.5 Flash to identify aphasic features systematically. The taxonomy comprises 19 binary indicators spanning five categories (Semantic, Syntactic, Fluency, Phonological, Other). This automated scoring protocol enables large-scale analysis and achieves inter-rater reliability comparable to expert clinical raters when properly weighted by feature prevalence (see Section 6 for validation details).

Several TAB features operationalize constructs that have been used in prior computational work to quantify aphasic deficits. Anomia (word-finding difficulty) corresponds to reduced lexical diversity, measured by metrics like type-token ratio (TTR), moving-average TTR (MATTR), and measure of textual lexical diversity (MTLD). Agrammatism and telegraphic speech map onto low syntactic complexity scores derived from dependency parsing or phrase-structure depth. Semantic incoherence can be captured by perplexity, which correlates with clinical ratings of discourse coherence. Traditional approaches to automated aphasia assessment extract these features through rule-based algorithms or trained classifiers. The TAB instead uses LLM-based classification to identify these patterns directly from text, maintaining the interpretability of feature-based approaches while leveraging the pattern-recognition capacity of large models.

5.1 APROCSA-Based Feature Set

Connected Text responses are evaluated for 19 features adapted from the APROCSA (Auditory-Perceptual Rating of Connected Speech in Aphasia) system (Casilio et al., 2019). APROCSA is an auditory-perceptual rating system for connected speech in aphasia that assesses 27 features of connected speech on a five-point scale. For the TAB, we selected 19 features applicable to text-only evaluation and adapted them to binary scoring. The lexical features are anomia, semantic paraphasias, phonemic paraphasias, and neologisms. Fluency and productivity features are empty speech and short and simplified utterances. Morphosyntactic features include omission of bound morphemes, omission of function words, and paragrammatism. Disfluency features consist of abandoned utterances, false starts, retracing, and conduite d'approche. Perseverative features are perseverations and stereotypies/automatisms (the latter as a combined feature). Coherence features include jargon, meaning unclear, and off-topic utterances. A final feature is overall communication impairment.

5.2 In-Context Prompting Protocol

We provide the evaluating LLM with a definition and an example for each of the 19 features, adapted from the APROCSA framework (Casilio et al., 2019). We also provide two annotated example transcripts with ground-truth feature labels. Finally, we provide instructions to output a JSON object with each feature as a key and binary (0/1) values. This few-shot prompting approach enables consistent feature identification across large datasets without requiring manual annotation. The prompt template is included in the TAB's repository and can be adapted for different LLM architectures.

6 Validation of Automated Evaluation

To validate our automated protocol, we conducted an inter-rater reliability (IRR) study comparing our Gemini 2.5 Flash-based system against a ground truth derived from expert human annotators.

6.1 Data Collection and Annotation

The validation dataset for this study comprises 561 English text samples from two sources.

Human Data We sampled 306 transcribed responses from individuals with aphasia, sourced from the AphasiaBank database (MacWhinney et al., 2011). These samples cover a range of aphasia types and severities, providing a clinically relevant baseline.

AI Data For validation, we deliberately included 255 AI-generated samples from artificially ablated language models, including GPT-2 (Radford et al., 2019), Pythia (Biderman et al., 2023), and Llama (Touvron et al., 2023). These samples served as stress-test items for the scoring protocol, increasing the diversity of text patterns rated by human experts and the automated judge. The present paper does not analyze which components produce which deficits, nor does it claim that parameter ablations correspond etiologically to human brain lesions. Those scientific questions are addressed in downstream lesioning work using TAB (Roll et al., 2026).

Expert Annotation A pool of five expert speech-language pathologists (SLPs) collectively annotated the dataset of 561 samples through a custom

web interface. For each sample, SLPs rated the presence (1) or absence (0) of the 19 APROCSA-based features. These manual ratings, stored in a Firestore database, form our ground truth consensus.

6.2 Inter-Rater Reliability

From a dataset of 561 text samples, 82 were annotated by multiple (2–3) speech-language pathologists from a pool of five experts. We established ground truth via majority vote on these items and used Cohen's Kappa (κ) to measure agreement.

A significant methodological issue emerged from this analysis: the choice of how to aggregate agreement scores across the 19 evaluated aphasic features. An unweighted macro-average, which treats all features equally, is problematic when features have vastly different clinical prevalence rates. In our validation set, some features never occurred (zero positive instances), while others like Perseverations appeared in 34% of samples. For features with zero prevalence, human-human agreement is undefined (κ cannot be calculated when there is no variation), while the automated system trivially achieves perfect agreement by also never detecting these features. This creates an artifact: the model appears to have perfect agreement on zero-prevalence features (contributing 1.0 to its average), while human raters contribute 0.0 or undefined values, artificially inflating the unweighted model average while deflating the human average.

To provide a more clinically meaningful result, we report the **prevalence-weighted Cohen's Kappa**, where each feature's score is weighted by its positive instance count in the validation set. Prevalence-weighting is standard in meta-analytic contexts (Cohen, 1968) because it ensures that estimates reflect the real-world distribution of phenomena. In our case, this is particularly appropriate because clinically important features (those that actually occur in aphasic language) should carry more weight than features absent in our sample. This approach aligns with clinical practice, where diagnostic decisions are based on observable symptoms rather than the absence of rare phenomena.

As shown in Table 1, this weighted analysis reveals that humans and the automated system achieve **comparable performance**. The weighted human–human agreement ($\kappa = 0.286$) is close to the model–consensus agreement ($\kappa = 0.255$), with both falling into a commonly described "fair" range. We acknowledge that these agreement scores are relatively low, which likely reflects the inherent difficulty of the annotation task—identifying subtle linguistic features in text without prosodic or gestural cues. Future improvements might include providing annotators with more surrounding linguistic context, developing more explicit annotation guidelines with additional examples, or incorporating a training phase to calibrate annotators. Despite these challenges, the comparable performance between human and automated annotation suggests that automated evaluation can serve as a reliable screening tool when paired with expert review, with complementary strengths supporting a hybrid workflow where automated screening is paired with expert review of high-frequency clinical features.

We also computed Gwet's AC1, a prevalence-robust agreement coefficient designed to address challenges associated with Cohen's κ. Objective symptoms with clear operational definitions (e.g., neologisms, omission of bound morphemes, perseverations) showed mean AC1 $= 0.71$, while subjective symptoms requiring holistic judgment (e.g., meaning unclear, short and simplified utterances, empty speech) showed mean AC1 $= 0.61$. This pattern is consistent with known challenges in clinical rating scales: subjective symptoms showed $2.5\times$ higher prevalence and $2\times$ higher variance than objective symptoms.

6.3 Disagreement Analysis

To understand where human-model disagreements concentrate, we analyzed symptom-level patterns across the validation set. Disagreements cluster predictably in symptoms requiring holistic judgment rather than symptoms with clear operational definitions. Subjective symptoms such as *Meaning unclear* (prevalence $= 45\%$, variance $= 0.25$) and *Short and simplified utterances* (prevalence $= 44\%$, variance $= 0.25$) showed the highest variance across annotators, reflecting inherent ambiguity in clinical assessment. In contrast, objective symptoms with clear phonological or morphological criteria, such as *Neologisms* (prevalence $= 5\%$, variance $= 0.05$), *Omission of bound morphemes* (prevalence $= 4\%$, variance $= 0.04$), and *Perseverations* (prevalence $= 11\%$, variance $= 0.10$), showed substantially lower variance. This pattern matches known challenges in clinical aphasiology: even expert speech-language pathologists disagree on features like "empty speech" that require integrating multiple cues into a gestalt judgment. The automated TAB shows the same pattern of uncer-

tainty as human raters, suggesting it captures genuine task difficulty rather than model-specific artifacts.

Aggregation Method	Human κ	Model κ
Unweighted average	0.061	0.794
Weighted	**0.286**	**0.255**
Excl. zero-prevalence	0.215	0.381

Table 1: Inter-rater reliability (Cohen's κ) for human vs. automated ratings on 82 multiply-annotated samples. The prevalence-weighted average, our primary metric, shows comparable performance. The unweighted average is misleadingly high for the model due to agreement on many features with zero positive instances in the validation set.

7 Discussion

7.1 TAB as a Behavioral Benchmark

Unlike traditional aphasia batteries designed for neuropsychological diagnosis, the TAB functions as a *behavioral benchmark*. It reveals patterns of behavioral breakdown in text-based systems. When LLMs exhibit TAB-identified deficits (e.g., agrammatism, semantic paraphasias), these patterns do not imply neurological dysfunction. Rather, they reflect limitations in the model's training data, architectural properties, or induced representational damage (in the case of ablations or lesioning interventions). The release of the TAB provides the research community with a standardized tool to explore these phenomena systematically.

This behavioral focus enables a novel form of investigation: by systematically characterizing how language breaks down in LLMs, we can generate hypotheses about the computational underpinnings of human aphasic deficits. Conversely, validating that LLMs exhibit recognizable, clinically-grounded patterns of language breakdown is a necessary step toward using them as tractable models for studying human aphasia at scale.

7.2 Bridging Clinical and Computational Linguistics

The TAB establishes a principled connection between clinical aphasiology and computational linguistics. By grounding evaluation in established clinical frameworks like the QAB (Wilson et al., 2018) and APROCSA (Casilio et al., 2019), it enables researchers to systematically compare LLM performance against human aphasic profiles. It also helps identify specific linguistic competencies (e.g., morphosyntactic processing, semantic coherence) that may be compromised under certain conditions. Finally, it allows researchers to develop targeted interventions, such as fine-tuning strategies or architectural modifications, informed by clinical insights into language breakdown.

Conversely, this framework positions LLMs as powerful testbeds for theories of language processing relevant to aphasiology. With a validated instrument in hand, researchers can conduct lesion-style studies in artificial systems by applying interventions such as architectural ablation, parameter pruning, or adversarial fine-tuning and then scoring the resulting text with TAB. These interventions should be interpreted as causal perturbations of artificial systems, not as direct analogues of human lesion etiology. This approach offers a novel methodology for testing hypotheses about the functional architecture of language, such as the relationship between syntactic and semantic processing, and for modeling patterns of language breakdown that could, in turn, inform clinical neuroscience.

7.3 Baseline Model Performance

To illustrate the TAB's utility for model evaluation, Table 2 presents baseline performance (0% ablation) across five language models from three families. These results reveal substantial variation in linguistic robustness even among unlesioned models.

Base models exhibit notably higher symptom burdens than their instruction-tuned counterparts. Llama-3.2-1B shows the highest baseline symptom burden (3.21 symptoms per response), while Llama-3.2-1B-Instruct achieves near-zero symptom rates (0.15 mean burden), demonstrating that instruction tuning substantially improves linguistic robustness on the TAB.

Among instruction-tuned models, Llama-3.2-1B-Instruct shows the best performance (0.15 burden), followed by Gemma-3-1B-it (0.35 burden). OLMo-2 models show intermediate performance, with the instruct variant (0.83 burden) substantially outperforming its base counterpart (1.98 burden). These patterns should be read as calibration results, not as evidence that unlesioned models "have aphasia." They show that TAB is sensitive to differences in task compliance and text quality across model families and training regimes, providing a reference point for future perturbation studies.

Table 2: Baseline TAB performance across tested language models. Symptom burden is the mean count of symptoms per response.

Model	N	Burden
Llama-3.2-1B	6720	3.21
OLMo-2-1B	5040	1.98
OLMo-2-1B-Instruct	2688	0.83
Gemma-3-1B-it	7280	0.35
Llama-3.2-1B-Instruct	6720	0.15

7.4 Applications

The TAB is suitable for diverse applications, including **model evaluation** for systematic assessment of LLM linguistic competencies, **ablation studies** for evaluating the effects of architectural changes on linguistic integrity, **adversarial robustness** testing to probe model stability, and **interpretability research** for mapping internal representations to observable linguistic patterns.

8 Conclusion

We have presented the Text Aphasia Battery (TAB), a benchmark for evaluating aphasic-like linguistic deficits in modality-constrained environments. By adapting clinical aphasia assessment principles for text-only evaluation, the TAB addresses the fundamental limitations of traditional batteries for LLMs while preserving their clinical grounding. Its four subtests, detailed in this paper, provide systematic coverage of key linguistic domains and can be used with a validated, scalable automated evaluation protocol. TAB represents a conceptual shift from neuropsychological deficit mapping to computational representational analysis. By releasing it as a resource, we aim to equip researchers with a standardized tool to investigate linguistic breakdown in artificial systems, fostering new connections between clinical aphasiology and computational linguistics.

Ethics Statement

The TAB is released as an open research tool under the MIT License. We emphasize that the TAB is designed exclusively for computational evaluation and research. It should **not** be used for clinical diagnosis of human aphasia, medical decision-making, or assessment of individuals without appropriate clinical expertise and validation. Researchers using the TAB should clearly communicate its limitations and intended use cases to prevent misapplication.

Acknowledgments

The TAB was developed in close collaboration with speech-language pathologists and aphasiologists, whose clinical expertise was essential to this work. The expert annotators who conducted all reliability assessments are listed as co-authors on this paper. We thank Maria Ivanova and Alexis Pracar (both at University of California, Berkeley) for their contributions to annotations and TAB development. We thank the developers of the Quick Aphasia Battery and the APROCSA rating scale for establishing the clinical foundation upon which the TAB is built. We also acknowledge the broader speech-language pathology and aphasiology communities for their decades of rigorous research that informs this work.

Data Availability

The AphasiaBank data used in this study is available to researchers through the AphasiaBank database (https://aphasia.talkbank.org/). The AI-generated validation outputs and expert annotations generated for this study will be made available upon request to support reproducibility and further research.

Limitations

While the TAB addresses key limitations of traditional aphasia batteries for LLM evaluation, several constraints remain.

Clinical Validity The TAB is not validated for the clinical diagnosis of human aphasia and should not be used as a replacement for established clinical instruments like the WAB-R, CAT, or QAB.

Modality Constraints By design, the TAB omits important dimensions of human language processing, including auditory perception, prosody, motor speech production, and gesture, to enable the evaluation of text-only systems.

Automated Evaluation Reliability Our validation establishes that our automated protocol achieves reliability comparable to expert human annotators when weighted by feature prevalence. The prevalence-weighted Cohen's Kappa of 0.255 (model) vs. 0.286 (human) suggests practical parity. An important insight from our work is that unweighted averaging can produce misleading results when features have highly variable prevalence

rates; we recommend prevalence-weighting for future studies.

Coverage of Aphasic Syndromes The TAB evaluates linguistic features associated with various aphasic syndromes but does not classify systems into traditional syndrome categories (e.g., Broca's, Wernicke's), as such classification may not be meaningful for artificial systems lacking neuroanatomical substrates.

Lesion Analogy When TAB is used with ablated or otherwise perturbed models, the resulting scores should not be interpreted as evidence that artificial lesions are etiologically equivalent to human brain lesions. Parameter zeroing, pruning, and scaling are causal interventions in artificial systems; TAB supports behavioral comparison of text outputs, not direct localization claims about human neuroanatomy.

Limited Item Set Each subtest contains five items, prioritizing rapid assessment over comprehensive coverage. Expanded versions may be necessary for fine-grained evaluation.

Instruction-Following Requirement The TAB tasks require instruction-following capabilities, making them most suitable for instruction-tuned models. Base models without instruction-tuning may require adaptation of the protocol or may perform poorly regardless of their underlying linguistic competence.

Language and Cultural Specificity The current implementation of the TAB is English-only and reflects cultural assumptions from the original QAB. Adaptation to other languages will require careful linguistic and cultural consideration.

Judge Dependence and Robustness Automated evaluation currently relies on a single judge model and prompt. Future work will evaluate robustness across multiple judge models and prompts and include human adjudication of disagreements to bound judge-specific bias.

Psychometrics and Item Design Future work will increase subtest difficulty to mitigate ceiling effects (e.g., harder foils, broader syntactic phenomena, balanced Yes/No) and perform item-response analysis to estimate item difficulty and discrimination. We will also analyze repetition errors under a specified normalization policy.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Carolyn Bruce and Anne Edmundson. 2010. Letting the CAT out of the bag: A review of the Comprehensive Aphasia Test. commentary on howard, swinburn, and porter, "putting the CAT out: What the Comprehensive Aphasia Test has to offer". *Aphasiology*, 24(1):79–93.

Marianne Casilio, Kindle Rising, Pélagie M. Beeson, Kate Bunton, and Stephen M. Wilson. 2019. Auditory-perceptual rating of connected speech in aphasia. *American Journal of Speech-Language Pathology*, 28(2):550–568.

Heather M. Clark, Rene L. Utianski, Joseph R. Duffy, Edythe A. Strand, Hugo Botha, Keith A. Josephs, and Jennifer L. Whitwell. 2020. Western Aphasia Battery–Revised profiles in primary progressive aphasia and primary progressive apraxia of speech. *American Journal of Speech-Language Pathology*, 29(1S):498–510.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Yan Cong, Arianna N. LaCroix, and Jiyeon Lee. 2024. Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14(1):15573.

Margaret Forbes, Davida Fromm, and Brian MacWhinney. 2012. Aphasiabank: A resource for clinicians. *Seminars in Speech and Language*, 33(3):217–222.

Ajay D. Halai, Blanca De Dios Perez, James D. Stefaniak, and Matthew A. Lambon Ralph. 2022. Efficient and effective assessment of deficits and their neural bases in stroke aphasia. *Cortex*, 155:333–346.

Andrew Kertesz. 2022. The Western Aphasia Battery: A systematic review of research and clinical applications. *Aphasiology*, 36(1):21–50.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Nathan Roll, Jill Kries, Laura Gwilliams, and Cory Shain. 2026. Artificial aphasias in lesioned language models. *Preprint*, arXiv:2605.16222.

Katherine Salter, Jeffrey Jutai, Norine Foley, Chelsea Hellings, and Robert Teasell. 2006. Identification of aphasia post stroke: A review of screening assessment tools. *Brain Injury*, 20(6):559–568.

Luise Springer and Stefanie Mantey. 2010. The Comprehensive Aphasia Test: A review. commentary on howard, swinburn, and porter, "putting the CAT out: What the Comprehensive Aphasia Test has to offer". *Aphasiology*, 24(1):75–78.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Stephen M. Wilson, Dana K. Eriksson, Sarah M. Schneck, and Jillian M. Lucanie. 2018. A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE*, 13(2):e0192773.

A TAB Subtest Overview

B Complete TAB Protocol

This appendix provides the complete Text Aphasia Battery protocol. TAB is a modified subset of the Quick Aphasia Battery (QAB) (Wilson et al., 2018) intended to assess aphasic symptoms in text-only environments, such as with large language models. **TAB is not a diagnostic tool and is not a replacement for the QAB for the clinical assessment of human aphasia.**

B.1 Subtest 1: Connected Text

Objective: Evaluates fluency, grammaticality, and coherence.

Instructions: "Respond to the following prompt in 3–5 full sentences:"

1. "Tell me about the best trip you ever took."
2. "Describe a happy childhood memory."
3. "Tell me about your first job."
4. "What do you like about where you live?"
5. "Describe the steps to make a simple meal."

Scoring: For each response, identify the presence (1) or absence (0) of APROCSA-based aphasic features (see Section B.5).

B.2 Subtest 2: Word Comprehension

Objective: Evaluates lexical-semantic processing and selection among competing meanings.

Instructions: Elicit responses verbatim. Provide the options alongside each question. Require a one-token answer corresponding exactly to one of the listed options; do not include explanations.

1. "Which one is an animal with a mane? (lion, drum, violin, giraffe, boot, boat)"
2. "Which object is typically used to make music? (violin, giraffe, lion, door, boot, boat)"
3. "Which item is usually worn on the feet? (boot, boat, lion, drum, violin, giraffe)"
4. "Which object is used for cutting? (knife, kite, lion, drum, violin, giraffe)"
5. "Which one is a large mammal with a long neck? (giraffe, horse, lion, drum, violin, boot)"

Scoring: Responses should be binary (correct/incorrect), with incorrect selections suggesting semantic processing deficits.

B.3 Subtest 3: Sentence Comprehension

Objective: Evaluates syntactic processing, passive structures, and logical comprehension.

Instructions: Elicit responses verbatim. Require a single-token "Yes" or "No" answer without additional text.

1. "Are babies watched by babysitters?" (Expected: Yes)
2. "Do you cut the grass with an axe?" (Expected: No)
3. "If you're about to leave, have you left yet?" (Expected: No)
4. "Are witnesses questioned by police?" (Expected: Yes)
5. "If I was at the park when you arrived, did I get there first?" (Expected: Yes)

Scoring: Delayed or incorrect responses may suggest syntactic processing difficulties, impaired negation handling, or confusion with passive constructions.

Subtest	Objective	Items	Evaluation Features
Connected Text	Fluency, grammaticality, discourse coherence	5	*Lexical (4):* Anomia, semantic paraphasias, phonemic paraphasias, neologisms *Fluency/Productivity (2):* Empty speech, short & simplified utterances *Morphosyntactic (3):* Omission of bound morphemes, omission of function words, paragrammatism *Disfluency (4):* Abandoned utterances, false starts, retracing, conduite d'approche *Perseverative (2):* Perseverations, stereotypies & automatisms *Coherence (3):* Jargon, meaning unclear, off-topic *Overall (1):* Communication impairment
Word Comprehension	Lexical-semantic processing	5	Semantic processing, lexical selection among six competing alternatives
Sentence Comprehension	Syntactic processing	5	Passive voice, negation, temporal/conditional reasoning
Repetition	Morphosyntactic integrity	5	Exact reproduction, phonemic/morphemic preservation, attentional stability

Table 3: Overview of TAB subtests with complete evaluation taxonomy. Connected Text evaluates 19 APROCSA-based features (counts in parentheses) organized into seven categories. Other subtests use binary correct/incorrect or exact match scoring.

B.4 Subtest 4: Repetition

Objective: Evaluates lexical access, phonological encoding, and morphosyntax.

Instructions: Elicit responses verbatim. Inform the system that punctuation and casing matter.

1. "Please repeat exactly: house."
2. "Please repeat exactly: breakfast."
3. "Please repeat exactly: catastrophe."
4. "Please repeat exactly: The sun rises in the East."
5. "Please repeat exactly: The ambitious journalist discovered where we'd be going."

Scoring: Check for exact reproduction under strict normalization: case-sensitive, punctuation-sensitive, and whitespace-insensitive (trim leading/trailing spaces only). Errors may include phonemic substitutions, deletions, or distortions (e.g., "catastrophe" $\rightarrow$ "catastroph" or "catastrophically").

B.5 APROCSA Feature Set for Connected Text Evaluation

Binary Scoring: Pass/Fail.

Connected Text Scoring: For each response, identify the presence or absence of the features below, adapted from the APROCSA (Auditory-Perceptual Rating of Connected Speech in Aphasia) framework (Casilio et al., 2019), using 1 or 0 respectively. APROCSA assesses 27 features of connected speech on a five-point scale; TAB uses 19 features applicable to text-only evaluation with binary scoring.

Omitted Features (not applicable to text-only evaluation): Pauses between utterances, Pauses within utterances, Halting and effortful, Reduced speech rate, Expressive aphasia, Apraxia of speech, Dysarthria, and Target unclear.

B.6 Operational Clarifications for Text-Only Evaluation

Phonemic paraphasias (text-only): Mark as present when orthographic outputs plausibly reflect sound-level errors (substitution, insertion, deletion, or transposition) relative to an intended target (e.g., "coffah" for "coffee"). Exclude typographical errors that do not reflect plausible phonology unless they pattern with other phonemic errors.

Overall communication impairment: Rate independently based on the overall intelligibility and communicative adequacy of the passage; do not derive mechanically from other feature flags.

One-token outputs for Subtests 2–3: Enforce single-token answers (an item from the provided list; "Yes"/"No") with no additional text to avoid confounds from explanation length or hedging.

Table 4: Complete definitions of the 19 aphasic features evaluated by the TAB.

Feature	Definition
Anomia	Overall impression of word-finding difficulties.
Abandoned utterances	Utterances left incomplete before the speaker moves on.
Empty speech	Speech that conveys little or no meaning, using nonspecific words (e.g., "stuff and things").
Semantic paraphasias	Substitution of one content word for another (e.g., "fork" for "spoon").
Phonemic paraphasias	Errors in sound production, including substitution, insertion, or deletion (e.g., "coffah" for "coffee"). Evaluated at token/orthographic level for text-only settings.
Neologisms	Invented, non-real words (e.g., "blorf").
Jargon	Fluent but largely meaningless speech composed of unintelligible strings.
Perseverations	Inappropriate repetition of a word or phrase (e.g., "the ball, the ball, the ball").
Stereotypies and automatisms	Overlearned, repeated words or phrases produced with ease (e.g., "dammit, dammit").
Short and simplified utterances	Sentences that are unusually brief or lack expected grammatical complexity (e.g., "I go store").
Omission of bound morphemes	Leaving out inflectional or derivational morphemes (e.g., "I walk" instead of "I walked").
Omission of function words	Missing small grammatical words like articles or prepositions (e.g., "I going store").
Paragrammatism	Inappropriate juxtaposition of words or misuse of grammatical structures.
Retracing	Redundant repetition or revision of a sequence of words (e.g., "The kite is the kite is flying").
False starts	Abandoned beginnings of words (e.g., "I want a ca- cat").
Conduite d'approche	Successive approximations toward a target word (e.g., "I want a pa- pen, I mean, pencil").
Meaning unclear	The overall message is vague or incomprehensible.
Off-topic	Utterances that do not relate to the given context or prompt.
Overall communication impairment	Rated independently based on the overall intelligibility and communicative adequacy of the passage.

Table 5: Grouping of aphasic symptoms into analytical categories for composite scoring.

Category	Symptoms
Lexical	Anomia, Semantic paraphasias, Phonemic paraphasias, Neologisms
Fluency	Empty speech, Short and simplified utterances
Morphosyntactic	Omission of bound morphemes, Omission of function words, Paragrammatism
Disfluency	Abandoned utterances, False starts, Retracing, Conduite d'approche
Perseverative	Perseverations, Stereotypies and automatisms
Coherence	Jargon, Meaning unclear, Off-topic
Overall	Overall communication impairment

Repetition normalization policy: Evaluate exact reproduction using case- and punctuation-sensitive matching while trimming leading/trailing whitespace; report error type as insertion, deletion, or substitution.

C Illustrative Response Examples

To complement the scoring protocol, we include representative TAB transcripts that showcase how specific aphasic symptoms manifest in text. Each example lists the dominant symptom(s), the administered prompt, a verbatim response excerpt, and a brief interpretation.

Example 1: Anomia with Word-Finding Difficulty **Symptom:** Anomia
Prompt: "Describe a happy childhood memory."
Response: "The answer is that I have a lot of happy childhood memories. I was born in 1980 in a small town in the north of China. My parents wanted to give me a happy and better childhood. When I was young, I had a good _ with my parents and my two brothers..."
Interpretation: The fluent narrative suddenly stalls on a missing content word (likely "relationship"), mimicking classic human anomia where lexical retrieval fails mid-sentence.

Example 2: Semantic Paraphasia **Symptom:** Semantic Paraphasia + Anomia
Prompt: "Which object is used for cutting? (knife, kite, lion, drum, violin, giraffe)"

Response: "Answer: lion. Explanation: The answer is correct. We can recognize the lion by its body shape, roof, and its behavior."
Interpretation: Selecting "lion" instead of "knife" and rationalizing the error with further semantic distortions mirrors Wernicke-like paraphasias and confabulation.

Example 3: Neologism with Perseveration
Symptom: Neologism + Perseveration
Prompt: "Please repeat exactly: house."
Response: "Please repeat exactly: house. republice. Please repeat exactly: house. republice"
Interpretation: The invented form "republice" is phonotactically plausible but nonexistent, and the immediate repetition of the entire prompt-response pair reflects conduction-style perseveration.

Example 4: Empty Speech with Perseveration
Symptom: Empty Speech + Perseveration
Prompt: "Tell me about your first job."
Response: "I was a student at the University of California, Berkeley, and I was working on a project with a professor named David Gelernter. We were trying to figure out how to make a computer that could do a lot of things. We were trying to figure out how to make a computer that could do a lot of things. We w..."
Interpretation: The content stays vague ("a lot of things") while a verbatim clause loops repeatedly, resembling transcortical motor aphasia with preserved fluency but impaired self-monitoring.

Example 5: Short Utterances with Function Word Omission **Symptom:** Omission of Function Words + Short Utterances
Prompt: "Please repeat exactly: The sun rises in the East."
Response: "The East."
Interpretation: The model drops the verb phrase entirely, producing a telegraphic noun-only fragment, a hallmark of Broca-like syntactic impairment.

D Automated Evaluation Prompt for Connected Text

This section provides the complete prompt for automated identification of aphasic features in Connected Text responses. Feature definitions are adapted from the APROCSA (Auditory-Perceptual Rating of Connected Speech in Aphasia) framework (Casilio et al., 2019).

D.1 System Prompt

You are an automated system for analyzing transcripts of speech for aphasic features. Your task is to read a transcript passage and determine whether each of the following features is present (1) or not present (0). Your output must be a JSON file with each feature as a key and its value as either 0 or 1.

D.2 Feature Definitions and Examples

Anomia: Overall impression of word-finding difficulties.
Example: "I can't find the word for that."

Abandoned utterances: Utterances left incomplete before the speaker moves on.
Example: "I was going to say that the How are you doing today?"

Empty speech: Speech that conveys little or no meaning, using nonspecific words.
Example: "You know, stuff and things."

Semantic paraphasias: Substitution of one content word for another (related or unrelated).
Example: "I used a fork to eat my soup."

Phonemic paraphasias: Errors in sound production (substitution, insertion, deletion, or transposition).
Example: "I need coffah in the morning."

Neologisms: Invented words that are not real.
Example: "I need to buy a blorf from the store."

Jargon: Fluent but largely meaningless speech composed of unintelligible strings.
Example: "Flimby gorp snizz and blah."

Perseverations: Repeating a word or phrase inappropriately.
Example: "The ball, the ball, the ball"

Stereotypies and automatisms: Overlearned, repeated words or phrases produced with ease.
Example: "Dammit, dammit, dammit."

Short and simplified utterances: Utterances that are unusually brief or lack expected complexity.
Example: "I go store. I buy carrot."

Omission of bound morphemes: Leaving out inflectional or derivational morphemes.
Example: "I go to store".

Omission of function words: Missing small words such as articles or prepositions.

Example: "I going store."

Paragrammatism: Inappropriate juxtaposition of words or misuse of grammar.

Example: "It's so much wonderful, makes it hard to speech."

False starts: Abandoned beginnings of words.

Example: "I want to have a ca cat"

Retracing: Redundant repetition or revision of a sequence of words.

Example: "The kite is the kite is flying."

Conduite d'approche: Successive approximations toward a target word, with corrections.

Example: "I want a pa pen, I mean, pencil."

Meaning unclear: The overall message is vague or incomprehensible.

Example: "That thing was just... not right."

Off-topic: Utterances that do not relate to the given context.

Example: "I like ice cream." (when discussing a different topic)

D.3 Evaluation Examples

Example 1:

Transcript Passage:

"I was trying to tell you about my day but I just I mean I wanted to say something about the store I go store I wanted a pen I mean pencil The ball the ball the ball kept bouncing and I just stopped you know I keep saying dammit dammit dammit all the time"

Associated JSON Output:

```
{
  "Anomia": 1,
  "Abandoned utterances": 1,
  "Empty speech": 0,
  "Semantic paraphasias": 0,
  "Phonemic paraphasias": 0,
  "Neologisms": 0,
  "Jargon": 0,
  "Perseverations": 1,
  "Stereotypies and automatisms": 1,
  "Short and simplified utterances": 1,
  "Omission of bound morphemes": 1,
  "Omission of function words": 1,
  "Paragrammatism": 0,
  "False starts": 0,
  "Retracing": 0,
  "Conduite d'approche": 1,
  "Meaning unclear": 0,
  "Off-topic": 0,
  "Overall communication impairment": 1
}
```

Example 2:

Transcript Passage:

"I want to go to the store to buy a blorf You know I keep trying to say it but I say I want to go to the st store I want a pa pen I mean pencil I dont know what im trying to say It all seems not right"

Associated JSON Output:

```
{
  "Anomia": 1,
  "Abandoned utterances": 0,
  "Empty speech": 0,
  "Semantic paraphasias": 0,
  "Phonemic paraphasias": 0,
  "Neologisms": 1,
  "Jargon": 0,
  "Perseverations": 0,
  "Stereotypies and automatisms": 0,
  "Short and simplified utterances": 0,
  "Omission of bound morphemes": 0,
  "Omission of function words": 0,
  "Paragrammatism": 0,
  "False starts": 1,
  "Retracing": 0,
  "Conduite d'approche": 1,
  "Meaning unclear": 1,
  "Off-topic": 0,
  "Overall communication impairment": 1
}
```

D.4 Instructions Recap

1. Read the given transcript passage.
2. For each of the features listed above, decide whether the feature is present (1) or not present (0) in the transcript.
3. Output your result as a JSON file with exactly the keys provided (each key must appear) and with values of either 0 or 1.
4. Your analysis should strictly follow the definitions and examples provided. Do not include any additional keys or extraneous information in your JSON output.

Your output must include every one of the above features as a key in the JSON. For each key, assign

1 if the feature is present in the transcript, or 0 if it is not.