

---

# Artificial Aphasias in Lesioned Language Models

---

Nathan Roll<sup>1</sup> Jill Kries<sup>2,3</sup> Laura Gwilliams<sup>2,3</sup> Cory Shain<sup>1</sup>

Department of {<sup>1</sup>Linguistics, <sup>2</sup>Psychology}, Stanford University

<sup>3</sup>Wu Tsai Neurosciences Institute, Stanford University

nroll@stanford.edu

## Abstract

*Aphasias*, selective language impairments which can arise from brain damage, reveal the functional organization of human language by providing causal links between affected brain regions and specific symptom profiles. Drawing on this literature, we introduce an aphasia-inspired technique to characterize the emergent functional organization of language models (LMs). We “lesion” (zero-out) model parameters and measure the effects of this intervention against clinical aphasia symptoms, as diagnosed by the Text Aphasia Battery (TAB). When applied to 112,426 outputs from five 1B-scale LMs, the full range of evaluated symptoms surface, but in distributions largely distinct from those of humans. Our method uncovers broad symptom-profile differences between attention components (query, key, value, output) and feed-forward components (up, gate, down), with weaker evidence for differences among components within the same mechanism. We also find an effect of depth, where lesions in early layers disproportionately cause syntactic and semantic symptoms while late-middle layers yield higher rates of phonological and fluency deficits. Although some LM lesions induce quantitatively more similar profiles to some human aphasia types than others, qualitative differences in symptom patterns between LMs and humans suggest that aphasia syndromes are heavily influenced by the details of learning and processing rather than being a domain-invariant consequence of disrupted language processing.

arXiv:2605.16222v1 [cs.CL] 15 May 2026

# 1 Introduction

For over a century, neuropsychologists have relied on naturally-occurring cases of brain damage to understand the brain’s functional organization [Broca, 1861, Wernicke, 1874], including language-selective deficits or *aphasias*. Neuropsychology has now provided a rich taxonomy of aphasia subtypes [Goodglass, 1993] and associated theories of brain function [Dronkers, 1996, Dronkers et al., 2004, 2017, Hickok and Poeppel, 2007, Friederici, 2011, Fedorenko et al., 2024], along with standardized assessments for a range of diagnostic symptoms [Kertesz, 2007, Goodglass et al., 2001, Swinburn et al., 2005, Wilson et al., 2018]. At the same time, researchers in natural language processing (NLP) and artificial intelligence (AI) increasingly intervene on language models (LMs)—including silencing weights and activations in ways reminiscent of brain lesions [LeCun et al., 1990, Michel et al., 2019, Voita et al., 2019, Meng et al., 2022, Geiger et al., 2021, Wang et al., 2025, 2026]—in order to study emergent structure-function correspondences in LMs. This effort falls under the broader program of *mechanistic interpretability* [Elhage et al., 2021, Olsson et al., 2022, Nanda et al., 2023, Conmy et al., 2023]. These two communities (neuropsychology and mechanistic interpretability) have to-date worked largely independently (though c.f., e.g., Wang et al., 2025, 2026). In this study, we show that both stand to gain by joining forces: neuropsychological methods that were developed for human aphasia reveal reliable and in some ways surprising emergent functional organization in LMs, and artificial LM “aphasias” constrain the hypothesis space about why human aphasia symptoms pattern as they do.

In brief, we induce artificial “aphasias” in LMs by “lesioning” (zeroing out) portions of key parameter matrices of a variety of open-weight LMs based on the Transformer architecture Vaswani et al. [2017]. We then produce a detailed symptom profile for each lesioned model by generating from it and passing the outputs through the recently-proposed Text Aphasia Battery [TAB; Roll et al., 2025]. Unlike traditional aphasia batteries such as the Western Aphasia Battery–Revised [WAB-R; Kertesz, 2007] and Boston Diagnostic Aphasia Examination [BDAE; Goodglass et al., 2001], the TAB is specifically engineered to apply to the text-only modality of decoder-only language models (LMs). Here, TAB symptom labels serve as a shared measurement system to describe model failures at a finer linguistic grain than e.g., benchmark accuracy [Wang et al., 2018, Warstadt et al., 2020, Gauthier et al., 2020] and can be directly compared to human productions from AphasiaBank [MacWhinney et al., 2011].

This approach respects three key desiderata for interpreting LM functional organization through the lens of aphasia. *First*, because human aphasia subtypes are not natural kinds [Wilson et al., 2023], models should not merely be compared to coarse categorical labels that can mask substantial heterogeneity (“Broca’s”, “Wernicke’s”, etc.), but should instead have their symptoms profiled in detail. *Second*, analyses of lesioned LMs should be sensitive not only to similarities to human aphasia, but also to differences. *Third*, analyses should prioritize functions that can be studied in text-only models, which remain the majority case in today’s technological landscape and thus the most generally relevant to the mechanistic interpretability community. These traits collectively distinguish our study from (and allow us to critically evaluate) related work claiming that lesioned LMs replicate aphasia subtypes [Wang et al., 2025, 2026].

We use this setup to build a causal behavioral map of lesion-induced failures in LMs. We zero-ablate seven *components* (weight matrices, i.e., Q, K, V, O, Gate, Up, Down) across layers at five severity levels in Llama 3.2, Gemma 3, and OLMo 2 model variants [Meta, 2024, Gemma Team et al., 2025, Team OLMo et al., 2025], producing 112,426 severity > 0 scored lesion records. The main output of our method is the *failure profile* for a given lesion, that is, the TAB-labeled symptom distribution induced by lesioning a given component at a given layer and severity. For comparison, we also score 6,000 AphasiaBank responses in the same label space. We find that attention and feed-forward-network (FFN) damage systematically produce different failure profiles across LMs and that those profiles vary with depth, with semantic and syntactic deficits reliably and surprisingly localized to *earlier* layers than phonological and fluency deficits. Moreover, we critically replicate prior reports that LM lesions convergently reproduce some human-like aphasias [Wang et al., 2025, 2026]. In particular, although we show in line with these reports that some LM lesions produce symptoms that look reliably more like some coarse human aphasia subtypes than others, human and LM lesions differ dramatically both in the composition of symptoms and in overall symptom burden, underscoring the importance of tracking detailed symptom profiles rather than merely mapping lesioned LMs to coarse aphasia category labels. The latter outcome indicates

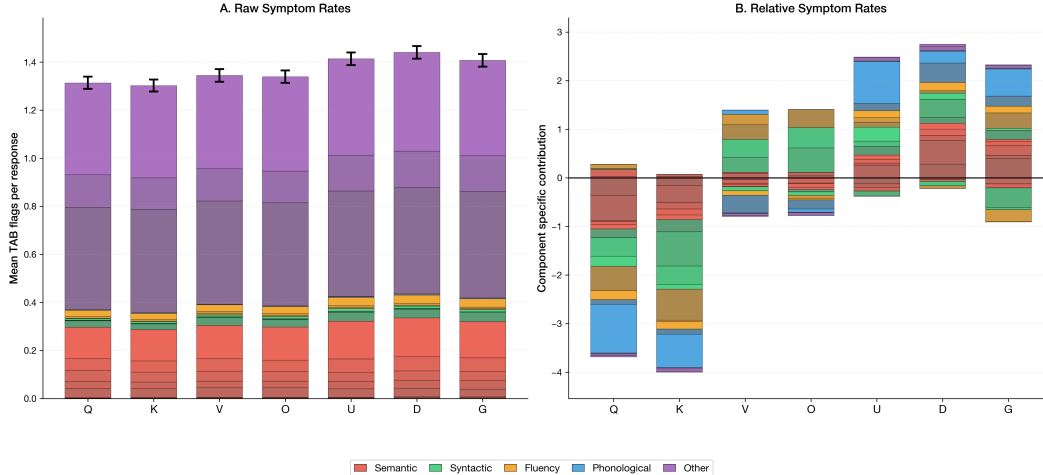


Figure 1: **Component lesions produce different TAB-symptom mixtures.** (A) Raw stacked TAB-symptom rates per component, colored by symptom class and shaded by specific symptom; bars can exceed 1 because each response can receive multiple symptoms. Error bars show 95% CIs for total symptom burden. (B) Component-specific contribution, computed as each component’s per-symptom rate as a multiple of the all-output ablated mean for that symptom, minus one. Values above 0 mark symptoms that are overrepresented for that component; values below 1 mark symptoms that are underrepresented.

that human aphasia symptoms are not inherent to the structure of language, as might be suggested by prior reports of human-LM convergence. Instead, our results are more consistent with the view that aphasia symptoms reflect contingent properties of the functional organization of the human brain.

We therefore make four key contributions: we (1) offer a proof of concept for an aphasia-inspired “neuropsychology” of LMs, (2) identify reliable functional differences between the attention and FFN mechanisms across LMs, (3) reveal (surprisingly) that semantic and syntactic deficits reliably precede phonological and fluency deficits as a function of depth, and (4) nuance prior reports of emergent alignment between human and LM aphasias.

## 2 Related work

Classic studies in connectionism used ablations to study neural network models of language. For example, [Hinton and Shallice \[1991\]](#) reproduced deep-dyslexia-like errors in attractor networks, and [Plaut and Shallice \[1993\]](#) showed that graded damage can produce crossed deficits without separate modules. These studies used small networks engineered to model specific human targets. We bring the same intervention logic to general-purpose pretrained transformers, where components can be systematically lesioned and scored in a common framework (TAB).

A second line of work uses *encoding* approaches to predict brain activity from LM hidden states [e.g., [Schrimpf et al., 2021](#)], typically in healthy individuals. This work is inherently correlational and thus of limited relevance to our target application (aphasia-inspired mechanistic interpretability).

A third line of work attempts to interpret LM components in linguistic terms, for example by linking attention heads to syntactic behavior [[Clark et al., 2019](#)] or feed-forward layers to lexical and conceptual information [[Geva et al., 2022](#)]. Other causal-abstraction and circuit methods test whether proposed internal mechanisms are behaviorally necessary [[Geiger et al., 2021](#), [Chan et al., 2022](#)]. Our work is closer to this causal tradition than to post-hoc feature attribution: we intervene on model components and measure the resulting behavioral distribution. At the same time, lesion experiments inherit a version of the out-of-distribution concern raised for feature-removal explanations by [Hase et al. \[2021\]](#): removing part of an input or model can create states unlike those seen during training. We therefore interpret a lesion profile as a controlled stress test of model behavior, not as the normal function of a component in isolation.

The most directly related prior studies are recent attempts to induce human aphasia categories in LMs through lesioning components or neurons [Wang et al., 2025, 2026]. We treat these studies as important evidence that aphasia-inspired interventions can produce non-random, clinically legible LM failures. Our test is stricter: we ask whether those failures organize like human aphasia when humans and LMs are scored in the same symptom space. Our results preserve part of the convergence story—some LM lesions are descriptively closer to some human diagnosis groups than others—while showing that coarse category matches can obscure large differences in the overall rate of symptoms (burden), proportion of symptoms (mixture), and co-occurrence tendencies between symptoms.

### 3 Methods

Section 3.1 defines the interventions and data, Section 3.2 defines the scoring rubric, and Section 3.3 links the statistical tests to these questions. TAB is the behavioral readout throughout.

#### 3.1 Experimental setup

The full sweep covers five  $\sim$ 1B-parameter decoder-only model variants across three families: Llama 3.2-1B and its instruction-tuned variant, Gemma 3-1B-Instruct, and OLMo-2-1B with its instruction-tuned variant. Layer counts differ by family (Gemma: 26; Llama and OLMo: 16), as do head counts (OLMo: 16; Gemma: 8; Llama: 32).

Each lesion targets one component at a given layer and severity. For a target component  $W$ , we define the ablated weights  $\widetilde{W}$  by an element-wise mask:

$$\widetilde{W} = W \odot M, \quad M_{ij} \sim \text{Bernoulli}(1 - s),$$

where  $s \in \{0, 0.25, 0.50, 0.75, 1.0\}$  is the lesion severity and  $\odot$  denotes the Hadamard product. Thus  $s$  is the expected fraction of weights zeroed, while  $1 - s$  is the retention probability. Although concerns about zeroing out as an LM intervention have been raised in prior work [Hase et al., 2021, Chan et al., 2022], this is not critical to our design, and we obtain qualitatively similar results when instead setting weights to their empirical means (Appendix D.1). We deliberately do not rescale retained weights by  $1/(1-s)$  because the intervention is a lesion rather than a dropout-style unbiased estimator: rescaling would preserve expected projection magnitude while amplifying the surviving weights. Because Q/K masks can also change attention-logit scale and softmax entropy, we use zero-ablations as a controlled component-damage experiment, which is more extreme but similar in distribution to replacement-value and perturbation-proxy controls (Appendices D.1 and D).

The seven components are the Q(uey), K(ey), V(alue), and O(utput) matrices of the attention mechanism and the G(ate), U(p), and D(own) matrices of the feedforward mechanism. The main analyses use a fixed 20-prompt TAB subset spanning connected text, word comprehension, sentence comprehension, and repetition (Appendix J); the full set of TAB prompts is in Roll et al. [2025]. When severity  $0 < s < 1$ , one Bernoulli mask is sampled for each condition; otherwise, the mask is deterministic. Appendix M reports comparisons across random seeds at  $s = 0.75$ ; results did not depend critically on the choice of seed.

Unless otherwise stated, analyses use deterministic greedy decoding in FP16 precision, with one beam, no sampling, no repetition penalty, and a maximum of 80 generated tokens. This gives a deterministic condition-level baseline. In a targeted comparison to nucleus-sampling using Gemma-3-1B-IT K/Gate/Up lesions (layers 5/10/15/20/25, severities 75/100;  $T = 0.7$ ,  $p = 0.9$ , repetition penalty 1.2), FFN-Gate and FFN-Up profiles remain qualitatively similar to their greedy profiles.

We compare lesioned LMs to 6,000 productions from AphasiaBank [MacWhinney et al., 2011] (licensed under TalkBank): 1,000 per diagnosis group for people with aphasia (PWA; Anomic, Broca’s, Conduction, Transcortical Motor, Wernicke’s) plus 1,000 control texts, drawn from 30 text corpora (see Appendix A.2 for the full list). Although our main analyses compare lesioned LMs to a range of human productions, we obtain similar results when comparing only to narratives produced by PWA (Appendix N).

The full analysis comprises 112,426 scored records from 2,528 LM lesions; “records” can include repeated generations later aggregated by model, layer, component, severity, and prompt. After prompt

deduplication, the primary condition-level analyses use 44,527 records from LMs with lesion severity  $> 0$ . The table of LM symptom rates used to compare LMs to humans contains 55,279 LM records including intact baselines.

### 3.2 Assessment

We score lesioned LMs with the previously released TAB [Roll et al., 2025], a text-native benchmark, initially adapted from the Quick Aphasia Battery [Wilson et al., 2018]. TAB scores text behavior alone and provides behavioral descriptions for text-only outputs.

The battery has four subtests: connected text generation, word comprehension, sentence comprehension, and repetition. For a given production, the TAB assigns 21 binary (yes/no) symptoms across five categories (semantic, syntactic, fluency, phonological, other). Human-vs-LM prevalence, co-occurrence, and scorer robustness analyses use 19 of these symptoms, excluding text-prosody pause symptoms because they are noisy to retrieve in text-only setups. Symptom classification is performed by Gemini 2.5 Flash [Comanici et al., 2025] at temperature 0.01 with a hashed response cache. Concrete output examples in Appendix Table 1 show how the TAB-generated symptom profile distinguishes repetition loops, prompt drift, jargon-like strings, formulaic output, and semantic drift. Roll et al. [2025] validated this Gemini-scored TAB framework against scores from human experts on 561 texts drawn from both humans and LMs.

### 3.3 Statistical methods

Our analyses quantify the behavioral consequences of lesions in language models and compare those consequences both to those of other lesion types and to productions by humans. We define a *symptom* as a positive value in one TAB column for one response. From this, we derive our foundational metrics: a *symptom rate* is the fraction of responses in a condition with that symptom, and a *symptom profile* is the vector of symptom rates across TAB columns. A *lesion condition* is one model, layer, component, and severity setting.

All lesioned LM outputs are expressed in decimal symptom-rate units, allowing us to look at aggregate distributions rather than individual one-hot arrays. To ensure tight experimental control, we only compare components or conditions (e.g., FFN vs. attention) when they are strictly matched on all other variables (model, layer, and severity).

Our overarching statistical framework relies on nonparametric methods to assess how these symptom profiles shift. We resample the model  $\times$  layer  $\times$  severity variables with replacement 5,000 times to obtain confidence intervals for profile distances and related statistics. Where appropriate, we supplement these with paired sign-flip tests to check alignment across matched conditions, and permutation tests to establish null distributions (e.g., permuting component labels or category names). We rely on the corpus/model-clustered bootstrap in Appendix C as the dependence-aware uncertainty check for human-vs-LM contrasts.

## 4 Results

### 4.1 Attention and feed-forward lesions cause distinct symptom profiles

We first asked whether lesions to different components simply make the model worse by different amounts. They do not. As shown in Figure 1A, FFN lesions produce a slightly higher symptom burden than attention lesions: at least one TAB symptom appears in 59.4% of FFN-lesioned responses and 56.8% of attention-lesioned responses, with means of 1.445 and 1.350 symptoms per response. This difference is real but small.

The larger effect is in the distribution of symptoms. Figure 1B compares each component’s symptom rates with the mean symptom rates across all lesioned LM outputs. FFN components are shifted toward *Meaning unclear* (+2.20 points), *Stereotypies and automatisms* (+1.70), *Off-topic* (+1.58), *Perseverations* (+1.23), *Short and simplified utterances* (+0.92), and *Target unclear* (+0.87). FFN lesions are more likely than attention lesions to show symptoms in the Semantic and Other categories, where Other refers to repetition-loop, stereotypy/automatism, and off-topic symptoms.

All other things being equal, the all-symptom profile distance  $\delta$  (between attention vs. FFN lesions) was  $L_2 = 3.73$  percentage points (95% CI [2.45, 5.66], sign-flip  $p = 0.0004$ ), with a restricted component-name permutation given in appendix B. When the component is entirely ablated ( $s = 1$ ) the all-symptom distance rises to 8.05 points (CI [4.45, 14.46],  $p = .018$ ).

In other words, attention and FFN lesions differ systematically not only in their effects on the overall rate of symptoms or the rate of one specific symptom, but in the pattern of covarying symptoms that they evoke. That pattern in turn carries reliable information about whether attention or FFN components were lesioned.

Within the broad attention vs. FFN distinction, there appears to be finer structure that we interpret speculatively. In particular, within the attention component, Q/K and V/O respectively form two distinct cohorts of components with similar behavioral effects (Fig 1B). This mirrors the transformer computation, where Q/K jointly shape which context positions are attended and V/O shape what information from the attended positions is carried forward and written back to the residual stream. In addition, within the FFN mechanism, we find that lesions to D create the highest overall symptom burden, possibly due to D’s role writing FFN updates back to the residual stream. And although this is not clearly evident from the TAB symptoms, qualitative inspection suggests that lesions to G often reduce outputs to punctuation-only or repeated function-word strings. Cosine similarities of these finer component-level effects to human productions are reported Appendix P.

Across multiple supplementary analyses, we confirmed that this TAB-based functional distinction between attention and FFN holds under different analysis choices. It is not explained by FFN lesions merely producing more symptoms overall: the profile distance remains after accounting for total symptom burden and after matching attention and FFN conditions with similar output length or symptom burden. The contrast remains even when repetition-loop, stereotypy, and off-topic symptoms are removed. It is not only a scorer artifact: scoring only based on the most reliable symptoms, using a heuristic (non-LLM) scorer, and using an independent scorer model all recover a coarse attention/FFN contrast, although individual symptom details are scorer-dependent. Finally, it is not obviously driven by one prompt, model, or random seed: the finding recurs across prompts, model variants, severities, and seeds, as well as in (deterministic) total ablations. See Appendix Tables 2, 3, and 38 for details.

Together, these results reveal in fine detail the specific impairments caused by lesions to different LM components. To gloss the differences in TAB symptoms reported above, FFN damage shifts outputs toward vague content, short or formulaic strings, and drift from the context set up by the prompt, where as attention damage disproportionately impairs phonology and fluency. The next question is whether this profile shape is specific to one model family or recurs across architectures and tuning variants.

## 4.2 Similar symptom profiles recur across model architectures

Looking across the model families (Llama, Gemma, and OLMo) of comparable size (1B), we find directionally consistent TAB symptom profiles. Between-family cosine similarities of raw symptom profiles are near ceiling (mean 0.936 for attention and 0.944 for FFN; range 0.91–0.97 across severity levels). Other ways of aggregating paint the same picture: in 77/97 model–prompt pairs, 18/20 prompts, 4/4 nonzero severities, 5/5 model variants, and 3/3 families, the symptom-level direction of difference for FFN vs. attention lesions matches the overall average reported above (Appendix E).

Targeted analyses of larger models suggest more severe interventions are needed in order to evoke measurable TAB symptoms. However, modulo this change, qualitatively similar effects are observed. At 3B (Qwen 2.5-3B-Instruct [Qwen et al., 2025]), targeted auxiliary analyses of hidden-state change and syntax/semantics probes preserve an attention/FFN contrast, while a 70-item syntax/semantics minimal-pair check shows an early-layer Gate vulnerability and K and Up stay near baseline (Appendix S). At 7B, single-layer ablation produces no detectable TAB symptoms, and found minimal results with ablated component pairs. The attention/FFN contrast reappears when multiple layers are ablated simultaneously, where FFN-Gate reaches a repetition-loop symptom rate of 0.41 at 25% severity while attention-K stays near baseline (0.02). For computational reasons, these analyses only manipulate the lesion severity and do not attempt to fully replicate the results from the main study.

In addition to effects of model size, we find that instruction tuning reduces symptom burden by 81%, although the symptom profiles for base and instruct variants of the same model are highly correlated ( $\bar{r} = 0.857$ ; all  $p < 0.001$ ). In other words, instruction-tuning greatly changes the overall rate of symptoms with little effect on the proportion of symptoms produced by lesions to different components.

### 4.3 Layer depth influences symptom profiles

We find that the depth at which a lesion occurs also shapes the symptom profiles. Figure 2 maps TAB-symptom prevalence against normalized layer position (0% = earliest layer, 100% = final layer; severity  $\geq 75\%$ , min-max normalized within rows/symptoms). Grouping symptoms by category labels (Semantic, Syntactic, Phonological, Fluency, Other), we find that within-category correlations between depth and symptom prevalence exceed cross-category correlations ( $r = 0.521$  vs. 0.226,  $\Delta r = 0.295$ ,  $p = 0.0004$ , 5,000 category label permutations). This structure is sharpest for the Other category (within-category  $r = 0.971$ ), intermediate for semantic and syntactic symptoms ( $r = 0.578$  and 0.541), and weakest for fluency ( $r = 0.295$ ).

We further observe two potentially surprising effects of depth. First, several symptoms in the Semantic category, a putatively high-level dimension of linguistic structure (Meaning Unclear, Neologisms, Target Unclear, and Semantic Paraphasias), are concentrated in relatively early layers (depth  $< 50\%$ ). This finding is even more pronounced for three of the Syntactic symptoms (Omission of Bound Morphemes, Omission of Function Words, and Short and Simplified Utterances), which peak strongly in the earliest layers (0%). In contrast, Phonological (Conduite d' Approche, Phonemic Paraphasias) and Fluency (Retracing, False Starts, Halting and Effortful) peak late by comparison (near 60%), even though these features concern relatively low-level aspects of utterance planning. Speculatively, one potential explanation for this asymmetry might be that the later layers of a Transformer network are closer to the output (next-token probability distributions, the closest analog within the model to articulation). It is possible that the network may privilege higher-level semantic and syntactic processing in earlier layers when possible so that later layers, which are downstream, can focus on the mechanics of realizing these features in text.

Second, we find that earlier layers are more necessary for function on average, and that after about 70% of the way through the model, lesions have almost no impact across TAB symptoms. This contrasts with the widespread tendency in e.g., brain decoding studies to find optimal human-LM alignment in mid to late layers [Caucheteux and King, 2022], but this difference is plausibly due to the fact that lesion studies prioritize causal necessity, rather than information content. This result suggests that models may preferentially do the most essential work as early as possible in the forward computation, plausibly to free up later layers for finer-grained refinements to the output.

### 4.4 Clinical reference profiles in TAB space

AphasiaBank productions ( $n = 6,000$ , Gemini scorer) let us compare LM symptom profiles against productions from PWA. Appendix Table 5 averages TAB symptoms within each clinical aphasia category across all scored outputs. The same TAB evaluation is applied to both corpora, but, as we show, the resulting profiles are different between humans and LMs.

We find that human productions are higher on semantic and syntactic TAB symptoms, while lesioned LMs are higher on Other-category symptoms. *Meaning unclear* appears in 45.2% of AphasiaBank productions versus 14.4% of LM outputs, anomia in 22.2% versus 0.3%, and empty speech in 14.1% versus 0.2%. LMs overproduce repetition-loop symptoms (43.4% vs. 10.9%) and off-topic symptoms (39.4% vs. 31.6%). Appendix Figure 5 separates symptom burden from symptom composition and finds clear human-LM differences in both: humans with different aphasia subtypes differ from each other both in total symptom burden and in symptom composition, whereas LM components primarily in burden and only subtly in composition, and humans have systematically higher symptom burdens and different symptom compositions from lesioned LMs. A corpus/model-clustered bootstrap demonstrates the reliability of the direction of difference between humans and LMs for semantic, syntactic, fluency, and phonological categories (Appendix C). A comparison using only narrative productions in Appendix N gives the same qualitative result. Appendix U places human and LM profiles in a single visual frame, and Appendix V shows that even lesioned LMs assign

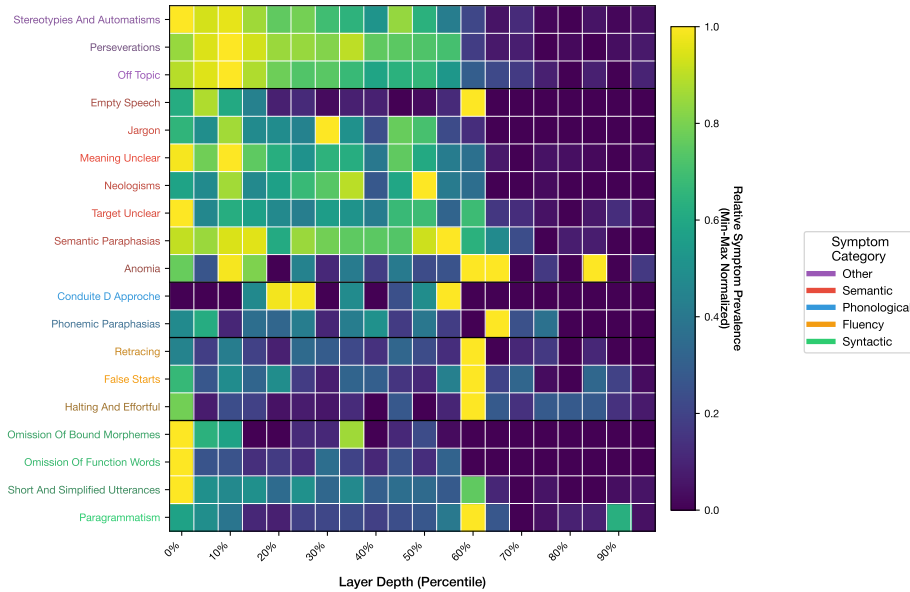


Figure 2: **Depth influences the symptom profile induced by lesions.** Each symptom row is min-max normalized across depth, so color marks where that symptom peaks instead of which symptom is most prevalent.

higher likelihood to productions from the control population than the PWA population, with no change in the strength of this preference as a function of LM lesion condition.

We also compared symptom co-occurrence patterns between humans and LMs. On the 561-text validation subset with direct SLP symptom annotations (306 human, 255 LM), AphasiaBank texts show dense symptom co-occurrence, whereas LMs show less clear co-occurrence structure: 33 human symptom pairs exceed  $\phi > 0.2$  (a modest positive co-occurrence relationship) versus only three LM pairs, and none of the 255 LM samples matches the classical-syndrome template rule from Roll et al. [2025], whereas 50 of 306 human samples do (Fisher exact  $p \approx 1.1 \times 10^{-14}$ ). Nonetheless, the co-occurrence patterns are not entirely dissimilar: human-LM co-occurrence alignment is significant under the full greedy assay, although it is decoder-dependent under nucleus sampling (Appendix O).

Importantly, there is some shared structure between human and LM symptom profiles. This was seen above in the evidences that human and LM symptom co-occurrence matrices are (weakly) correlated, and Appendix Figure 8 further shows that LM symptom profiles can sometimes resemble human ones, at least in their relative effects, when specifically selected to do so. Thus, our results are not entirely incompatible with prior reports of convergence between lesioned LMs and human aphasia subtypes [Wang et al., 2025, 2026]. But when interpreted within the full context of strong evidence of large qualitative differences between human aphasias and artificial “aphasias” in lesioned LMs, it becomes clear that these dimensions of convergence do not tell the full story, and that the emergent functional organization of Transformer LMs remains at least qualitatively distinct from that of the human brain.

## 5 Discussion

In this work, we adapt methodologies and theoretical insights from the study of human aphasia in order to understand the emergent functional organization of language models, leveraging the recently developed Text Aphasia Battery [TAB Roll et al., 2025]. The result is not a claim that LMs have aphasia in the clinical sense, but rather a way to ask a neuropsychological question of a transformer: when a component is damaged, what kind of language breakdown follows?

The main finding is that attention and FFN lesions differ in their symptom profiles. FFN lesions push LM outputs toward vague content, short or formulaic strings, drift from the prompt, and, in the case of some lesions to the Gate matrix, punctuation-only or repeated function-word output. Attention lesions occupy a different region of TAB space, with Q/K and V/O forming attention-like subclusters. This is consistent with prior component analyses linking attention heads to syntactic and relational structure [Clark et al., 2019] and FFN layers to lexical or conceptual information [Geva et al., 2022]. Our contribution is to show that those internal differences surface as different “symptoms” that causally arise from damage to specific parts of the model.

This reframes earlier LM-aphasia results. Prior lesion studies asked whether LM damage can resemble e.g., Broca- or Wernicke-like aphasia subtypes [Wang et al., 2025, 2026]. We indeed find evidence suggestive of such a result, when we deliberately look for it. But we also find that, when considering the full pattern of symptoms, any convergences between LM “aphasias” and human aphasias are swamped by the divergences: the same symptoms appear in both systems, yet they combine strikingly differently. In other words, human aphasia gives us a clinically grounded coordinate system for studying emergent function in LMs, but based on our empirical results, this is not a license to equate transformer components with human brain areas or transformer impairments with human aphasias. Still, glimmers of convergence warrant further research into the contribution of architecture, training paradigm, and task as we seek to decouple the biological and computational aspects of language processing.

## 6 Limitations

A major limitation of our study is our dependence on the LM scorer (Gemini). We have attempted to mitigate this concern based on prior validation of the scorer against expert annotations [Roll et al., 2025], as well as by conducting secondary analyses that focus only on the most reliably-scored symptoms, or by using a different model [Gemma 4 E2B Google DeepMind, 2026] as the scorer. These additional analyses yield qualitatively similar findings, although details differ at the level of individual symptoms. We therefore treat individual symptoms as descriptive coordinates under a fixed measurement instrument, not as scorer-invariant clinical facts.

The intervention design is also limited. Bernoulli weight masks are a blunt lesioning tool, and ablations can create model states outside the distribution encountered during training (c.f. Hase et al. [2021]). We ran a follow-up analysis that mitigates this concern by replacing weights with their empirical means (rather than zero), resulting in highly similar findings (Appendix D.1). Beyond this, the fact that we fixed the random seed across some comparisons gives us tight experimental control at the cost of potentially increased sensitivity to the seed itself. We attempted to mitigate this concern by varying the seed in targeted *post hoc* analyses and by comparing to deterministic interventions like 100% ablations, but it was not feasible to sweep our entire analysis pipeline over random seeds. We also focused exclusively on the LM weight matrices and not on other plausible intervention sites (e.g., activations in the residual stream), and we therefore cannot comment on the functions supported by other candidate intervention sites.

Our study favors controlled coverage over ecological breadth. Primary inference uses a 19-prompt TAB subset and greedy decoding, which can amplify repetition loops. Targeted analyses of the choice of decoding algorithm show that the attention/FFN contrast survives when repetition is suppressed, but repetition-heavy symptoms and human-LM co-occurrence patterns are decoder-sensitive. These controls make a simple “FFN lesions are just larger lesions” explanation less likely, but they do not perfectly match attention and FFN lesions by functional impact across all models and layers.

Finally, the fact that we have explicitly attempted to bridge LM interpretability research and clinical aphasia research could lead to misinterpretation (e.g., that we are claiming that lesioned LMs have “aphasia” in the usual sense, which is not the case, and if anything the opposite of our argument). Clinical symptom names can invite anthropomorphic readings, model diagnoses, or unsupported claims about human neurobiology. They could also be used carelessly in public-facing evaluations of model behavior, where “aphasia” may be mistaken for a medical or cognitive state. We avoid that reading explicitly: raw AphasiaBank text is not redistributed, TAB symptoms are used only as operational descriptions of LM behavior, and no attempt is made to map LM components to human brain circuitry.

## References

- Paul Broca. Remarques sur le siège de la faculté du langage articulé, suivies d’une observation d’aphémie (perte de la parole). *Bulletins de la Société Anatomique de Paris*, 36:330–357, 1861.
- Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1.
- Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL <https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing>.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 2019. doi: 10.18653/v1/W19-4828.
- Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.
- Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- Nina F. Dronkers. A new brain region for coordinating speech articulation. *Nature*, 384(6605): 159–161, 1996. doi: 10.1038/384159a0.
- Nina F. Dronkers, David P. Wilkins, Robert D. Van Valin, Brenda B. Redfern, and Jeri J. Jaeger. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1–2):145–177, 2004. doi: 10.1016/j.cognition.2003.11.002.
- Nina F. Dronkers, Maria V. Ivanova, and Juliana V. Baldo. What do language disorders reveal about brain–language relationships? from classic models to network approaches. *Journal of the International Neuropsychological Society*, 23(9–10):741–754, 2017. doi: 10.1017/S1355617717001126.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>.
- Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, 2024. doi: 10.1038/s41583-024-00802-4.
- Angela D. Friederici. The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4):1357–1392, 2011. doi: 10.1152/physrev.00006.2011.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, 2020. doi: 10.18653/v1/2020.acl-demos.10.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586, 2021.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.

- Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 2022. doi: 10.18653/v1/2022.emnlp-main.3.
- Harold Goodglass. *Understanding Aphasia*. Academic Press, San Diego, CA, 1993. ISBN 0122900405.
- Harold Goodglass, Edith Kaplan, and Barbara Barresi. *The Assessment of Aphasia and Related Disorders*. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition, 2001. Includes the Boston Diagnostic Aphasia Examination.
- Google DeepMind. google/gemma-4-E2B-it. Hugging Face model card, 2026. URL <https://huggingface.co/google/gemma-4-E2B-it>. Accessed: 2026-05-06.
- Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 3650–3666, 2021.
- Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007. doi: 10.1038/nrn2113.
- Geoffrey E. Hinton and Tim Shallice. Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991. doi: 10.1037/0033-295X.98.1.74.
- Andrew Kertesz. *WAB-R: Western Aphasia Battery–Revised*. Pearson, San Antonio, TX, 2007.
- Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2, pages 598–605. Morgan Kaufmann, 1990. URL <https://proceedings.neurips.cc/paper/1989/hash/6c9882bbac1c7093bd25041881277658-Abstract.html>.
- Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011. doi: 10.1080/02687038.2011.589893.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Meta AI Blog, September 2024. URL <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>. Accessed: 2026-05-06.
- Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=9XFSbDPmdW>.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- David C. Plaut and Tim Shallice. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10(5):377–500, 1993. doi: 10.1080/02643299308253469.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.

- Nathan Roll, Jill Kries, Flora Jin, Catherine Wang, Ann Marie Finley, Meghan Sumner, Cory Shain, and Laura Gwilliams. The text aphasia battery (TAB): A clinically-grounded benchmark for aphasia-like deficits in language models. *arXiv preprint arXiv:2511.20507*, 2025.
- Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.
- Kate Swinburn, Gill Porter, and David Howard. *Comprehensive Aphasia Test*. Psychology Press, 2005.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 furious, 2025. URL <https://arxiv.org/abs/2501.00656>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018. doi: 10.18653/v1/W18-5446.
- Chengcheng Wang, Zhiyu Fan, Zaizhu Han, Yanchao Bi, and Jixing Li. Emergent modularity in large language models: Insights from aphasia simulations. *bioRxiv*, 2025. doi: 10.1101/2025.02.22.639416. Preprint.
- Yifan Wang, Jichen Zheng, Jingyuan Sun, Yunhao Zhang, Chunyu Ye, Jixing Li, Chengqing Zong, and Shaonan Wang. Component-level lesioning of language models reveals clinically aligned aphasia phenotypes. *arXiv preprint arXiv:2601.19723*, 2026.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl\_a\_00321.
- Carl Wernicke. *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Max Cohn & Weigert, Breslau, 1874.
- Stephen M. Wilson, Dana K. Eriksson, Sarah M. Schneck, and Jillian M. Lucanie. A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLOS ONE*, 13(2):e0192773, 2018. doi: 10.1371/journal.pone.0192773.
- Stephen M. Wilson, Jillian L. Entrup, Sarah M. Schneck, Caitlin F. Onuscheck, Deborah F. Levy, Maysaa Rahman, Emma Willey, Marianne Casilio, Melodie Yen, Alexandra C. Brito, Wayneho Kam, L. Taylor Davis, Michael de Riesthal, and Howard S. Kirshner. Recovery from aphasia in the first year after stroke. *Brain*, 146(3):1021–1039, 2023. doi: 10.1093/brain/awac129.



Table 2: **What each analysis is allowed to support.** This table separates the primary fixed-assay claim from validation, controls, clinical-reference analyses, and exploratory summaries.

Role	Analyses	Claim licensed
Primary test	Paired all-symptom profile distance over 360 model-layer-severity strata; deterministic 100% subset	Attention and FFN components separate as aggregate profiles in this fixed intervention/scorer/prompt design.
Validation and stress tests	Burden removal, reliability filters, decoder-sensitive symptom removal, scorer-free features, independent rerating, prompt/model spread, surface seed grid, scored mask-seed slice	The profile split is not explained by one simple burden, decoder, scorer, prompt, model, or mask artifact within the targeted coverage.
Perturbation controls	Matched-random, visible-damage matching, next-token KL matching, residual-state matching, joint KL+residual matching	Generic damage magnitude is less plausible, but full functional-dose matching remains open.
Clinical reference	AphasiaBank prevalence, co-occurrence, and likelihood contrasts	TAB symptoms describe both corpora; LM failure profiles organize differently from the human reference distribution.
Exploratory summaries	Depth maps, broader minimal-pair prompt probes, 3B/7B probes, human-reference cosines	Hypotheses about profile recurrence, scale, depth, and human-reference similarity for future work.

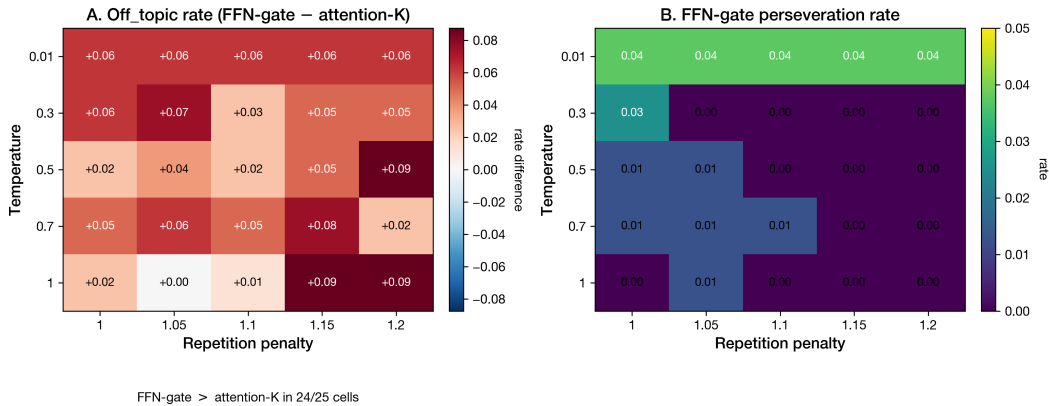


Figure 4: **FFN-vs-attention split persists across decoding settings.**  $5 \times 5$  grid of temperature and repetition penalty (Gemma-3-1B-IT). (A) Gate-minus-K off-topic rate difference: nonnegative in every cell and strictly positive in 24/25 cells. (B) FFN-Gate repetition-loop symptom rates are concentrated at low-temperature/no-penalty settings and suppressed by higher repetition penalties; they are zero at  $T = 0.7$ , repetition penalty 1.2.

Table 3: **Calibration of the attention-vs-FFN profile contrast.** Rows separate primary Gemini/TAB evidence, scorer-independent text features, independent-rerater checks, and targeted stress tests. TAB distances are all-symptom FFN-vs-attention profile separations in percentage-point symptom-rate space unless otherwise noted.

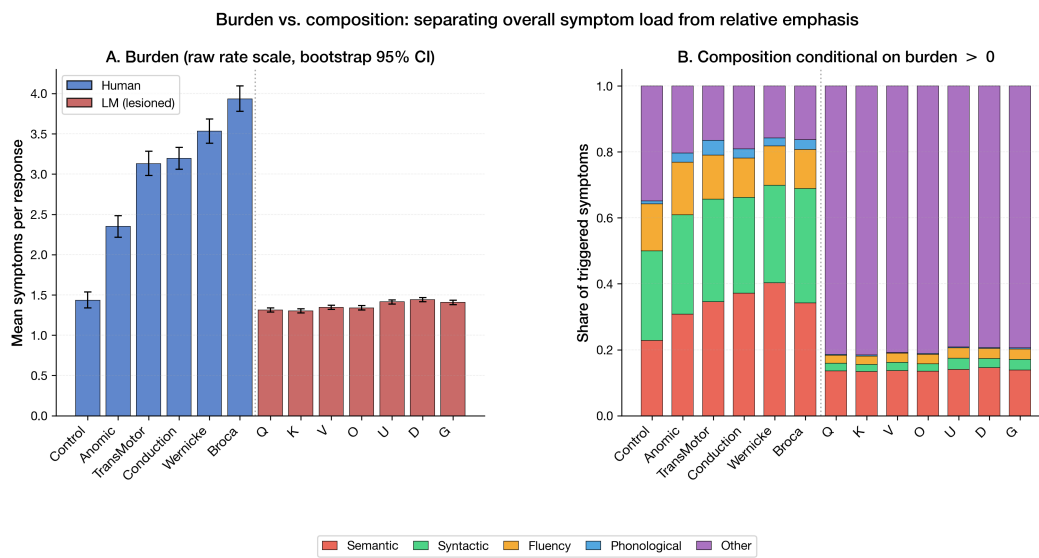
Status	Check	Unit and effect	Takeaway
Primary scored	Primary fixed assay	360 paired strata; 3.73 [2.45, 5.66], $p = .0004$	Attention and FFN condition profiles separate under the main scorer.
Primary scored	Burden removed	360 paired strata; 3.21 [2.24, 4.79], $p = .0002$	Separation is not only a higher overall symptom rate.
Seed-free scored	Deterministic 100%	90 paired strata; 8.05 [4.45, 14.46], $p = .018$	The profile contrast sharpens when mask randomness is absent.
Symptom reliability	Reliability filtered	12 symptoms / 11 stricter symptoms; $\kappa > .30$ : 3.50, $p = .0002$ ; AC1 > .65: 1.54, $p = .052$	Moderate filter preserves the effect; strict filter is directional but borderline.
Independent scorer	Independent rerating	300+480 Gemma 4-scored outputs; agreement 93.6%; slice L2 9.86 pp	Supports a coarse component read-out; individual symptom directions remain scorer-dependent.
Scorer-free text	Deterministic features	360 paired strata; $z\text{-}L_2 = 0.295$ [0.205, 0.433], $p = .0002$	Some component separation is visible in raw generated-text features before TAB scoring.
Symptom stress	Decoder-sensitive symptoms removed	18 symptoms; 3.20 [2.17, 4.61], $p = .0002$	Profile split is not only repetition, stereotypy, or off-topic symptoms.
Surface matched	Visible damage matched	360 matched strata; 3.25 [2.03, 5.57], $p = .016$	Profile shape remains after length matching.
Surface residual	Length/diversity conditioned	360 paired strata; 2.52 [2.00, 3.85], $p = .015$	Residualized TAB profiles still separate after simple surface features are regressed out.
Dependence check	Grouped spread	97 model-prompt cells; 20 prompts; 5 models; 77/97 cells; 18/20 prompts; 4/4 severities; 5/5 models aligned	Broad recurrence over realized groups; still not population-level family inference.
Dose proxy	KL-matched slices	96 same-layer FFN-to-attention matches; median L2 28.3; median log-KL gap 0.021	Two-model dose-proxy check; next-token KL, not generated-output KL.
Dose proxy	Residual-state matched slices	96 same-layer FFN-to-attention matches; median L2 28.3; median log-gap 0.004	Two-model hidden-state proxy check; not a full activation-norm match.
Dose proxy	Joint KL+residual match	96 same-layer FFN-to-attention matches; median L2 27.2; log gaps 0.028/0.023	Simultaneous proxy matching still leaves distinct profiles; generated-output matching remains open.
Mask-seed slice	Mask-seed rerated slice	480 Gemma responses; K-vs-FFN L2 0.173 vs. within-component seed L2 0.135	Not a single extreme mask draw; not full seed marginalization.
Mask-seed grid	Scorer-free seed grid	1,600 Gemma responses; K-vs-FFN z-distance 2.51 vs. seed distance 2.42	Seed jitter is substantial in surface features, so mask claims stay bounded.

Table 4: **Example outputs show different lesion patterns.** Lesion conditions are at 100% severity on Gemma-3-1B-IT for the prompt “Describe what is happening in the following scene: The boy is pushing the girl.” The AphasiaBank row is a de-identified Cinderella narrative excerpt from the Anomic group for orientation, not a paired example on the same prompt.

Condition	Primary TAB symptoms	Output
Intact LM	none in this excerpt	The girl is looking down at him. The girl is holding a small, blue toy. This scene depicts a moment of conflict and potential tension.
Attn-K (L12)	repetition-loop symptom	The boy is pushing the girl. The boy is pushing the girl. The boy is pushing the girl. The boy is pushing the girl.
FFN-Gate (L0)	Stereotypies and automatisms; Short and simplified utterances	the. the. the. the. the. the. the. the. the.
FFN-Up (L8)	Meaning unclear	The girl is the girl. She has a thing. The boy has a thing. They are doing the thing now.
AphasiaBank ref. (Anomic group)	Anomia-oriented word-retrieval symptom	The girl uh she dropped the the the slipper at the ball and the prince he found the the thing and he was looking for her. . .

Table 5: **Humans skew semantic/syntactic; lesioned LMs skew Other.** Mean symptom prevalence within category for AphasiaBank ( $n = 6,000$ ) vs. raw severity  $> 0$  LM scored records ( $n = 112,426$ ), restricted to the 19 symptoms shared with human analyses. These raw prevalence rows are separate from the 55,279-row prompt-deduplicated common table used for co-occurrence. Other = repetition-loop symptoms, stereotypies/automatisms, and off-topic responses.

Category	Human (%)	LM (%)	$\Delta$	Cohen’s $d$
Semantic	17.6	4.4	+13.2	0.61
Syntactic	18.5	1.1	+17.4	1.30
Fluency	13.9	1.4	+12.5	0.91
Phonological	4.4	0.2	+4.2	0.70
Other	15.1	32.3	-17.2	-0.37



**Figure 5: Humans and lesioned LMs differ in both burden and composition.** (A) Human diagnosis groups vary in mean TAB-symptom count, whereas LM components have similar overall burdens. (B) Conditional on at least one positive TAB symptom, category shares sum over triggered symptoms and are not raw prevalence rates; human outputs distribute across semantic, syntactic, and fluency categories, while lesioned-LM outputs are dominated by Other-category symptoms.

### A.1 Analysis denominators

Table 6: **Analysis denominators.** The paper reports several datasets because different questions require different row units and symptom inventories.

Use	Row unit	$n$	Notes
Primary lesion profiles	Scored LM output record, severity > 0	112,426	21 TAB symptoms; used for raw prevalence and full-sweep profile summaries.
Condition-level bootstrap	Model $\times$ layer $\times$ component $\times$ severity profile	2,528	Prompt-deduplicated first, then paired within 360 model $\times$ layer $\times$ severity strata.
Human-LM raw prevalence	Scored LM output record, severity > 0	112,426	Same raw LM records as primary lesion profiles, but restricted to the 19 symptoms shared with AphasiaBank.
Common occurrence table	co- Deduplicated LM row on common symptom inventory	55,279	Includes intact baselines; used for human-LM co-occurrence analyses on 19 symptoms.
Filtered LM sensitivity	Raw LM records excluding repetition/stereotypy positives	62,862	Used to test whether the attention/FFN split is only a repetition artifact.
Human reference	AphasiaBank production	6,000	1,000 per aphasia diagnosis group plus controls; common human/LM analyses use 19 symptoms.

### A.2 AphasiaBank corpora

The 6,000-production human-reference sample draws from these 30 AphasiaBank corpus labels: ACWT, Adler, BU, Baycrest, CMU, Capilouto, Elman, Fridriksson, Fridriksson-2, Garrett, Kansas, Kempler, Kurland, MSU, NEURAL, NEURAL-2, Richardson, SCALE, STAR, TAP, TCU, TCU-bi, Thompson, Tucson, UMD, UNH, Whiteside, Williamson, Wozniak, and Wright.

## B Paired condition-level bootstrap

To reduce dependence on raw row counts, we ran a condition-level bootstrap for the main attention/FFN contrast. We first removed duplicate model  $\times$  layer  $\times$  component  $\times$  severity  $\times$  prompt rows, averaged symptoms within each model  $\times$  layer  $\times$  component  $\times$  severity condition, and then compared FFN against attention within each shared model  $\times$  layer  $\times$  severity stratum. This gives 2,528 condition profiles organized into 360 paired strata. Bootstrap intervals resample these strata with replacement 5,000 times.

Table 7: **Condition-level clustering preserves the selected FFN-shift contrast.** Values are FFN-minus-attention differences in percentage points after pairing within model  $\times$  layer  $\times$  severity strata. Selected FFN-shift symptoms are *Meaning unclear*, *Stereotypies and automatisms*, *Off-topic*, *Perseverations*, *Short and simplified utterances*, and *Target unclear*.

Contrast	Difference (pp) $\uparrow$	Bootstrap 95% CI
Selected FFN-shift symptoms	<b>1.25</b>	[0.56, 1.99]
All remaining symptoms	0.08	[-0.03, 0.19]
Difference-in-differences	<b>1.17</b>	[0.53, 1.86]

The same condition-level table also supports tests that do not depend on the selected set. We report an all-symptom paired  $L_2$  profile distance and a restricted permutation that shuffles component names only within each model  $\times$  layer  $\times$  severity stratum, preserving the design and component counts. The selected-set effect is positive in 10 of 12 family  $\times$  severity cells (negative only at 25% severity in Gemma and Llama), stays positive when any family is left out (1.16–1.37 points), is positive within each subtest (1.03–2.01 points), and stays positive in all 20 leave-one-prompt analyses (1.18–1.38 points). Intact-subtracted deltas give a similar difference-in-differences of 1.20 points (95% CI [0.46, 1.94]).

Table 8: **All-symptom paired profile test.** FFN-minus-attention differences are estimated after prompt deduplication and pairing within model  $\times$  layer  $\times$  severity strata.

Statistic	Estimate	Interval / $p$
All-symptom $L_2$ profile distance	3.73 pp	[2.45, 5.66], $p = 0.0004$
Selected-set minus remaining symptoms	1.17 pp	[0.53, 1.86], $p = 0.0004$
Restricted within-stratum permutation	3.73 pp	$p = 0.0002$

Table 9: **Largest symptom-level FFN-minus-attention effects.** Differences are percentage points;  $q$  is BH-adjusted over 21 symptoms.

Symptom	Diff.	95% CI	$q$
Meaning unclear	2.72	[1.55, 3.95]	0.002
Off topic	1.36	[0.10, 2.69]	0.151
Stereotypies and automatisms	1.05	[0.07, 2.06]	0.163
False starts	0.98	[0.53, 1.45]	0.002
Perseverations	0.86	[-0.39, 2.08]	0.285
Short and simplified utterances	0.79	[0.16, 1.43]	0.092
Target unclear	0.71	[-0.02, 1.48]	0.166
Neologisms	0.52	[-0.06, 1.13]	0.197

Table 10 reports the same contrast under coarser aggregation units. The point is to clarify the dependence structure, not to enlarge the claim: prompt aggregation is robust inside the fixed prompt set, while model and family aggregation are directionally aligned but underpowered.

Table 10: **Dependence-sensitive views of the attention-vs-FFN profile contrast.** The same cached condition profiles are summarized at increasingly conservative units. The condition-stratum row is the primary fixed-assay test; coarser rows show how far the direction persists when prompt, severity, model, and family dependence are made visible.

Unit	Count	Effect	Interpretation
Condition strata	360 paired model-layer-severity strata	$L_2 = 3.73$ pp [2.45, 5.66], $p = .0004$	Primary fixed-assay estimate over realized lesion conditions.
Model-prompt cells	97 cells	$L_2 = 4.25$ pp [2.55, 8.75], 77/97 aligned	Coarser clustered view across the realized models and prompts; descriptive, not population inference.
Prompts	20 prompts	$L_2 = 3.20$ pp, 18/20 aligned, $p < .0001$	Broad over the fixed prompt subset; not a claim about all possible TAB prompts.
Severities	4 nonzero levels	$L_2 = 3.28$ pp [1.64, 5.98], 4/4 aligned	The direction is not confined to one nonzero lesion severity.
Model variants	5 realized variants	$L_2 = 3.93$ pp, 5/5 aligned, $p = .091$	Directionally recurrent, but exact top-level inference is underpowered.
Families	3 families	$L_2 = 3.64$ pp, 3/3 aligned, $p = .333$	Useful stress test, not population-level architectural generalization.

## C Visible-damage matching and clustered human-reference checks

Matched-random controls ask whether targeted lesions differ from random lesions with the same parameter count. We also ran a cached condition-level stress test that keeps the original target interventions but compares each FFN condition with the closest attention condition in the same model  $\times$  layer  $\times$  severity stratum under simple visible-damage measures. This is weaker than an activation-norm or KL-matched control, but it asks a useful narrow question: does the profile contrast disappear when attention and FFN conditions are matched on output length or observed symptom burden?

Table 11: **FFN-vs-attention profile contrast after matching visible damage proxies.** Each row matches each FFN condition to the closest attention condition within the same model  $\times$  layer  $\times$  severity stratum, averages the resulting FFN-minus-attention vectors within stratum, then applies the same sign-flip profile test over 360 strata. Match distance is Euclidean distance after standardizing the listed matching variables within stratum.

Matching variables	$L_2$ distance (pp) $\uparrow$	95% CI; $p$	Mean match dist. $\downarrow$
Mean word count	3.25	[2.03, 5.57]; .0158	0.70
Mean symptom burden	<b>4.60</b>	[3.37, 6.33]; .0002	<b>0.69</b>
Word count, lexical diversity, alphabetic ratio	4.09	[2.63, 6.11]; .0002	1.84
Word count and symptom burden	4.22	[2.81, 6.20]; .0002	1.42

The top symptom differences after word-count matching are *Meaning unclear* (+2.49 pp), *Off-topic* (+1.06 pp), *Stereotypies and automatisms* (+0.94 pp), *Perseverations* (+0.71 pp), and *Short and simplified utterances* (+0.51 pp). These checks do not replace perturbation-matched controls. They support the narrower claim that, within the fixed TAB/Gemini assay, attention and FFN lesions differ in profile shape even when compared at similar visible output length or symptom burden.

Table 12: **Human-reference category differences under corpus/model clustered bootstrap.** The bootstrap resamples AphasiaBank corpora and LM model variants before recomputing category means. Positive values indicate higher human prevalence; negative values indicate higher LM prevalence.

Category	Human (%)	LM (%)	Clustered diff. (pp)
Semantic	17.6	4.0	+13.6 [8.4, 18.3]
Syntactic	18.5	1.1	+17.4 [15.8, 19.6]
Fluency	10.1	1.5	+8.7 [7.4, 10.5]
Phonological	4.4	0.2	+4.3 [3.4, 6.0]
Other	15.1	30.9	-15.8 [-41.3, 2.2]

The clustered bootstrap preserves the broad human>LM direction for semantic, syntactic, fluency, and phonological categories. The Other-category contrast is less stable because it is driven by decoder-sensitive repetition-loop and off-topic symptoms in a small set of model variants; we therefore use the human comparison mainly as a clinical reference for profile organization.

## D Targeted internal-perturbation calibration slices

To probe whether the attention-vs. FFN profile contrast disappears when attention and FFN lesions are compared at similar internal perturbation magnitude, we ran two output-KL calibrations. Both measure next-token KL divergence from the intact model on the same 20 TAB prompts, then read the already-scored TAB profile from the main checkpoint for the matched condition. The original Gemma-3-1B-IT slice compared K with Gate/Up at layers 5/10/15/20 and severities 25/50/75/100; the 32 FFN-to-K matches have a median absolute  $\log_{10}$  KL gap of 0.087 and a median single-condition TAB-profile  $L_2$  distance of 21.2 percentage points (range 7.1–41.5).

We then broadened the calibration to the two locally available 1B models with cached weights (Gemma-3-1B-IT and Llama-3.2-1B), all seven components, four representative layers per model,

and severities 25/50/75/100. Each FFN condition was matched to the closest attention condition (Q/K/V/O) in the same model and layer by log-KL. Across 96 same-layer FFN-to-attention matches, the median absolute  $\log_{10}$  KL gap is 0.021 and the median TAB-profile  $L_2$  distance is 28.3 percentage points; 84/96 matches fall within a 0.25  $\log_{10}$  KL gap and still have median  $L_2 = 28.7$  points. The per-model medians are 15.9 points for Gemma and 41.1 for Llama. Allowing the matched attention condition to come from any inspected layer reduces the median log-KL gap to 0.004 and gives a similar median profile distance of 26.9 points.

We repeated the same matched-profile analysis with a residual-state proxy. For each inspected condition, the calibration script measures the mean normalized change in the final-token, final-layer hidden state on the same 20 TAB prompts, then matches each FFN condition to the nearest attention-family condition by log residual-state change. Across the same 96 same-layer matches, the median absolute  $\log_{10}$  residual gap is 0.004 and the median TAB-profile  $L_2$  distance is 28.3 percentage points; every match falls within a 0.25  $\log_{10}$  gap. Allowing any inspected layer reduces the median residual gap to 0.001 and gives median  $L_2 = 28.7$  points. The same-layer per-model medians are 17.0 points for Gemma and 46.2 for Llama.

Table 13: **Residual-stream matched profile calibration.** FFN conditions are matched to the closest attention-family condition by mean final-token residual-state change on the same 20 TAB prompts, then compared using already-scored TAB profiles from the main checkpoint.

Matching set	$n$	Median log gap	Median TAB $L_2$
Same model and layer	96	0.004	28.3 pp
Same model, any inspected layer	96	0.001	28.7 pp
Gemma same-layer subset	48	0.008	17.0 pp
Llama same-layer subset	48	0.002	46.2 pp

Table 14: **Joint dose-matched profile calibration.** FFN conditions are matched to attention-family conditions by simultaneous next-token KL and final-residual-state perturbation on the same 20 TAB prompts, then compared using already-scored TAB profiles.

Matching set	$n$	Median KL gap	Median residual gap	Median TAB $L_2$
Same model and layer	96	0.028	0.023	27.2 pp
Same model, any inspected layer	96	0.011	0.010	28.0 pp
Gemma same-layer subset	48	0.071	0.043	16.7 pp
Llama same-layer subset	48	0.014	0.016	39.8 pp

Finally, we matched on both proxies at once. Same-layer joint matching gives median log gaps of 0.028 for KL and 0.023 for residual-state change, while the median TAB-profile distance remains 27.2 points; 84/96 matches are within 0.25  $\log_{10}$  units on both proxies and have median  $L_2 = 27.6$  points. These calibrations are not the primary inferential test. Output KL uses next-token rather than full generated-output distributions; residual-state matching uses a final-layer hidden-state proxy rather than full layerwise activation-norm matching; both cover two locally cached models; and partial-severity masks use deterministic seeds from the calibration scripts rather than necessarily the original main-sweep masks. Even with those limits, the result is informative: matching on model-internal perturbation proxies does not make FFN profiles resemble attention profiles in this assay. A full perturbation-matched control across all models, layers, severities, decoding policies, and generated-output distributions remains future work.

### D.1 Replacement-value control for zero-ablation

Zero-ablation sets damaged weights to a fixed value, so one possible concern is that the result depends on choosing zero rather than a more typical value for the matrix. We ran a bounded local check on Gemma-3-1B-IT using the same 20 TAB prompts, K/Gate/Up/Down components, layers 5/10/15/20, severities 75% and 100%, and greedy decoding. For each mask, we compared zero replacement with three mean-replacement rules: global mean, row mean, and column mean. We then measured generated-text features and next-token KL from the intact model; this check does not use an external symptom scorer.

Table 15 gives the result. Global-mean replacement is very close to zero-out: the replacement values are tiny relative to the weights being removed, the median KL gap is 0.008  $\log_{10}$  units, and the K-vs-FFN generated-feature vector remains strongly aligned with zero-ablation. Row- and column-mean replacement are still small in weight space and track the zero-ablation KL ordering, but they introduce more surface-feature variance, especially under full ablation. The takeaway is therefore narrow: zero-out is not uniquely responsible for the perturbation scale, and global-mean replacement gives the same qualitative check; row/column means are less clean as behavioral controls. We use zero-out in the main sweep because it is the most transparent fixed replacement rule.

Table 15: **Replacement-value control for zero-ablation.** Gemma-3-1B-IT, K/Gate/Up/Down, layers 5/10/15/20, severities 75% and 100%, 20 TAB prompts, greedy decoding, no external scorer. The same mask is used for zero and mean replacement within each condition. Tensor ratio is the norm of the replacement values divided by the norm of the weights being removed. Feature distance is the average z-scored generated-text distance from the matched zero-ablation condition; lower values indicate closer behavior. Cosine compares the K-vs-FFN feature-difference vector with the zero-ablation vector. KL columns compare next-token KL from the intact model.

Replacement	Tensor ratio	Feature dist.	K-vs-FFN cosine	Median $ \Delta \log_{10} \text{KL} $	KL corr.	
Global mean	0.002	1.37	0.81	0.008	0.85	The zero-ablation
Row mean	0.025	2.46	0.38	0.070	0.79	
Column mean	0.049	2.70	0.40	0.190	0.68	

K-vs-FFN feature distance is 1.16 in the same z-scored space. Global-mean replacement is closest to zero; row and column means preserve perturbation scale but introduce more generation-level variance, especially under full ablation.

## E Reliability, burden, and dependence stress tests

To test whether the attention/FFN result depends on low-reliability symptoms, degenerate outputs, overall symptom burden, or treating correlated condition strata as fully exchangeable, we repeated the paired all-symptom profile statistic on cached subsets and conservative aggregations (Tables 16–17). These are stress tests of the profile-level contrast, not replacement validation for any individual symptom.

Table 16: **Profile contrast under reliability, output-form, and burden filters.** Each row reports the paired FFN-vs-attention  $L_2$  profile distance after prompt deduplication and pairing within model  $\times$  layer  $\times$  severity strata. The burden-adjusted row mean-centers each condition profile across symptoms before pairing, removing overall symptom burden by construction.

Subset	Symptoms	$L_2$ distance (pp) $\uparrow$	95% CI; $p$
All primary symptoms	21	3.73	[2.45, 5.66]; .0004
$\kappa > .30$ symptoms	12	3.50	[2.19, 5.29]; .0004
AC1 $> .65$ symptoms	11	1.54	[0.93, 2.88]; .052
Remove repetition/stereotypy/off-topic symptoms	18	3.20	[2.17, 4.61]; .0002
Drop rows with those symptoms positive	21	2.65	[1.46, 4.79]; .0052
Length $\geq 10$ chars and $\geq 2$ words	21	3.99	[2.51, 6.19]; .0004
Nondegenerate text filter	21	3.81	[2.27, 6.23]; .0004
Connected-text prompts only	21	<b>4.12</b>	[2.68, 6.68]; .0036
Non-connected prompts only	21	3.84	[2.69, 5.85]; .0006
Burden-adjusted profiles	21	3.21	[2.29, 4.81]; .0002

The stricter AC1 subset removes several high-prevalence subjective symptoms and should be read as a conservative lower-signal check; it preserves the direction but not conventional significance. The  $\kappa > .30$  subset, decoder-sensitive symptom removal, output-form filters, and burden-adjusted profiles preserve the contrast. The main all-symptom result is therefore not explained solely by low-reliability symptoms, repetition/off-topic symptoms, punctuation-only strings, connected-text repetition loops, or a small overall FFN burden increase.

All five model variants and all three model families show nonzero FFN-vs-attention profile separation in the same broad direction as the pooled effect (model-level cosine range with the pooled

Table 17: **Conservative top-level aggregations.** Each row first averages the FFN-minus-attention profile vectors within the listed unit, then computes the  $L_2$  distance of the unit-level mean vector. Exact sign-flip tests use only the unit count in the row, so the family-level test is intentionally severe with three units.

Aggregation unit	Units	$L_2$ distance (pp) $\uparrow$	Exact sign-flip $p$ $\downarrow$
Model variant	5	<b>3.93</b>	.091
Model family	3	3.64	.333
Prompt	20	3.20	< .0001

vector: 0.69–0.88; family range: 0.69–0.95). At the prompt level, 18 of 20 prompts have positive cosine with the pooled vector (range  $-0.009$  to  $0.870$ ), so the contrast is not driven by one prompt family. We report these unit-level tests to show the dependence structure, but treat model/family aggregation as sensitivity checks rather than primary inference because the number of top-level model units is small.

## F Effect-size calibration

The paired  $L_2$  statistic is useful but abstract, so we add two simple calibrations. First, we compare the attention-vs-FFN grouping with every same-size split of the seven components into a three-component group and a four-component complement. Second, we ask whether individual condition profiles are large enough to classify held-out attention-vs-FFN membership. Table 18 gives the intended reading: Gate/Up/Down vs. Q/K/V/O is the largest 3-vs-4 component partition in the pooled profile space, but held-out classifiers remain weak. The effect is meaningful as an aggregate profile signature, not as a reliable per-condition diagnostic.

Table 18: **Effect-size calibration for the profile contrast.** The FFN-vs-attention partition is compared with all same-size 3-vs-4 component partitions. Classifier rows are intentionally conservative: they ask whether single condition profiles are enough to recover attention-vs-FFN membership out of sample.

Calibration	Result	Interpretation
Severity-step anchor	$L_2 = 3.73$ pp is $1.01\times$ the median adjacent severity-step shift (3.68 pp)	The contrast is about one within-assay severity step: visible as a profile shift, but not a large per-condition diagnostic.
Component partition rank	Gate/Up/Down vs Q/K/V/O rank 1/35; $L_2 = 3.73$ pp	The transformer FFN grouping is the strongest same-size component split in the pooled profile space.
Next-best split	K/Q/V vs complement; $L_2 = 3.27$ pp	The observed anatomical split exceeds the nearest arbitrary grouping; median non-FFN split is 1.16 pp.
Leave-model classifier	balanced accuracy 0.553; AUROC 0.582	Individual condition profiles are only weakly classifiable; the effect is aggregate/profile-level.
Leave-family classifier	balanced accuracy 0.550; AUROC 0.570	Cross-family generalization is modest, matching the fixed-assay scope.
Leave-severity classifier	balanced accuracy 0.571; AUROC 0.586	Severity does not by itself make the profile split a strong per-condition classifier.

### F.1 Length/diversity-conditioned TAB profiles

Visible-damage matching compares conditions with similar output length or symptom burden. As a complementary check, Table 19 residualizes each condition-level symptom rate against simple generated-text features and then recomputes the same paired profile statistic. This is not a causal adjustment for all surface-form differences, but it tests a narrower alternative explanation: whether the

TAB profile separation is only a relabeling of shorter or less diverse text. The profile distance shrinks after conditioning on word count and lexical diversity, and shrinks further after adding repeated-token mass, but remains positive in both cases. The residual coordinates also change sign for some repetition-heavy symptoms, which is why we treat individual symptoms descriptively and rely on the full profile vector for the main claim.

Table 19: **TAB profile contrast after conditioning on simple surface features.** Each conditioning row residualizes the condition-level TAB-symptom rates against the listed generated-text features, then recomputes the paired FFN-minus-attention profile distance in percentage-point symptom-rate units.

Profile readout	Conditioned features	$L_2$ pp	95% CI; $p$
Raw TAB profile	None	3.73	[2.44, 5.66]; .0004
Condition on word count	Word count	4.09	[2.62, 6.16]; .0004
Condition on word count + lexical diversity	Word count, Lexical diversity	2.52	[2.00, 3.85]; .015
Condition on word count + lexical diversity + repetition mass	Word count, Lexical diversity, Repeated-token mass	2.10	[1.77, 3.33]; .044
<i>Largest residual FFN-minus-attention symptoms after word-count + lexical-diversity conditioning</i>			
Perseverations		-1.74 pp	residual coordinate
Meaning_unclear		+1.13 pp	residual coordinate
False_starts		+0.85 pp	residual coordinate
Jargon		-0.84 pp	residual coordinate

## G Scorer-free behavioral feature profile

To test whether the component split is visible without the TAB rubric, we computed deterministic features from generated text and prompts alone. These features are deliberately simple and do not replace the TAB profile: they cannot adjudicate meaning loss or clinical-style symptoms. They answer a narrower measurement question: whether attention and FFN lesions leave different surface traces before any LLM scorer is applied.

Table 20: **Scorer-free text-feature profiles also separate attention and FFN lesions.** Features are deterministic functions of the prompt and generated text, with no TAB symptoms or LLM scorer. Distances are Euclidean lengths of the mean FFN-minus-attention vector after z-scoring condition-level feature means.

Subset	Paired strata	Feature-profile $L_2$	95% CI; $p$
All nonzero severities	360	0.295	[0.205, 0.433]; < .0003
Severity $\geq$ 75%	180	0.528	[0.364, 0.797]; < .0003
Deterministic 100%	90	0.741	[0.469, 1.238]; .0022
<i>Largest all-severity raw FFN-minus-attention feature shifts</i>			
Alphabetic character share		-1.46 pp	z shift -0.212
Tiny response rate		+0.93 pp	z shift +0.107
Prompt-content recall		-1.83 pp	z shift -0.102
Prompt-bigram copy rate		+0.41 pp	z shift +0.084
Unique-word ratio		-1.87 pp	z shift -0.076
Repeated-token mass		+1.27 pp	z shift +0.045

## H Depth stratification details

Table 21: **Within-category depth correlations exceed cross-category.** Mean Pearson  $r$  between depth profiles, by TAB symptom category.

Correlation Type	Mean $r$	N Pairs
Within-category	<b>0.521</b>	41
Semantic	0.578	15
Syntactic	0.541	6
Fluency	0.295	10
Other	0.971	10
Cross-category	0.226	169

## I High-reliability symptom subset

Restricting to the high-AC1 symptoms *increases* the Mantel correlation from  $r = 0.373$  to  $r = 0.470$  ( $p = 0.002$ , 5,000 permutations). Shared co-occurrence structure is therefore not merely measurement noise in low-reliability symptoms.

## J Prompt manifest

Table 22: **Primary 20-prompt TAB subset.** The full TAB inventory is reported in [Roll et al. \[2025\]](#); the lesion sweep uses this fixed subset.

Subtest	Prompts
Connected text	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; Describe what is happening in the following scene: The boy is pushing the girl.
Repetition	Please repeat exactly: <i>house; breakfast; catastrophe; The sun rises in the East; The ambitious journalist discovered where we'd be going.</i>
Sentence comprehension	Are babies watched by babysitters? Do you cut grass with an axe? If you are about to leave, have you left yet? Are witnesses questioned by police? If I was at the park when you arrived, did I get there first?
Word comprehension	Which one is an animal with a mane? Which object is typically used to make music? Which item is usually worn on the feet? Which object is used for cutting? Which one is a large mammal with a long neck? Each item is presented with six choices.

## K Scorer audit

Table 23: **Per-symptom reliability for the six descriptive FFN-shift symptoms.** TAB columns come from the SLP validation analysis; Gemma 4 E2B and Qwen columns compare independent automatic reraters with the Gemini scorer on targeted 10-symptom subsets where available. Dashes indicate symptoms not included in that rerating subset.

Label	TAB $\kappa$	TAB AC1	Gemma 4 E2B $\kappa$	Qwen $\kappa$
Meaning unclear	0.467	0.469	0.069	0.365
Stereotypies/automatisms	0.115	0.793	–	–
Off-topic	0.436	0.503	0.276	–0.031
Perseverations	0.357	0.700	0.658	0.215
Short/simplified	0.465	0.471	0.201	–0.024
Target unclear	0.420	0.517	–	–

The scorer audit motivates two choices in the main text. First, we use Gemini as a fixed measurement instrument rather than treating any single symptom as ground truth. Second, we report profile-

level and category-level contrasts, not only individual symptoms, because several high-prevalence symptoms are subjective and rerater-sensitive.

## L Interpreting TAB symptoms for LM outputs

TAB symptom names come from aphasiology, but their LM-side use in this paper is operational. Table 24 gives the intended reading: a symptom describes a text pattern produced under a fixed prompt and scorer, not a clinical syndrome, communicative intent, or human cognitive mechanism.

Table 24: **What TAB symptoms can and cannot mean for LM text.** Symptom names are rubric shorthand; the main claims use profile-level contrasts over these operational measurements.

Symptom group	LM-side reading	Not implied
Semantic symptoms	The output lacks requested content, is vague, or makes the target unclear under the prompt.	Lexical-semantic impairment, word-retrieval failure, or preserved communicative intent.
Syntactic/morphology symptoms	The generated string omits function words or shows disrupted local form under the rubric.	Agrammatism, paragrammatism, or a human production-system deficit.
Phonological symptoms	The text contains nonwords, malformed strings, or near-word substitutions.	Speech-sound planning errors or phonological paraphasia in the clinical sense.
Fluency and output-form symptoms	The response is short, formulaic, halting, or low-information as text.	Motor-speech fluency, prosody, effort, or bedside fluency classification.
Perseveration, stereotypy, and off-topic symptoms	The decoder repeats, emits fixed formulas, or drifts from the requested content.	Clinical perseveration, stereotyped utterance, pragmatic impairment, or syndrome membership.

## M Partial-mask seed checks

To probe mask-seed variance without launching a full Gemini rescoring of the corpus, we ran three targeted checks on cached Gemma-3-1B-IT with 75% masks. The first check uses five independent Bernoulli seeds for representative layer-10 K and Gate components on an eight-prompt subset and reports only surface metrics. It asks whether coarse length/diversity effects are dominated by a single partial-mask draw.

Table 25: **Small 75% mask-seed surface check.** Values are mean  $\pm$  std across five Bernoulli masks. Repeated-token mass is the share of generated tokens whose type appears more than three times in the eight-prompt sample.

Component	Repeated mass	Unique ratio	Mean length	Tiny rate
Gate	0.187 $\pm$ 0.040	0.702 $\pm$ 0.038	16.80 $\pm$ 2.79	0.275 $\pm$ 0.094
K	0.136 $\pm$ 0.028	0.729 $\pm$ 0.012	17.23 $\pm$ 0.74	0.375 $\pm$ 0.000

We then expanded the surface-only check to four layers (5/10/15/20), K/Gate/Up/Down, five seeds, and all 20 prompts, generating 1,600 local responses with no external scorer. In z-scored surface-feature space, same-seed K-vs-FFN distances averaged 2.51, while within-component across-seed distances averaged 2.42 (ratio 1.04; Table 26). Surface-form differences are not dominated by one representative seed, but seed jitter is large enough that these features alone provide weak attention/FFN separation after seed marginalization.

We then scored a follow-up subset with a local Gemma 4 E2B rerater on the 10-symptom TAB-style subset used in Appendix K: K/Gate/Up/Down, layers 5 and 10, three seeds, all 20 prompts, 480 total responses. This check asks whether the representative partial-mask draw is an obvious outlier under TAB symptoms in layers where the main Gemma slice shows visible K-vs-FFN differences. Same-seed K-vs-FFN profile distances averaged  $L_2 = 0.173$  across 18 comparisons, while within-component across-seed distances averaged 0.135 across 24 comparisons (ratio = 1.27). The largest

Table 26: **Broader scorer-free mask-seed grid.** Gemma-3-1B-IT, 75% partial masks, four layers, five Bernoulli seeds, and 20 TAB prompts. Values are mean  $\pm$  std across layer–seed cells.

Projection	Mean length	Unique ratio	Repeated mass	Tiny rate
K	15.77 $\pm$ 0.91	0.550 $\pm$ 0.023	0.384 $\pm$ 0.033	0.352 $\pm$ 0.043
Gate	17.34 $\pm$ 2.86	0.565 $\pm$ 0.023	0.370 $\pm$ 0.035	0.268 $\pm$ 0.117
Up	16.14 $\pm$ 2.66	0.554 $\pm$ 0.026	0.366 $\pm$ 0.044	0.323 $\pm$ 0.094
Down	16.29 $\pm$ 2.27	0.561 $\pm$ 0.028	0.372 $\pm$ 0.040	0.310 $\pm$ 0.101

distance averaged 2.51, versus 2.42 for within-component across-seed distance (ratio 1.04).

same-seed contrasts were K–Gate at layer 5 (mean  $L_2 = 0.218$ ) and K–Down at layer 10 (mean 0.249). The component contrast is not created by one extreme partial-mask seed, but seed variance is large enough that this remains a targeted robustness check rather than full marginalization over masks.

Table 27: **Two-layer TAB-symptom mask-seed check.** Gemma 4 E2B rerates 480 responses from 75% masks at layers 5 and 10 on the full 20-prompt subset and a 10-symptom TAB-style rubric. Values are mean  $\pm$  std across six layer–seed cells.

Comp.	Mean symp-toms	Any symptom	Mean symp-tom SD	Top rerated symptoms
K	0.242 $\pm$ 0.067	0.242 $\pm$ 0.067	0.011	Short/simplified 0.225; Off-topic 0.017
Gate	0.275 $\pm$ 0.118	0.267 $\pm$ 0.118	0.026	Off-topic 0.133; Short/simplified 0.092; repetition-loop 0.050
Up	0.267 $\pm$ 0.055	0.258 $\pm$ 0.055	0.018	Short/simplified 0.158; Off-topic 0.067; repetition-loop 0.042
Down	0.192 $\pm$ 0.045	0.192 $\pm$ 0.045	0.015	Off-topic 0.083; Short/simplified 0.083; repetition-loop 0.025

## N Task-narrowed human comparison

To test whether the main prevalence contrast depends on heterogeneous elicitation contexts, we repeated the category-level comparison on a narrower slice: the three largest AphasiaBank narrative sections (Stroke, Cinderella, Important\_Event;  $n = 3,055$ ) versus lesioned LM connected-text outputs ( $n = 31,411$ ), with intact connected-text outputs shown for reference ( $n = 7,980$ ). The qualitative ordering is unchanged: humans remain higher on semantic and syntactic TAB symptoms, while LMs remain higher on Other-category symptoms. The Other-category gap is modestly larger under lesioning than in the intact greedy baseline, which suggests that lesions amplify an existing decoder-model repetition tendency rather than creating a new clinical pattern de novo.

Table 28: **Task-narrowed comparison preserves the human/LM ordering.** Mean symptom-level prevalence within each category for AphasiaBank narratives, lesioned LM connected text, and intact LM connected text.

Category	Human narrative	LM connected text	LM intact
Semantic	18.3	4.3	4.4
Syntactic	17.7	1.3	1.1
Fluency	15.0	1.3	0.0
Phonological	4.6	0.1	0.0
Other	14.2	33.4	32.3

## O Error-type co-occurrence analysis

We tested whether the *pattern* of which symptoms co-occur corresponds between AphasiaBank productions and LM outputs in the common table using a Mantel test on  $19 \times 19$  phi-coefficient matrices ( $n = 55,279$  LM rows including intact baselines,  $n = 6,000$  human). Under the primary

greedy assay, the correspondence is significant:  $r = 0.373$  ( $p = 0.0005$ , 50,000 permutations; Figure 6). Restricting to cross-category pairs gives  $r = 0.296$  ( $p = 0.006$ ). Removing rows with repetition-loop symptoms doubles the correlation to  $r = 0.662$  ( $p = 0.0001$ ). The alignment strengthens on the high-agreement symptoms ( $r = 0.470$ ; Section I).

Under nucleus sampling ( $T = 0.7$ ,  $p = 0.9$ , repetition penalty 1.2), the global Mantel alignment does not persist ( $r = 0.035$ ,  $p = 0.36$ ). The FFN-Gate/Up component dissociation remains qualitatively stable under the same decoding change (cosine to greedy: 0.68 Gate, 0.67 Up vs. 0.07 K). The human-LM co-occurrence correspondence changes with the decoder, while the attention/FFN contrast is more stable.

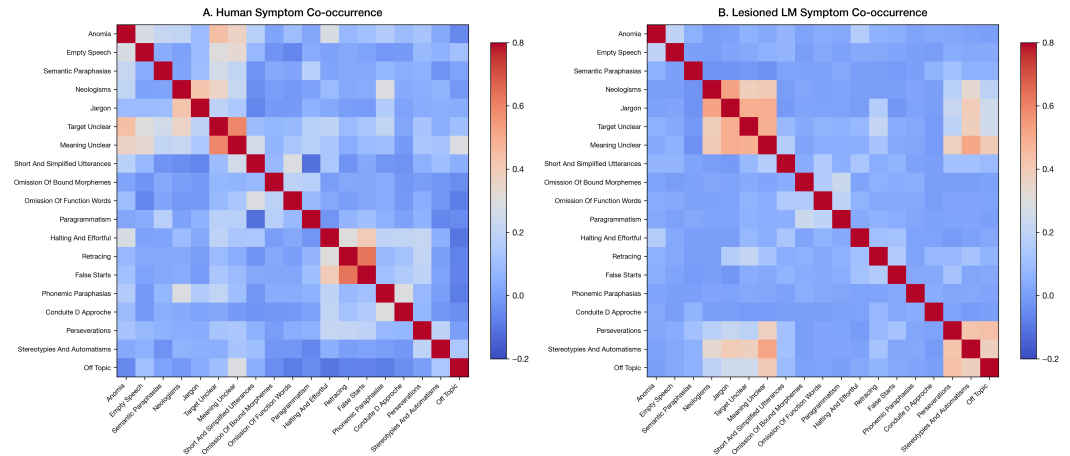


Figure 6: **Humans show denser TAB-symptom co-occurrence than LM rows in the common table.** (A) AphasiaBank ( $n = 6,000$ ). (B) LM rows including intact baselines ( $n = 55,279$  on the common 19-symptom inventory). Greedy assay: Mantel  $r = 0.373$ ,  $p = 0.0005$ ; repetition-filtered  $r = 0.662$ . Sampling decoding removes the alignment.

## P Exploratory human-reference profile cosines

To relate lesion profiles to the human contrast distribution descriptively, we compute cosine similarity between the mean-centered symptom profile of each human diagnosis group and each lesioned component. Table 29 reports the full matrix. These similarities are exploratory contrasts. They are not model diagnoses, and they do not show that a component implements a human syndrome.

Table 29: **Exploratory human-reference cosine summaries; only Anomic-K passes table-wide correction.** Cosine similarity between mean-centered human diagnosis-group profiles and lesioned-LM component profiles. Bold = largest descriptive cosine per row. Bootstrap 95% CIs (5,000 resamples) and selection-corrected  $p$ -values (max- $T$ , 10,000 permutations).

	Q	K	V	O	U	D	G	Largest [CI; $p_{row}$ ]
Anomic	.74	<b>.81</b>	.62	.69	.41	.40	.33	K [.41,.84]; .002
Broca's	.52	.60	.49	.54	.62	.59	<b>.67</b>	G [.19,.73]; .051
Wernicke's	.48	.53	.47	.50	<b>.59</b>	.54	.51	U [.16,.69]; .108
Conduction	.29	.38	.30	.32	.41	<b>.45</b>	.39	D [.11,.61]; .298
TC Motor	.23	.30	.26	.27	.29	.31	<b>.34</b>	G [-.08,.54]; .500
Control	.62	<b>.70</b>	.54	.61	.40	.39	.30	K [.42,.78]; .030

Only Anomic-K passes table-wide correction ( $p_{table} = 0.014$ ).

## Q Additional figures

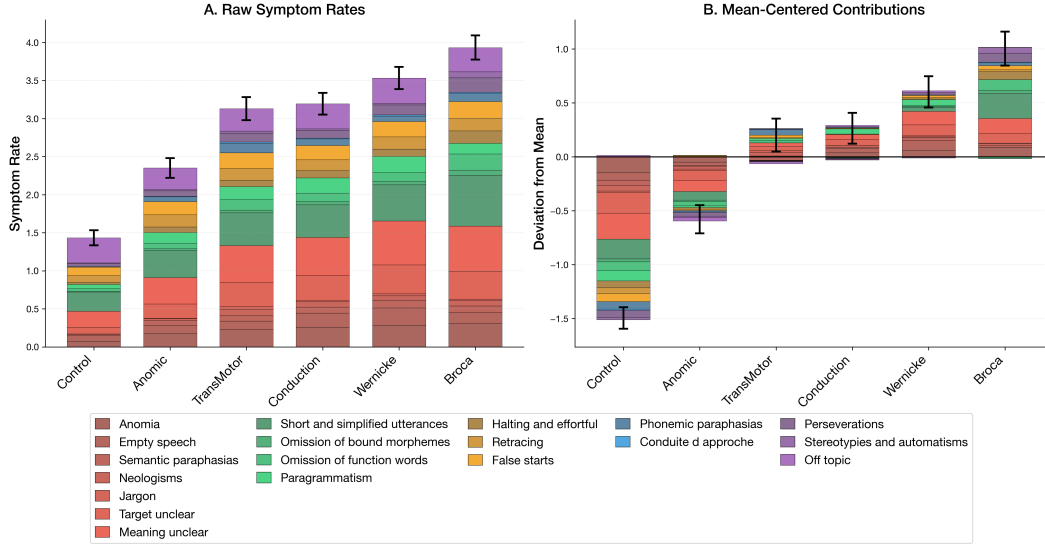


Figure 7: **Human TAB profiles by diagnosis.** (A) Raw symptom rates per diagnosis. (B) Deviations from the human corpus mean; panel A holds absolute magnitudes.

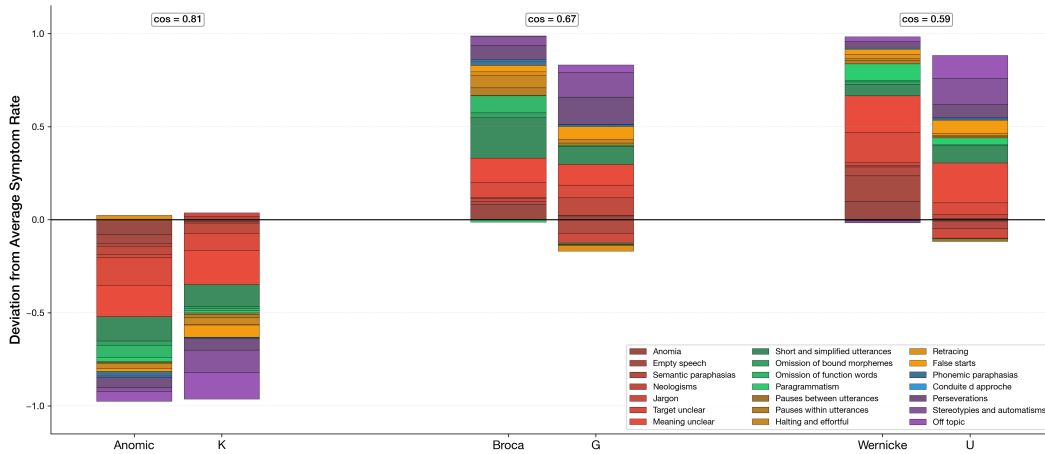


Figure 8: **Selected descriptive contrasts between human diagnosis profiles and LM component profiles.** Anomic vs. Attention-K (cosine 0.808); Broca’s vs. FFN-Gate (0.665); Wernicke’s vs. FFN-Up (0.592). These are exploratory profile similarities, not syndrome mappings.

## R Decoding sensitivity checks

The main text reports a  $5 \times 5$  decoding grid (Figure 4). In a separate targeted check, we re-ran K, Gate, and Up at layers 5, 10, 15, 20, and 25 and severities 75% and 100% under nucleus sampling ( $T = 0.7$ ,  $p = 0.9$ , repetition penalty 1.2) on Gemma-3-1B-IT and computed cosine similarity to greedy counterparts. FFN-Gate and FFN-Up profiles retain cosine  $> 0.67$ ; the attention-K profile diverges (cosine = 0.07). The FFN-vs-attention contrast is the more stable cross-decoder signal.

## S Scaling to 7B parameters

Single-layer ablation at 7B (Qwen2.5-7B-Instruct [Qwen et al., 2025]) produces no detectable TAB symptoms in this assay. Multi-layer ablation recovers the FFN-vs-attention dissociation: FFN-Gate repetition-loop symptom rate reaches 0.41 at 25% severity while attention-K remains at 0.02 (Figure 9). We also ran broader minimal-pair prompt probes to separate “TAB prompt” concerns from

scale concerns. On Qwen2.5-3B-Instruct, a 70-item syntax/semantics probe across layers 7/18/28 and K/Gate/Up at 75% severity shows a localized early-Gate effect: at layer 7, Gate drops syntax accuracy by 52.5 pp and semantic accuracy by 30.0 pp relative to the intact baseline, while K and Up stay near baseline; averaged across the three layers, Gate drops syntax by 17.5 pp and semantics by 10.0 pp. On Qwen2.5-7B-Instruct, a deliberately small 20-item mid-layer smoke test at 75% severity shows no accuracy drop for K/Gate/Up. These probes broaden the prompt surface and suggest a higher intervention threshold at scale, but they are not TAB-symptom replications.

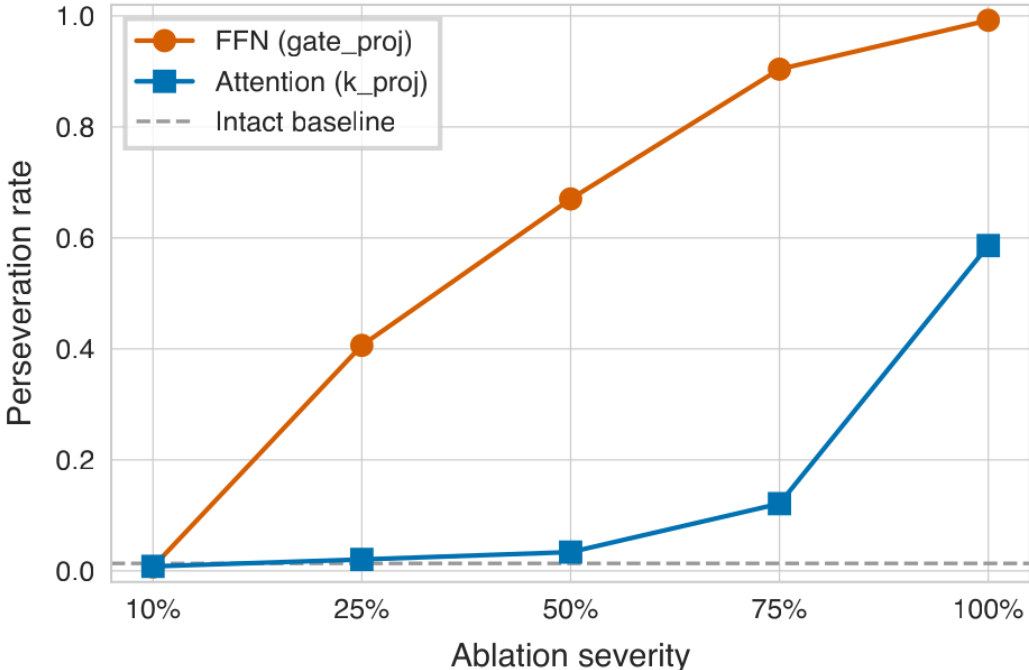


Figure 9: **At 7B, the FFN-vs-attention contrast appears only under multi-layer ablation.** FFN-Gate (orange) reaches a repetition-loop symptom rate of 0.41 at 25% severity; attention-K (blue) stays near baseline until 75%.

## T Matched-random ablation control

Table 30: **Targeted lesions reduce length and lexical diversity; matched random does not.** Gemma-3-1B-IT, layers 0–12, 100% severity.

Component	Condition	Length ↓	Unique ↓	Persev. ↑
k_proj (3.8M)	Targeted	<b>2.9</b>	<b>0.155</b>	<b>0.050</b>
	Random (±std)	27.3 ± 1.0	0.475 ± 0.018	0.000
gate_proj (103.5M)	Targeted	<b>1.0</b>	<b>0.050</b>	0.000
	Random (±std)	36.3 ± 1.6	0.500 ± 0.073	0.017
Intact baseline		27.4	0.471	0.000

To move beyond the surface metrics, we rescored the existing 100% matched-random responses with Gemma 4 E2B on a 10-symptom subset (*Perseverations, Off\_topic, Short and simplified utterances, Meaning unclear, Empty speech, Anomia, Semantic paraphasias, Neologisms, Jargon, Omission of function words*). The intact baseline stays low on this subset (any-symptom rate 0.20; other-like mean 0.025). For attention-K, the targeted lesion still exceeds matched-random on other-like symptoms (0.100 vs. 0.008), driven by off-topic responses (0.100 vs. 0.017) and repetition-loop symptoms (0.100 vs. 0.000; Table 32). For Gate, the targeted outputs collapse to punctuation-only

Table 31: **Targeted and sparsity-matched random lesions differ across K, Gate, and Up.** Gemma-3-1B-IT, layers 0–12, 20 prompts, local full-TAB-style scoring. Random rows average three sparsity-matched draws. Values are targeted / random mean.

Component	Severity	Any symptom $\uparrow$	Length $\downarrow$	Unique ratio $\downarrow$	Other $\uparrow$
K	75%	0.70 / 0.80	18.2 / 25.9	0.475 / 0.492	0.150 / 0.122
K	100%	0.80 / 0.83	2.9 / 26.9	0.155 / 0.484	0.183 / 0.083
Gate	75%	1.00 / 0.88	22.1 / 29.1	0.365 / 0.289	0.800 / 0.595
Gate	100%	0.90 / 0.97	1.0 / 26.5	0.050 / 0.565	0.617 / 0.717
Up	75%	1.00 / 0.88	40.7 / 29.1	0.309 / 0.289	1.000 / 0.595
Up	100%	0.90 / 0.97	1.0 / 26.5	0.050 / 0.565	0.617 / 0.717

strings, which the TAB rubric summarizes poorly; Table 30 is the more faithful summary of that condition.

Table 32: **Gemma 4 E2B rerating preserves the targeted-vs-random gap on attention-K.** 100% attention-K matched-random responses, 10-symptom subset.

Condition	Any $\uparrow$	Other-like $\uparrow$	Off_topic $\uparrow$	Persev. $\uparrow$
K targeted	<b>0.20</b>	<b>0.100</b>	<b>0.100</b>	<b>0.100</b>
K random ( $\pm$ std)	0.167 $\pm$ 0.024	0.008 $\pm$ 0.012	0.017 $\pm$ 0.024	0.000 $\pm$ 0.000
Intact baseline	0.20	0.025	0.050	0.000

## U Combined human and LM TAB-symptom panel

Figure 10 combines the human and model TAB-symptom profiles from Figures 7 and 1. Panels A and B show raw symptom rates for human diagnoses and LM lesion components; panels C and D show deviations from the corresponding corpus mean. The centered panels in C and D use each corpus mean separately, so vertical scales show within-corpus emphasis rather than cross-population comparable rates. The legend groups symptoms by category (Semantic, Syntactic, Fluency, Phonological, Other) so that the color map can be read column-wise.

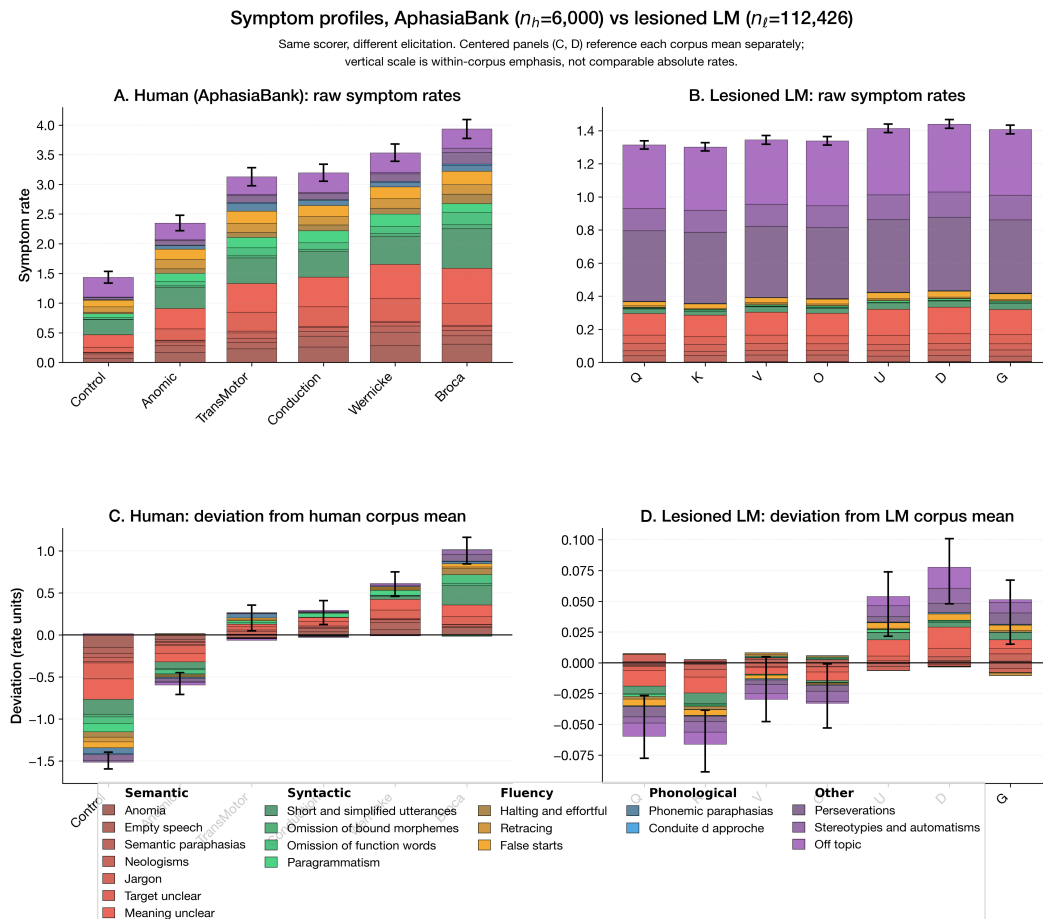


Figure 10: **Combined human and lesioned-LM TAB-symptom panel.** AphasiaBank ( $n_h = 6,000$ ) and lesioned LMs ( $n_l = 112,426$ ). (A, B) Raw symptom rates. (C, D) Deviations from the corresponding corpus mean. Centered panels show within-corpus emphasis only. Colors are shared across all four panels; the legend groups symptoms under their TAB category headers.

## V Likelihood of AphasiaBank productions under model lesions

As a model-native complement to the rubric-based human comparison, we score actual AphasiaBank productions under intact and lesioned Gemma-3-1B-IT and report mean per-token log-probability. Following Wang et al. [2025], we ask whether lesions that drive off-topic and short model outputs also assign lower likelihood to PWA productions, or whether they affect PWA and Control text similarly.

We sample 50 PWA and 50 Control rows from the sampled AphasiaBank interaction file (responses with  $\geq 10$  characters; seed 42). For each row, we tokenize the AphasiaBank prompt with the model’s chat template and compute conditional log-probabilities of the human response token-by-token under three conditions: intact, attention-K at layer 10 with full Bernoulli mask, and FFN-Gate at layer 10 with full Bernoulli mask. Bootstrap 95% CIs use 2,000 resamples and Cohen’s  $d$  [Cohen, 1988] contrasts PWA against Control within each condition.

PWA productions are less likely than Control productions in all three conditions, with a stable Cohen’s  $d$  around  $-0.5$  (Table 33). Lesions raise the mean per-token log-probability instead of lowering it, consistent with a peakier predictive distribution that concentrates mass on common tokens such as articles and copulas; FFN-Gate ablation produces the largest upward shift. The PWA–Control gap is preserved across conditions: lesions do not differentially *disadvantage* PWA text. We read this as evidence that the lesions characterized in the main paper affect distributional fit broadly, not

in a way that mimics the lexical-semantic and morphosyntactic contrasts that distinguish PWA from Control text under this scorer.

Table 33: **PWA productions are less likely than Control under every condition; the gap is preserved across lesions.** Mean per-token log-probability of AphasiaBank responses under Gemma-3-1B-IT with bootstrap 95% CIs (2,000 resamples);  $n = 50$  per population per condition; lesion configurations apply 100% Bernoulli masks at layer 10. Cohen’s  $d$  contrasts PWA against Control within each condition.

Condition	Control	PWA	Cohen’s $d$
Intact	-11.27 [-12.12, -10.48]	-12.99 [-13.85, -12.10]	-0.56
Attn-K, L10, 100%	-10.58 [-11.30, -9.88]	-11.99 [-12.77, -11.18]	-0.52
FFN-Gate, L10, 100%	-8.54 [-9.25, -7.89]	-9.86 [-10.54, -9.16]	-0.54

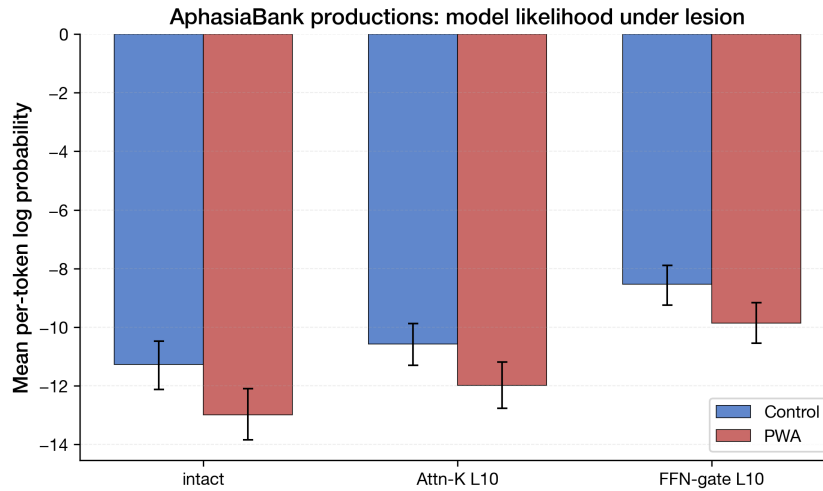


Figure 11: **Lesions shift overall likelihood but preserve the PWA–Control gap.** Mean per-token log-probability assigned by Gemma-3-1B-IT to 50 PWA and 50 Control AphasiaBank responses under three lesion conditions, with bootstrap 95% CIs.

## W Activation patching

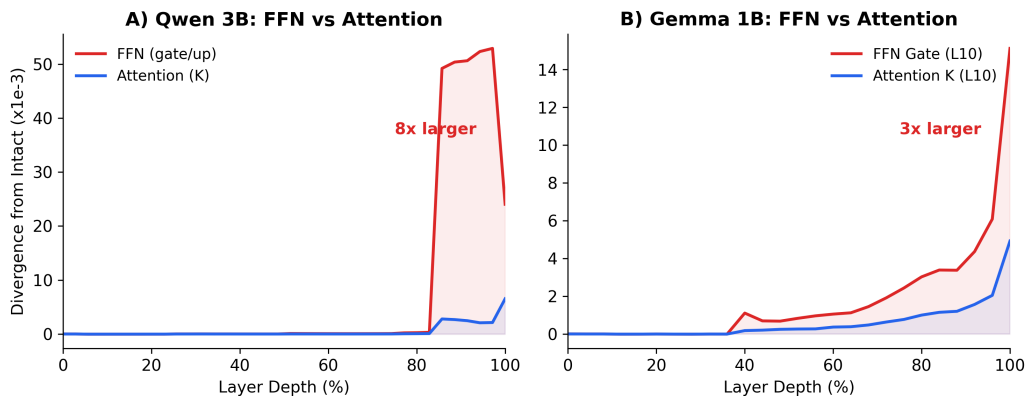


Figure 12: **FFN ablation perturbs the residual stream more than attention-K.** Hidden-state divergence from the intact model after ablation. (A) Qwen 3B bridge probe: about  $10\times$  larger peak for FFN (red) than attention-K (blue),  $p = 3 \times 10^{-6}$ . (B) Gemma 1B bridge probe: about  $2\times$  larger.

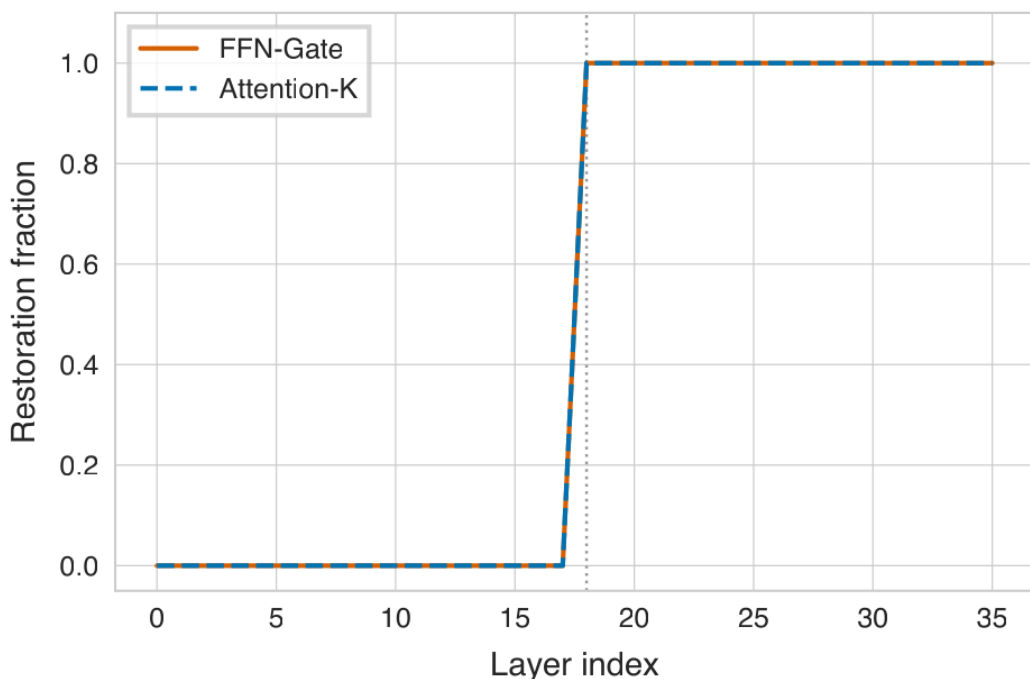


Figure 13: **Gate ablation starts from a larger baseline KL perturbation than attention-K in this small bridge probe.** Activation patching on Qwen 3B (3 prompts; mean  $\pm$  std). Gate starts from a higher baseline KL divergence at the inspected layer, while restoration curves are otherwise similar; this is an auxiliary perturbation-size check instead of evidence for a distinct circuit mechanism.

## X Representational analysis

Mean post-ablation CKA is lower for FFN than for attention-K (0.933 vs. 0.976), which indicates a larger representational shift on this probe; Gate and Up CKA trajectories at the inspected layer nearly coincide, so within-FFN separation is supported by behavior, not by this measure. FFN ablation reduces rank  $\sim 4\times$  more than attention (drop 0.40 vs. 0.10). Coarse linear probes change by  $< 2\%$  under either intervention. We read this as evidence that lesions do not erase linearly decodable features at the resolution of these probes, not as positive evidence for a specific computational mode.

Appendix Table 34 reports the independent-scorer validation slice.

Table 34: **Independent-scorer validation is coarse, not symptom-complete.** Cached rerater slices use local non-primary scorers and no new Gemini calls. They support a coarse component-profile readout while marking individual symptom directions as scorer-dependent.

Check	Coverage	Result	Interpretation
Gemma4 agreement slice	300 outputs; 10 symptoms	93.6% agreement; $\kappa = 0.455$	Supports coarse rerating consistency; not a full replacement for the primary scorer.
Gemma4 component slice	18 balanced model/severity/depth strata	Gemma4 K-vs-FFN $L_2 = 9.86$ pp; Gemini same rows $L_2 = 9.16$ pp	Independent scoring detects a component profile contrast, but the symptom-direction cosine is -0.38; individual coordinates remain scorer-dependent.
Gemma4 mask-seed slice	480 responses; 2 layers, 3 seeds	K-vs-FFN $L_2 = 0.173$ vs. within-component seed $L_2 = 0.135$	A targeted seed check sees a component signal slightly larger than mask-seed variation.
Qwen rerater boundary	500 outputs; 10 symptoms	weighted $\kappa = 0.026$ ; scorer-invariant = no	Non-Google rerating is unstable, so symptom names are descriptive coordinates rather than scorer-invariant clinical categories.

## Y Traceability, scorer sensitivity, and scope

This appendix collects audit tables for the most important boundaries of the paper: which files regenerate the headline numbers, how scorer reliability changes the readout, what scorer-free checks show, and which perturbation-matching or generalization checks have and have not been run. These tables are not additional claims; they make the fixed-assay scope explicit.

Table 35: **Headline claim-to-artifact map.** The anonymous artifact includes generated LM outputs, cached summary files, and the scripts below. Raw AphasiaBank text and proprietary scorer caches are excluded for licensing/API reasons, but derived summaries and final LM symptoms are included.

Claim	Reported quantity	Regeneration path	Boundary
Primary profile split	$L_2 = 3.73, p = .0004$ over 360 paired strata	Profile stress-test script and cache	Fixed Gemini/TAB scorer; 1B realized models; greedy decoding.
Burden-adjusted split	$L_2 = 3.21, p = .0004$	Same stress-test script and cache	Removes overall symptom-rate burden, not all possible surface differences.
Visible-damage matching	$L_2 = 3.25$ after length matching; $4.22$ after length+burden matching	Visible-damage analysis script and cache	Matches observed output proxies, not activation norm.
Output-KL matching	96 same-layer FFN-to-attention matches; median $L_2 = 28.3$ ; median log-KL gap 0.021	Multi-model KL script and cache	Two locally cached models; next-token KL, not generated-output KL.
Residual-state matching	96 same-layer FFN-to-attention matches; median $L_2 = 28.3$ ; median log-residual gap 0.004	Residual-stream calibration script and cache	Two locally cached models; final-token/final-layer hidden-state proxy, not full activation-norm matching.
Joint dose matching	96 same-layer FFN-to-attention matches; median $L_2 = 27.2$ ; median log gaps 0.028/0.023	Joint dose-matching script and cache	Simultaneous proxy match; still not generated-output or layerwise activation matching.
Mask-seed checks	1,600-response surface grid; 480 rerated responses; K-vs-FFN $L_2 = 0.173$ vs. seed $L_2 = 0.135$ in the scored slice	Mask-seed surface and scoring scripts/caches	Surface seed jitter is substantial; scored check remains targeted, not full seed marginalization.
Independent scorer boundary	300-output Gemma 4 rerating plus 500-output Qwen check	Independent-scorer summary script and cache	Coarse component readout has support; exact symptom directions are scorer-dependent.
Scorer-free text features	360 strata; feature $L_2 = 0.295, p < .0003$ ; full-ablation $L_2 = 0.741$	Scorer-free behavior script and cache	Surface trace only; does not score semantic categories.
Length/diversity-conditioned TAB profile	Residualized symptoms; $L_2 = 2.52, p = .0148$ after word count + lexical diversity	Length-conditioned profile script and cache	Does not condition on every possible surface or semantic covariate.
Decoder-sensitive symptom removal	18-symptom vector; $L_2 = 3.20, p = .0002$	Profile robustness script and cache	Removes repetition-loop, stereotypy, and off-topic symptoms as dimensions, not all decoder concerns.
Effect-size calibration	FFN split rank 1/35; leave-model balanced accuracy 0.553	Effect-size calibration script and cache	Aggregate profile signature; weak per-condition classifier.
Human boundary result	6,000 AphasiaBank productions; shared symptoms, different organization	Human-reference analysis script plus manuscript caches	Rubric-matched, not task-matched; raw text requires TalkBank access.

Table 36: **Dependence and effect-size reading.** The headline profile distance is a fixed-assay estimate. Conservative aggregations are included to expose dependence rather than to imply that three model families are enough for a population-level family test.

Reading	Evidence	Interpretation
Fixed-assay profile split	360 paired strata; $L_2 = 3.73$ , $p = .0004$	Primary estimate over the realized prompt/model/decoder/scorer design.
Single-symptom magnitude	Selected FFN-shift symptoms: +1.25 pp; remaining symptoms: +0.08 pp	Modest marginal changes combine into a reliable multivariate signature.
Seed-free anchor	100% ablation; 90 strata; $L_2 = 8.05$ , $p = .018$	The strongest contrast does not depend on a partial-mask draw.
Top-level units	Five model variants and three families align directionally; exact $p = .091$ and $.333$	Encouraging recurrence, but not a population-level family claim.
Prompt dependence	18/20 prompt vectors align with pooled contrast; prompt-level $p < .0001$	The pooled result is not driven by a single prompt family.

Table 37: **Scorer-sensitivity ladder.** The main estimand is a profile-level contrast under a fixed scorer. Individual symptom names are descriptive coordinates, and the rows below show how much of the conclusion survives increasingly conservative scorer views.

Scorer view	Coverage	Result	Interpretation
Fixed Gemini/TAB instrument	21 primary symptoms	$L_2 = 3.73$ , $p = .0004$	Primary fixed-assay result.
Moderate reliability filter	12 symptoms with $\kappa > .30$	$L_2 = 3.50$ , $p = .0002$	Profile split is not driven only by the least reliable symptoms.
Strict reliability filter	11 symptoms with $AC1 > .65$	$L_2 = 1.54$ , $p = .052$	Direction survives, but strict filtering is borderline.
Scorer-free feature profile	10 deterministic text features	$L_2 = 0.295$ , $p < .0003$ ; 100% subset $L_2 = 0.741$	Shows a surface-level component trace without any TAB symptoms or LLM scorer.
Length/diversity-conditioned profile	21 TAB symptoms after residualizing generated-text features	$L_2 = 2.52$ , $p = .0148$ after word count + lexical diversity; $L_2 = 2.10$ , $p = .044$ after adding repetition mass	Shows the TAB profile split is not only shorter or less diverse text.
Decoder-sensitive symptoms removed	18 TAB symptoms	$L_2 = 3.20$ , $p = .0002$	The profile split is not only repetition-loop, stereotypy, or off-topic symptoms.
Gemma 4 rerater subset	300-output, 10-symptom subset plus 480-response seed slice	Agreement 93.6%, $\kappa = .455$ ; K-vs-FFN slice $L_2 = 9.86$ pp; seed slice $L_2 = 0.173$	Supports coarse profile readout; symptom directions differ from Gemini on the balanced slice.
Qwen rerater subset	Targeted subset	10-symptom	Less stable, especially off-topic/repetition-loop symptoms
			Main reason symptom-specific clinical readings stay descriptive.

Table 38: **Perturbation-matching ladder.** Each row asks whether the attention-vs-FFN profile split could be a proxy for generic damage magnitude. Later rows are stricter, but none is a full generated-output-matched causal estimate.

Control	What is matched	What it shows	Remaining gap
Matched random ablation	Number of zeroed parameters in targeted Gemma slice	Targeted component lesions change length/diversity more than random sparsity-matched lesions.	Does not match functional perturbation size.
Visible-damage matching	Output length and/or total TAB symptom burden within strata	Profile shape remains at similar visible output damage.	Surface proxies may miss internal disruption.
Surface-feature conditioning	Word count, lexical diversity, and repeated-token mass residualized from TAB symptoms	Profile distance remains positive after conditioning.	Residualization is descriptive, not a matched causal estimator.
Next-token KL matching	Intact-vs-lesioned next-token distributions on TAB prompts	FFN profiles remain distinct from attention profiles across 96 same-layer matches.	Next-token KL is not full generated-output KL.
Residual-state matching	Final-token hidden-state change on TAB prompts	FFN profiles remain distinct from attention profiles across 96 same-layer matches with median log-gap 0.004.	This is a proxy, not full layerwise activation-norm matching.
Joint KL+residual matching	Both proxy distances simultaneously	FFN profiles remain distinct across 96 same-layer matches with median $L_2 = 27.2$ pp.	Still not full generated-output or layerwise activation matching.
Not yet done	Generated-output KL and full layerwise activation-norm matching over all models/severities	–	Strongest remaining perturbation-matching extension.

Table 39: **Scope of generalization.** The paper’s language should be read as a fixed-assay claim with broad within-assay recurrence, not as population-level proof over arbitrary prompts, scorers, scales, or model families.

<b>Dimension</b>	<b>Evidence in this paper</b>	<b>Not claimed</b>
Prompt surface	20 TAB prompts; 18/20 prompt-level vectors align with pooled contrast; connected and non-connected subsets remain positive; Qwen 3B/7B minimal-pair probes broaden the prompt surface.	Prompt-universal behavioral profile over arbitrary tasks.
Model families	Five realized 1B variants across Llama, Gemma, and OLMo align directionally; exact top-level tests are underpowered.	A population-level family effect or broad scale law.
Severity and masks	Full severities in main sweep; deterministic 100% subset strengthens effect; 75% seed checks include a 1,600-response surface grid and 480 scored responses.	Full marginalization over partial-mask seeds.
Decoding	Greedy is primary; targeted sampling/repetition-penalty checks preserve FFN profiles and reduce attention-K repetition loops.	Decoding-invariant syndrome organization.
Human comparison	AphasiaBank gives a common symptom vocabulary and a boundary test for co-occurrence structure.	Task-matched clinical equivalence or aphasia subtype mapping.

## Z Experimental details

For each configuration, the target component’s weight tensor is cloned, ablated with element-wise masking, used for generation, then restored. The released driver supports CPU, CUDA, and Apple MPS. It derives deterministic per-condition mask seeds from the model, component, layer, severity, strategy, and user-supplied base seed; restores weights even if generation fails; and marks generation or scorer failures rather than converting them to absent symptoms.

The anonymous supplemental ZIP contains a checksum manifest, generated LM outputs and symptoms, table-regeneration scripts, cached summary files, and the manuscript source. The primary scored LM table and derived human-reference summaries are included where licensing permits; raw AphasiaBank-derived text is excluded from the artifact and must be reconstructed under Talk-Bank access rules. The included scripts regenerate the condition-level profile summaries, robustness filters, dependence-sensitive uncertainty summaries, visible-damage and clustered human checks, scorer-free features, length/diversity conditioning, effect-size calibration, independent-scorer validation, perturbation-proxy calibrations, replacement-value controls, evidence summaries, and mask-seed checks. The README lists the short regeneration sequence for the main analysis tables and PDF. The exact prompt subset is listed in Appendix J; model identifiers and decoding settings are in Section 3.1; scorer validation and rerater summaries are in Appendix K and Table 34. Primary generation used the HuggingFace causal-language-model loader with batch size 20 where memory allowed, greedy decoding, and an 80-token generation cap. Scorer scripts used Gemini 2.5 Flash, temperature 0.01, SHA256 response caching, retry logic, and five LM-scoring workers. The runnable dependency specifications cover both cached no-API regeneration and the broader generation/scoring environment.

There are three reproduction tiers. First, the main cached analyses are exactly reproducible: included output/symptom tables regenerate manuscript statistics and generated tables, and the PDF rebuilds from included checksummed figure files without model downloads or API calls. Second, fresh generation with the released driver is deterministic going forward because the script records the base seed, mask seed, and device for new runs. Third, exact fresh regeneration of the historical partial-mask rows in the full-run checkpoint is not bitwise guaranteed because that checkpoint predates per-condition mask-seed logging; deterministic 100% ablations and the targeted seed experiments do not have this ambiguity. Generated LM outputs, symptoms, targeted seed metadata, and analysis scripts are included where source-model licenses permit. The included outputs allow independent re-judging of the same generations with another scoring model without rerunning the lesion sweep. Proprietary scorer caches, API credentials, model weights, and raw AphasiaBank text are not redistributed. Scripts read scorer credentials and optional gated-model HuggingFace tokens from environment variables rather than embedded secrets.

Table 40 reports the recoverable compute metadata. The main 1B sweeps and most targeted appendix checks used local Apple Silicon via PyTorch MPS on an M-series Max laptop with 128 GB unified memory; the released driver also supports CUDA and CPU. Table-level statistical analyses run in seconds to minutes on a laptop CPU. Exact consolidated wall-clock and peak-memory logs for the original full 1B sweep were not preserved, so we report the worker/device category and dataset scale rather than implying precision the logs do not support. Repeating the full scorer pass would require API quota for at least the 112,426 severity > 0 generated outputs plus intact/common-table rows, but the cached output/symptom tables regenerate the main statistics and generated tables, then rebuild the PDF from included figures, without API calls. Targeted checks report elapsed seconds where available.

Table 40: **Compute/resource audit.** We report exact elapsed time only when the generating script wrote it to the recovered JSON artifact; otherwise the entry records the worker/device category and scale needed to rerun the analysis.

Analysis	Worker/device	Scale	Recovered runtime
Primary 1B lesion sweep	Apple Silicon M-series Max MPS, FP16, 128 GB unified memory; Gemini scorer cache/API	112,426 severity > 0 scored records; 2,528 nonzero condition profiles	Consolidated wall-clock and peak memory not logged
Condition-level profile tests	Laptop Python/pandas	CPU, 360 paired model $\times$ layer $\times$ severity strata; 5,000 bootstrap/permutation draws; length/diversity residualization	Length-conditioned check: 2.4 s; other table checks seconds to minutes; no accelerator needed
Visible-damage and clustered human checks	Laptop Python/pandas	CPU, 360 matched strata; 1,080 FFN-attention matches; 5,000 corpus/model bootstraps	Minutes; no accelerator needed
Output-KL/residual calibration slices	Gemma-3-1B-IT and Llama-3.2-1B on Apple Silicon MPS	Single-model K/Gate/Up KL slice plus two-model Q/K/V/O/Gate/Up/Down KL and residual-state slices, four layers/model, four severities, 20 prompts; joint matching reads the two cached proxy files and TAB profiles	KL Gemma-only: 185.8 s; KL two-model rerun: 65.7 s; residual two-model rerun: 104.4 s; joint synthesis: seconds
Replacement-value control	Gemma-3-1B-IT on Apple Silicon MPS	K/Gate/Up/Down, layers 5/10/15/20, severities 75% and 100%, zero/global-mean/row-mean/column-mean replacement, 20 prompts; scorer-free features and next-token KL	128 conditions; 2,560 generations; no scorer/API calls
Matched-random control	Gemma-3-1B-IT on Apple Silicon MPS	26 layers, 20 prompts, targeted and sparsity-matched random masks	1,322.9 s
Decoding grid	Gemma-3-1B-IT on Apple Silicon MPS plus scorer calls	25 decoding configurations, 13 conditions/configuration, 20 prompts	6,429.5 s
AphasiaBank likelihood	Gemma-3-1B-IT on Apple Silicon MPS	50 PWA and 50 Control productions across three conditions	13.7 s
Scaling/bridge checks	Qwen2.5-7B-Instruct and Qwen2.5-3B-Instruct on Apple Silicon MPS where logged	7B: 12 TAB conditions plus 20-item minimal-pair smoke; 3B: 98-condition ablation slice, hidden-state probes, and 70-item minimal-pair check	7B TAB slice: 1,010.0 s; 7B minimal-pair smoke: seconds after load; 3B minimal-pair check: seconds-to-minutes after load; 3B bridge wall-clock not logged
Mask-seed checks	Cached Gemma-3-1B-IT generation plus local Gemma 4 E2B rerater on Apple Silicon MPS	Surface: five seeds, K/Gate, eight prompts; broader surface grid: four layers, five seeds, K/Gate/Up/Down, 1,600 responses; scored: layers 5/10, three seeds, K/Gate/Up/Down, 480 responses	Broader surface grid: 2,325.5 s; expanded scored check: 3,224.8 s; small surface-only wall-clock not logged

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction state four claims: (i) attention vs. feed-forward lesions separate in an operational TAB-symptom behavioral space, (ii) the contrast is supported by restricted profile permutations, an all-symptom paired profile test, deterministic 100% ablation, targeted matched-random, decoding, intact-subtracted, scorer-stress, visible-damage-matching, and condition-level bootstrap controls, (iii) component profiles recur directionally across the realized families and tuning comparisons while depth effects are descriptive, and (iv) human AphasiaBank productions provide a contrast distribution with shared symptoms but divergent mixtures and co-occurrence structure. Each is supported by Results and qualified in Discussion.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Discussion covers measurement (shared Gemini scorer; per-symptom reliability; TAB as operational text-behavior measurements, not diagnoses), design and assay (greedy decoding can amplify perseveration; one fixed mask per nonzero condition; targeted 480-response mask-seed check instead of full-corpus seed marginalization; co-occurrence is decoder-dependent; AphasiaBank is rubric- not task-matched), and external validity (matched-random scope, scaling and tuning heterogeneity, dense English decoder-only models, and bridge-model state-level analyses).

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This is an empirical study. No theorems or formal proofs are presented.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 specifies HuggingFace model identifiers, the Bernoulli weight-zero ablation procedure with a fixed mask per nonzero condition, severity levels, decoding parameters, the Gemini 2.5 Flash scoring configuration, and references the exact 20-prompt TAB subset. The Appendix separates exactly reproducible cached analyses from fresh generation/scoring, notes that historical partial-mask regeneration is not bitwise guaranteed because the primary checkpoint predates per-condition mask-seed logging, and documents the deterministic seed behavior in the released driver.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The anonymous supplemental artifact includes analysis scripts, generated LM outputs and symptoms where source-model licenses permit, targeted seed metadata, cached summary tables, all manuscript build assets, a pinned cached-analysis environment file, and a SHA256 manifest that is rebuilt after table/PDF regeneration. Human data from AphasiaBank requires a separate license from TalkBank and cannot be redistributed; proprietary scorer caches and API credentials are also excluded.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 reports model identifiers, realized coverage, inference parameters (greedy decoding, FP16, max tokens), ablation severity levels, component definitions, and scorer settings. The Appendix details statistical methods and decoder/scaling sensitivity checks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Bootstrap 95% CIs and selection-corrected  $p$ -values are reported for exploratory human-reference cosine summaries (Appendix P). A paired condition-level bootstrap, all-symptom sign-flip test, restricted within-stratum permutation, and intact-subtracted sensitivity are reported for the main attention/FFN contrast (Appendix B). Permutation tests are reported for component-profile separation, depth coherence, and Mantel co-occurrence. FDR correction is applied to category-level prevalence comparisons and symptom-level component effects.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix Table 40 reports recovered worker/device categories, memory category, dataset scale, and exact elapsed seconds for targeted checks whose scripts wrote runtime metadata. The surrounding reproducibility manifest reports the generation script, batch size, scorer model/settings, cache/retry behavior, worker count, dependency specification, pinned cached-analysis environment, and artifact checksum manifest. The table also states the remaining gap honestly: exact consolidated wall-clock and peak-memory logs for the original full 1B sweep were not preserved, so the manuscript reports the local Apple Silicon M-series Max MPS/128 GB unified-memory setting, run scale, and scorer-call scale instead of implying precision the logs do not support; cached tables reproduce the statistics without rerunning that sweep.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The study uses publicly available models and a licensed research corpus (AphasiaBank). No human subjects were recruited. The TAB is a research benchmark, not a clinical diagnostic tool.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Discussion describes the positive interpretability use of component-lesion profiles and cautions against misinterpretation: TAB symptoms are operational text-behavior measurements, not bedside diagnoses; weight ablation is not homologous to brain damage; lesion phenotypes are model-specific.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper analyzes publicly available open-weight models under their respective licenses and does not release new models or high-risk datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models are cited with HuggingFace identifiers. AphasiaBank is cited and accessed under its TalkBank license. The TAB is cited as prior work.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The lesion corpus is a new asset and is documented in the paper, with a prompt manifest, row-unit audit, and anonymous supplemental artifact containing generated LM outputs, symptoms, targeted seed metadata, cached summaries, analysis scripts, and a checksum manifest where licenses permit. Human AphasiaBank data cannot be redistributed and must be obtained through TalkBank.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or human subjects research was conducted. Human data comes from the existing AphasiaBank corpus. Expert SLP annotations for TAB validation are described in the TAB paper.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects were recruited. AphasiaBank data was collected under IRB protocols at contributing institutions.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Gemini 2.5 Flash is used as an automated scorer for TAB symptom classification (Section 3, Assessment). This usage is a core methodological component, described in detail in the TAB paper and summarized in Section 3 and the Appendix.

**Guidelines:**

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.