Artificial Aphasias in Lesioned Language Models

Nathan Roll[1] **Jill Kries**[2,3] **Laura Gwilliams**[2,3] **Cory Shain**[1]

Department of {[1]Linguistics, [2]Psychology}, Stanford University
[3]Wu Tsai Neurosciences Institute, Stanford University

nroll@stanford.edu

Abstract

Human aphasia motivates a detailed way to describe language-model failures: measure the kinds of errors that follow an intervention, rather than only its effect on accuracy or loss. We apply the previously released Text Aphasia Battery (TAB) to a weight-lesion sweep centered on five approximately 1B-parameter model variants, with 112,426 scored lesion-output records. TAB labels are operational descriptions of generated text, not clinical diagnoses. Attention- and feed-forward-targeted lesions produce a modest pooled profile shift ($L_2 = 3.73$ percentage points), with weak prediction of component class on held-out models and scorer-dependent individual label directions. Comparing the outputs with 6,000 AphasiaBank productions reveals substantial differences in label prevalence and co-occurrence, alongside partial, decoder-dependent overlap. The resulting framework profiles lesion-induced failures in the evaluated small models and makes the distinction between shared error labels and shared error organization empirically testable.

1 Introduction

Neuropsychological studies use selective language impairments, or *aphasias*, to relate brain damage to patterns of behavior [Broca, 1861, Wernicke, 1874, Goodglass, 1993]. Clinical batteries distinguish errors in meaning, grammatical form, fluency, and other dimensions [Kertesz, 2007, Goodglass et al., 2001, Swinburn et al., 2005, Wilson et al., 2018]. This motivates a useful question for language-model (LM) evaluation: after an intervention, *how* does generated language fail?

Intervening on model weights or activations is already central to mechanistic interpretability [Michel et al., 2019, Voita et al., 2019, Geiger et al., 2021, Meng et al., 2022]. Many evaluations summarize an intervention through accuracy, loss, or task performance. Here we retain a vector of text-behavior labels, using the previously released Text Aphasia Battery [TAB; Roll et al., 2025]. We then compare label prevalence and co-occurrence across model components and with human productions from AphasiaBank [MacWhinney et al., 2011]. These are causal interventions on weights with a behavioral readout; they do not by themselves identify the circuits implementing a function.

We make three study-design choices. We profile individual labels rather than assigning models to coarse aphasia subtypes, whose boundaries can be unstable even in humans [Wilson et al., 2023]. We examine divergence as well as similarity between human and LM outputs. And we restrict the comparison to observable text. Throughout, *aphasia-like* means surface similarity in this label space, not clinical diagnosis, shared cognitive processes, or biological equivalence. Human-derived labels are a common coordinate system, not an exhaustive taxonomy of LM errors.

Our empirical contribution is a projection $\times$ depth $\times$ severity sweep that retains the prevalence and co-occurrence of generated-language failures without selecting a target syndrome in advance. It yields a small pooled attention/FFN profile difference across the evaluated approximately 1B-scale variants. The human comparison shows why coarse matches can be misleading: overlapping labels

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

can have different prevalences and combine differently. The depth analyses and larger-model probes are exploratory. We do not infer a general early-to-late linguistic hierarchy or use these experiments to constrain theories of human aphasia.

2 Related work

Lesioning networks to study language predates transformers. Hinton and Shallice [1991] modeled deep-dyslexia-like errors, and Plaut and Shallice [1993] showed that graded damage can produce crossed deficits without separate modules. Modern work causally masks attention heads [Michel et al., 2019, Voita et al., 2019], relates attention to syntactic dependencies [Clark et al., 2019], and analyzes lexical and conceptual information in FFN layers [Geva et al., 2022]. Probing also reveals depth-dependent linguistic representations [Tenney et al., 2019]. Attention/FFN differences and depth effects are therefore not architectural discoveries of this paper. Our contribution is to characterize their behavioral expression as failure profiles rather than a single performance score.

Recent LM-aphasia studies report similarities to human aphasia categories [Wang et al., 2025, 2026]. Fridriksson et al. [2026] compare layer-perturbed model errors on clinical tasks with human lesion maps. In post-submission work, Yang et al. [2026] search perturbation configurations to match individual picture-naming error profiles. Those positive-alignment results concern different tasks and estimands: optimized matching to a specified human target can coexist with divergence in a target-agnostic free-generation comparison. We retain both label prevalence and co-occurrence, including differences that a coarse subtype match can obscure.

Two cautions shape this interpretation. Removing weights can create states outside the training distribution [Hase et al., 2021]; our lesions are stress tests, not direct readouts of normal component function. And human-derived tests need not transfer unchanged to LM errors [Mancoridis et al., 2025]. TAB gives a shared description of text behavior, while loops, prompt echo, drift, malformed strings, and punctuation-only output may require a richer LM-native taxonomy.

3 Methods

3.1 Experimental setup

The full sweeps cover five approximately 1B-parameter decoder-only variants: Llama 3.2-1B and its instruction-tuned variant, Gemma 3-1B-Instruct, and OLMo-2-1B with its instruction-tuned variant [Meta, 2024, Gemma Team et al., 2025, Team OLMo et al., 2025]. Gemma has 26 layers; Llama and OLMo have 16. We intervene on the Q(uery), K(ey), V(alue), and O(utput) attention matrices and the Gate, Up, and Down FFN matrices. For target weights W, a lesion is an unscaled element-wise zero mask:

$$\widetilde{W} = W \odot M, \qquad M_{ij} \sim \text{Bernoulli}(1-s), \qquad s \in \{0, .25, .50, .75, 1\}.$$

Here s is the expected zeroed fraction. Partial lesions use one sampled mask per condition; intact and complete lesions are deterministic. We do not rescale surviving weights. Replacement-value, mask-seed, and perturbation-proxy checks appear in Appendices D.1, M, and D.

Generation uses a fixed 20-prompt TAB subset spanning connected text, word comprehension, sentence comprehension, and repetition (Appendix J). Unless stated otherwise, decoding is greedy in FP16, with one beam, no sampling or repetition penalty, and at most 80 new tokens. The recovered nonzero-lesion table contains 112,426 scored records, including repeated generations and 260 records from an auxiliary, incomplete Gemma-base slice. Prompt deduplication yields 44,527 rows and 2,528 component-condition profiles. The paired attention/FFN analysis uses 360 model $\times$ layer $\times$ severity strata from the five full-sweep variants; the auxiliary base-only slice has no paired attention condition. Appendix A.1 distinguishes these denominators.

For the human comparison, we score 6,000 AphasiaBank productions: 1,000 each from Anomic, Broca's, Conduction, Transcortical Motor, and Wernicke's groups, plus 1,000 controls, drawn from 30 corpora (Appendix A.2). Human and model data share a scoring rubric, not identical elicitation tasks. Appendix N narrows the comparison to narrative/connected-text productions.

3.2 Assessment and operational definitions

TAB is a previously released text-native assessment adapted from the Quick Aphasia Battery [Roll et al., 2025, Wilson et al., 2018]. A *TAB label* is a binary text-behavior judgment; a *label rate* is its prevalence within a condition; a *profile* is the vector of those rates. A *lesion condition* specifies model, layer, component, and severity. The rubric has 21 labels in five categories: Semantic, Syntactic, Fluency, Phonological, and Other. The descriptive component plot, human/LM prevalence and co-occurrence, scorer-robustness, and raw-depth comparisons use 19 labels, excluding the two pause labels because text alone does not reliably reveal within- or between-utterance pauses. Other comprises Perseverations, Stereotypies and automatisms, and Off-topic.

Gemini 2.5 Flash performs scoring at temperature .01 with a hashed response cache [Comanici et al., 2025]. Prior TAB validation compared automatic and expert scores on 561 human and LM texts [Roll et al., 2025]; that does not make every label scorer-invariant. Table 1 gives actual cached outputs before the aggregate analyses. The excerpts illustrate measurements, not distinct component syndromes. The Gate example also exposes a rubric blind spot: a degenerate repeated string can receive no cached positive label. Figure 1 shows why decoding sensitivity must also be considered.

Table 1: **Representative cached outputs.** Gemma-3-1B-IT on the same scene-description prompt ("The boy is pushing the girl); lesions are complete. Line breaks are normalized and excerpts are truncated. These individual examples do not estimate population-level separation.

Condition	TAB labels	Output excerpt
Intact	None	The girl is looking down at him. The girl is holding a small, blue toy.
K, layer 12	Perseverations	The boy is pushing the girl. The boy is pushing the girl. The boy is pushing the girl…
Up, layer 12	Perseverations; Short/simplified	The girl is looking down, and she is crying. The boy is pushing the girl. The girl is looking down, and she is crying…
Gate, layer 0	No cached positive labels	the. the. the. the. the. the. the. the. the. the.…

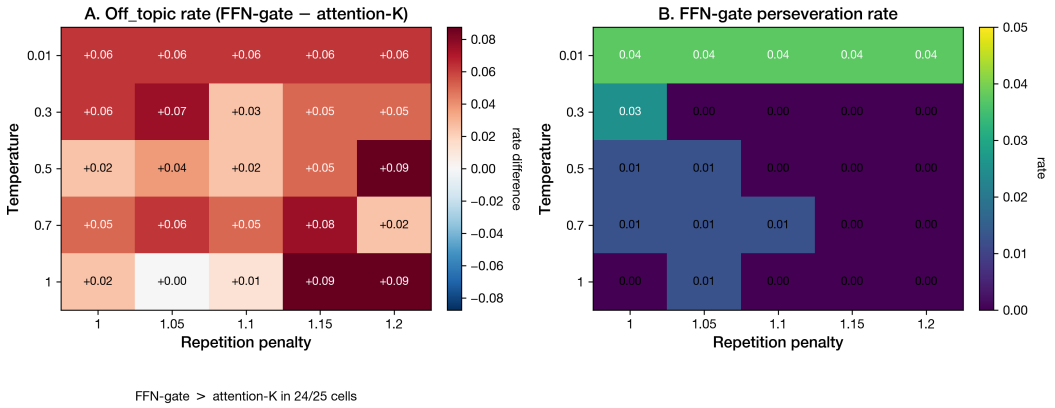


Figure 1: **Decoding changes the observed failure profile.** Gemma-3-1B-IT temperature/repetition-penalty grid. (A) Gate-minus-K off-topic prevalence is positive in 24/25 settings. (B) Gate repetition-loop labels diminish with repetition penalties and reach zero at $T = .7$, penalty 1.2. This targeted check supports a cross-decoder contrast, not invariant individual label rates.

3.3 Statistical methods

We average labels within prompt-deduplicated component conditions, then compare attention and FFN profiles within matched model × layer × severity strata. The primary comparison uses all 21 labels. The pooled L_2 distance is the norm of the mean FFN-minus-attention profile vector, expressed in percentage points. Confidence intervals resample the 360 strata 5,000 times; paired

sign-flip tests assess the mean contrast. Restricted component-name permutations, coarser aggregation, and effect-size calibrations are in Appendices B and F. For human/LM contrasts, we use the corpus/model-clustered bootstrap in Appendix C. Table 2 separates the roles of the analyses.

Table 2: **Claim-to-evidence summary.** Effect sizes accompany significance tests; supporting analyses have narrower scope than the main fixed-assay comparison.

Role	Evidence	Interpretation
Primary	Pooled $L_2 = 3.73$ pp, 95% CI [2.45, 5.66]; $p = .0004$	A modest attention/FFN shift in the fixed assay; adjacent-severity distance is 3.68 pp.
Validation	Held-out balanced accuracy .553; strict-label $p = .052$; independent-scorer cosine $-.38$	Weak per-condition prediction; individual label directions are scorer-dependent.
Controls	Decoding, burden, matched damage, replacement values, mask seeds	Targeted checks address particular alternatives, without full functional-dose matching.
Human comparison	Clinician-rated pairs/templates and corpus-level prevalence/co-occurrence	Shared labels with different organization; partial overlap depends on decoding.
Exploratory	Depth, 3B/7B probes, selected profile cosines	Hypotheses for follow-up, not a general hierarchy or cross-scale replication.

4 Results

4.1 Attention- and FFN-targeted interventions yield a modest pooled profile shift

Figure 2 displays broad overlap across all seven components. Overall label burden is the mean number of positive labels per response. Across all 21 labels, FFN-lesioned outputs have a slightly higher burden: at least one label appears in 59.4% of FFN responses and 56.8% of attention responses, with means of 1.445 and 1.350 labels per response. Differences in profile shape are also measurable. The pooled all-label distance is $L_2 = 3.73$ percentage points (95% CI [2.45, 5.66], paired sign-flip $p = .0004$), close to the median adjacent-severity distance of 3.68 points. Complete ablation yields $L_2 = 8.05$ points (CI [4.45, 14.46], $p = .018$).

Under the primary scorer, FFN-minus-attention differences include Meaning unclear (+2.20 pp), Stereotypies and automatisms (+1.70), Off-topic (+1.58), Perseverations (+1.23), Short and simplified utterances (+.92), and Target unclear (+.87). These are descriptive directions in this assay. They do not establish separate syndromes for Q, K, V, O, Gate, Up, and Down. Leave-one-model-out classification is weak (balanced accuracy .553; AUROC .582), despite the reliable pooled contrast (Appendix F).

The contrast survives several targeted checks: removing Other-category labels gives $L_2 = 3.20$ points ($p = .0002$), and deterministic surface features without an LM scorer give a standardized-profile distance of .295 ($p < .0003$). But scorer dependence matters. Independent and primary reraters give similar K-versus-FFN distances (9.86 and 9.16 points) while disagreeing on the direction of the profile difference (cosine $= -.38$). On the strict 11-label reliability subset, $L_2 = 1.54$ points and $p = .052$. Thus the evidence is stronger for an aggregate difference under specified measurement choices than for invariant label-level signatures (Appendix K).

4.2 Coverage across the evaluated variants; limits at larger scale

The descriptive FFN-shift direction recurs in 5/5 full-sweep variants, 18/20 prompts, and 77/97 model–prompt cells. This summarizes the realized design, not independent replications across a model population. Raw profile cosines across families are high (means .936 for attention and .944 for FFN), but common high-prevalence labels can drive similarity. Instruction tuning reduces symptom burden by 81%, while base/instruct profile shapes remain correlated ($\bar{r} = .857$). These comparisons do not establish scale-general transformer organization.

The 3B and 7B probes are exploratory. The single-layer 7B slice (K/Gate/Up, layers 10/20, severities 75/100, 20 prompts) was not TAB-scored or matched for functional disruption. Its multi-layer .406/.021 Gate/K values are *mean repeated-subword-trigram fractions*, not clinical or TAB symptom rates. The 3B minimal-pair probe shows a localized early-Gate vulnerability, while a small 7B

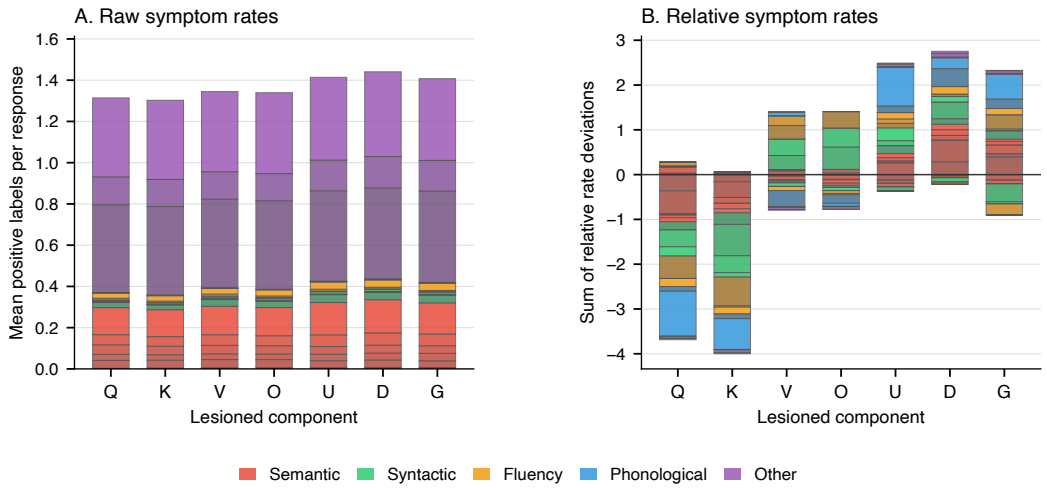


Figure 2: **Component profiles overlap, with a modest pooled attention/FFN shift.** (A) Pooled raw rates for 19 labels (pause labels excluded); bars sum label prevalences and may exceed 1 because labels can co-occur. (B) Each component's label rate divided by that label's all-lesioned-output mean, minus one, with positive and negative deviations stacked separately. These are descriptive summaries; uncertainty for the matched attention/FFN contrast is reported in the text.

minimal-pair test shows no accuracy drop. Appendix S reports these outcomes directly. A cross-scale test would need to match functional disruption, such as output KL or residual-state change, rather than nominal mask percentage alone.

4.3 Depth-dependent differences are descriptive, not a linguistic hierarchy

Layer position changes observed profiles. Figure 3 shows unnormalized category-level prevalence with uncertainty for the five full-sweep variants at 75% and 100% severity. The normalized per-label view is retained in Appendix H for exploration. Row-wise min–max normalization makes a rare label's small fluctuation as bright as a common label's large change, so it cannot establish an early-to-late ordering of linguistic functions. Fluency and Phonological labels have mean all-severity prevalences of 1.4% and .2%, respectively; these are not the severe-lesion subset plotted in the figure. Their normalized peaks can be sparse or localized to one layer. We therefore do not infer a general semantic/syntactic-to-phonological/fluency progression.

4.4 Human and LM labels differ in prevalence and co-occurrence

AphasiaBank and LM outputs share labels without sharing their full distribution. Meaning unclear appears in 45.2% of human productions versus 14.4% of LM outputs, anomia in 22.2% versus .3%, and empty speech in 14.1% versus .2%. LMs show more repetition-loop labels (43.4% versus 10.9%) and off-topic labels (39.4% versus 31.6%). Appendix Table 5 reports mean *per-label prevalence within each category*, not shares of all flags. Narrative-only and corpus/model-clustered comparisons retain the broad direction of these differences (Appendices N and C).

The clinician-rated 561-text subset provides a comparison independent of automatic rerating. We quantify co-occurrence with ϕ, the Pearson correlation between two binary label indicators: 33 human label pairs exceed $\phi = .2$, versus three LM pairs. Fifty of 306 human samples match the syndrome-template rule from Roll et al. [2025], versus none of 255 LM samples (Fisher exact $p \approx 1.1 \times 10^{-14}$). These templates summarize label patterns; they do not diagnose a model. In the larger automatically scored table, co-occurrence matrices nevertheless align partially under greedy decoding (Mantel $r = .373$, $p = .0005$). This alignment does not persist under nucleus sampling ($r = .035$, $p = .36$; Appendix O). The supported result is a difference in observed label organization with partial, decoder-dependent overlap, not universal human/LM separation.

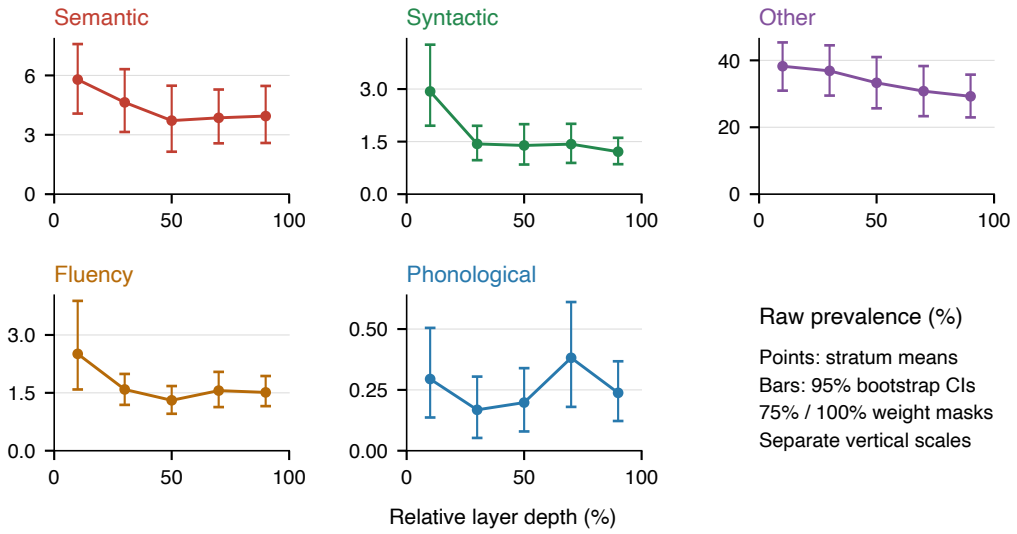


Figure 3: **Raw label prevalence across relative depth.** Using the shared 19-label inventory, each point averages within-category rates across component conditions, then across model × layer × severity strata in a 20-percentage-point depth bin. Depth is normalized within each model; masks are 75% or 100%. Bars are 95% intervals from 5,000 within-bin stratum bootstrap resamples, conditional on the five full-sweep variants and fixed prompts. Vertical scales differ and all start at zero. No row-wise normalization is applied.

5 Discussion

The useful transfer from aphasiology is a measurement strategy: retain the form and combination of errors after an intervention. In the evaluated small models, this reveals a modest attention/FFN shift and a human/LM comparison that coarse subtype labels can miss. It complements prior causal ablation and probing work rather than establishing a new theory of projection specialization or depth. Behavioral similarity and mechanistic similarity remain separate questions.

Because TAB scores generated text, the same profiling approach could be applied to steering, adversarial prompting, or other interventions. These are future applications; this paper tests weight masking. A richer inventory of LM-native failures and scorer-independent validation would improve that extension. Larger-model comparisons additionally require matched functional disruption. None of these directions is evidence that the current experiments identify a human-like lesion mechanism.

6 Limitations

The main results concern five approximately 1B-scale variants, 20 fixed English prompts, and mostly greedy decoding. One historical mask draw per partial-lesion condition is not full marginalization over mask seeds. Targeted seed, replacement-value, decoding, visible-damage, and internal-perturbation checks address particular alternatives, but do not equate functional damage across all components, layers, models, or scales. Historical partial-mask outputs are not guaranteed bitwise reproducible; the artifact distinguishes cached reproduction from fresh generation.

Gemini is a fixed measurement instrument, not ground truth for each label. Strict reliability filtering and independent rerating narrow the label-level claims. Human data are rubric-matched rather than task-matched; narratives reduce but do not remove this difference. TAB labels can miss LM-native failures and should not be read as clinical diagnoses or used to infer human neurobiology. Raw AphasiaBank text is not redistributed. Our causal weight interventions support claims about generated behavior under this design, not an identified circuit mechanism or the normal function of each projection.

Acknowledgments and disclosures

We thank the Stanford NLP Group. This work received no external funding and was conducted independently of Oruk. Nathan Roll is a co-founder of Oruk, Inc.

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.

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.

Julius Fridriksson, Roger D. Newman-Norlund, Saeed Ahmadi, Regan Willis, Nadra Salman, Kalil Warren, Xiang Guan, Yong Yang, Srihari Nelakuditi, Rutvik Desai, Leonardo Bonilha, Jeff Charney, and Chris Rorden. Stroke lesions as a rosetta stone for language model interpretability. *arXiv preprint arXiv:2602.04074*, 2026. doi: 10.48550/arXiv.2602.04074.

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.

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.

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.

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.

Marina Mancoridis, Bec Weeks, Keyon Vafa, and Sendhil Mullainathan. Potemkin understanding in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 42857–42881, 2025. URL https://proceedings.mlr.press/v267/mancoridis25a.html.

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.

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.

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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019. doi: 10.18653/v1/P19-1452.

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.

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.

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.

Yong Yang, Xiang Guan, Sophie Arheix-Parras, Saeed Ahmadi, Roger Newman-Norlund, Leonardo Bonilha, Christopher Rorden, Julius Fridriksson, Rutvik H. Desai, and Srihari Nelakuditi. Lesioned multimodal language models reproduce aphasic picture-naming patterns. *arXiv preprint arXiv:2607.11621*, 2026. doi: 10.48550/arXiv.2607.11621.

A Claim map and analysis denominators

Table 3: **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.

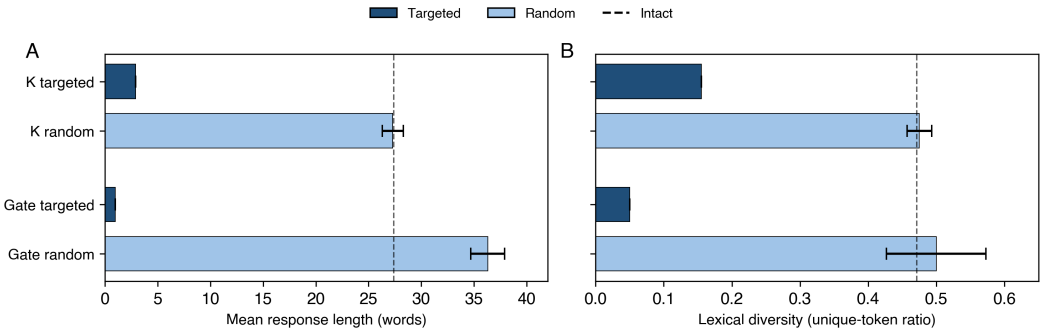


Figure 4: **Matched random does not mimic targeted lesions.** Gemma-3-1B-IT, layers 0–12, 100% severity, 3 seeds. (A) Mean response length. (B) Lexical diversity. Targeted lesions reduce both metrics; sparsity-matched random controls do not show the same targeted length/diversity reduction.

Table 4: **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 readout; individual symptom directions remain scorer-dependent.
Scorer-free text	Deterministic features	360 paired strata; z-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 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 (%)	Δ	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

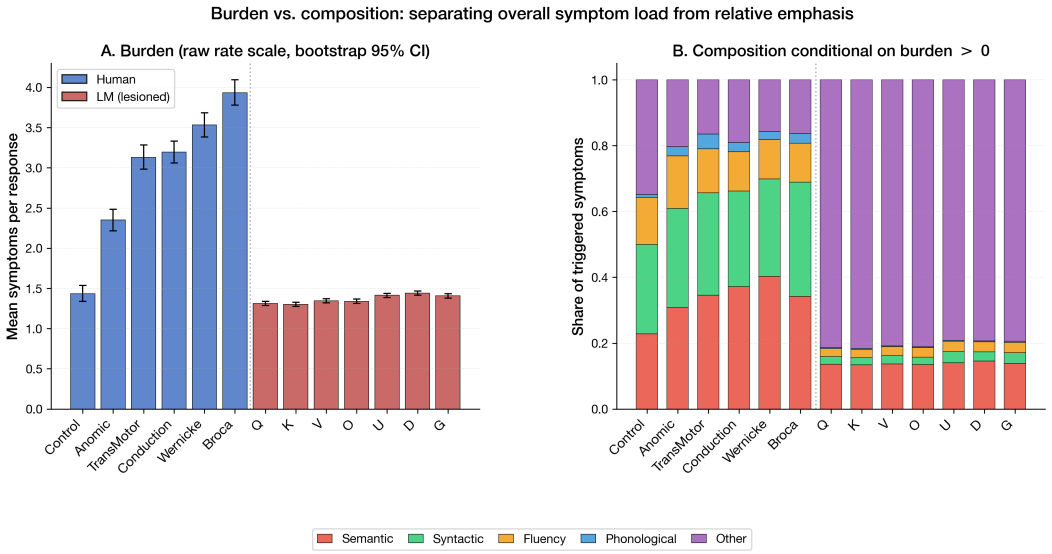


Figure 5: **Humans and lesioned LMs differ in both burden and composition.** (A) Human diagnosis groups vary in mean TAB-symptom count, whereas LM components have similar overall burdens. (B) Conditional on at least one positive TAB symptom, category shares sum over triggered symptoms and are not raw 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; includes 260 auxiliary Gemma-base records in addition to the five full sweeps.
Condition-level bootstrap	Model $\times$ layer $\times$ component $\times$ severity profile	2,528	Prompt-deduplicated first; 8 auxiliary Gemma-base profiles have no attention pair. Paired tests use 360 strata from five full sweeps.
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 co-occurrence table	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.

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 7: **Condition-level clustering preserves the selected FFN-shift contrast.** Values are FFN-minus-attention differences in percentage points after pairing within model × layer × 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)** ↑	**Bootstrap 95% CI**
Selected FFN-shift symptoms	**1.25**	[0.56, 1.99]
All remaining symptoms	0.08	[−0.03, 0.19]
Difference-in-differences	**1.17**	[0.53, 1.86]

Table 8: **All-symptom paired profile test.** FFN-minus-attention differences are estimated after prompt deduplication and pairing within model × layer × 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 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 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: **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	**4.60**	[3.37, 6.33]; .0002	**0.69**
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.

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

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

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** $\lvert\Delta \log_{10} \mathrm{KL}\rvert$	**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 ≥ 10 chars and ≥ 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	**4.12**	[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.

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	**3.93**	.091
Model family	3	3.64	.333
Prompt	20	3.20	$< .0001$

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 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× 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
Largest residual FFN-minus-attention symptoms after word-count + lexical-diversity conditioning			
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
Largest all-severity raw FFN-minus-attention feature shifts			
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

The normalized map below is exploratory: min–max scaling emphasizes within-label location and hides absolute prevalence. In the historical map, depth bins use the maximum layer index across the pooled table rather than each model's own layer count. Figure 3 instead normalizes depth within each model and shows raw rates with uncertainty. The correlations below summarize the historical normalized analysis; they do not establish an early-to-late linguistic hierarchy.

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	**0.521**	41
Semantic	0.578	15
Syntactic	0.541	6
Fluency	0.295	10
Other	0.971	10
Cross-category	0.226	169

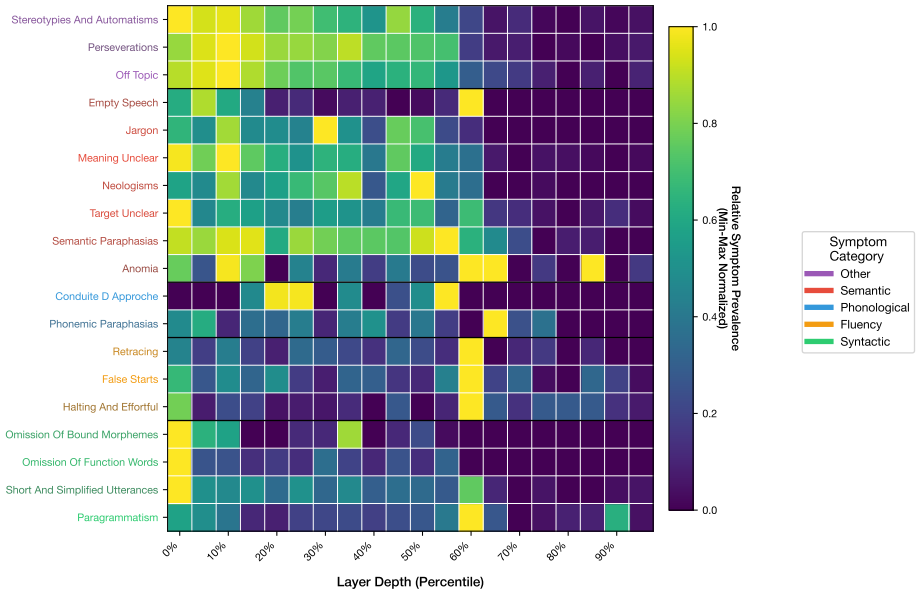


Figure 6: **Exploratory row-normalized depth map.** Each label is min–max scaled separately, so colors show relative location rather than absolute prevalence. Sparse labels can have bright peaks. This display is not evidence for a general linguistic progression across layers.

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: *house*; *breakfast*; *catastrophe*; *The sun rises in the East*; *The ambitious journalist discovered where we'd be going*.
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.

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 κ	TAB AC1	Gemma 4 E2B κ	Qwen κ
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	–	–

K Scorer audit

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.

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

Table 25: **Small 75% mask-seed surface check.** Values are mean ± 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 ± 0.040	0.702 ± 0.038	16.80 ± 2.79	0.275 ± 0.094
K	0.136 ± 0.028	0.729 ± 0.012	17.23 ± 0.74	0.375 ± 0.000

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.

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 ± std across layer–seed cells.

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

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

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 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 ± std across six layer–seed cells.

Comp.	Mean symptoms	Any symptom	Mean symptom SD	Top rerated symptoms
K	0.242 ± 0.067	0.242 ± 0.067	0.011	Short/simplified 0.225; Off-topic 0.017
Gate	0.275 ± 0.118	0.267 ± 0.118	0.026	Off-topic 0.133; Short/simplified 0.092; repetition-loop 0.050
Up	0.267 ± 0.055	0.258 ± 0.055	0.018	Short/simplified 0.158; Off-topic 0.067; repetition-loop 0.042
Down	0.192 ± 0.045	0.192 ± 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 19×19 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 7). 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.

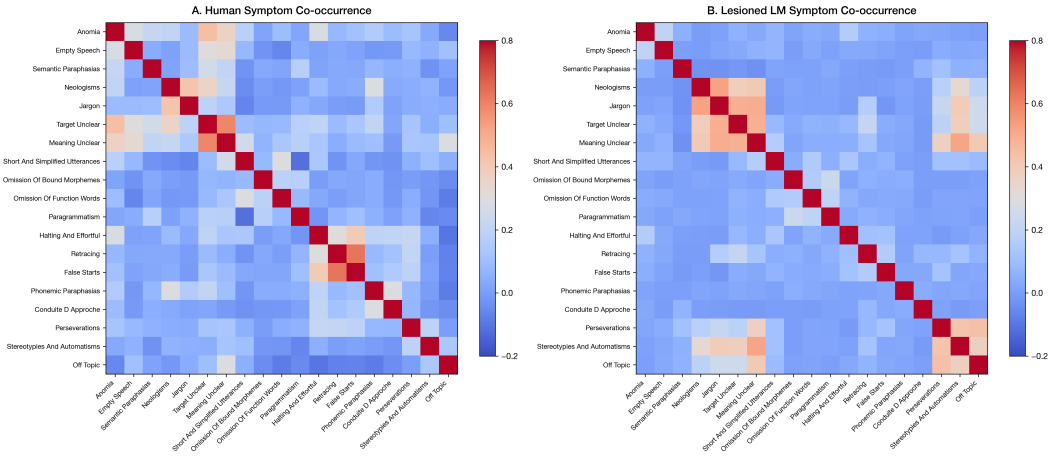


Figure 7: **Humans show denser TAB-symptom co-occurrence than LM rows in the common table.** (A) AphasiaBank ($n = 6{,}000$). (B) LM rows including intact baselines ($n = 55{,}279$ on the common 19-symptom inventory). Greedy assay: Mantel $r = 0.373$, $p = 0.0005$; repetition-filtered $r = 0.662$. Sampling 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	**.81**	.62	.69	.41	.40	.33	K [.41,.84]; .002
Broca's	.52	.60	.49	.54	.62	.59	**.67**	G [.19,.73]; .051
Wernicke's	.48	.53	.47	.50	**.59**	.54	.51	U [.16,.69]; .108
Conduction	.29	.38	.30	.32	.41	**.45**	.39	D [.11,.61]; .298
TC Motor	.23	.30	.26	.27	.29	.31	**.34**	G [−.08,.54]; .500
Control	.62	**.70**	.54	.61	.40	.39	.30	K [.42,.78]; .030

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

Q Additional figures

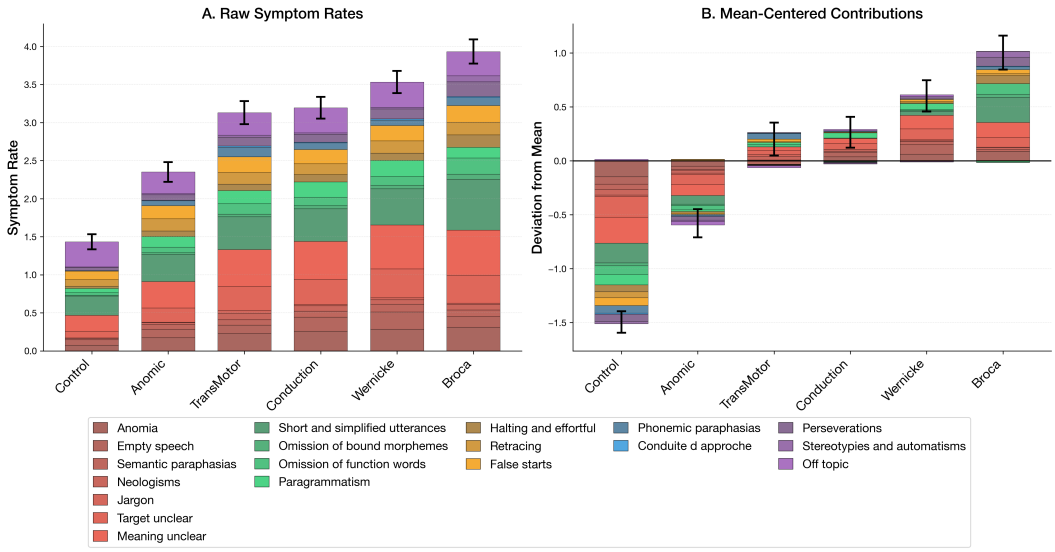


Figure 8: **Human TAB profiles by diagnosis.** (A) Raw symptom rates per diagnosis. (B) Deviations from the human corpus mean; panel A holds absolute magnitudes.

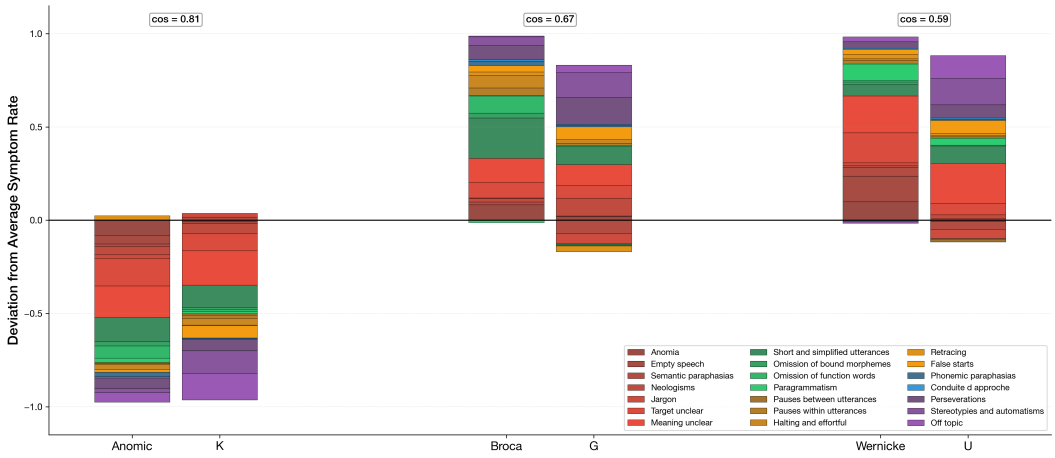


Figure 9: **Selected descriptive contrasts between human diagnosis profiles and LM component profiles.** Anomic vs. Attention-K (cosine 0.808); Broca's vs. FFN-Gate (0.665); Wernicke's vs. FFN-Up (0.592). These are exploratory profile similarities, not syndrome mappings.

R Decoding sensitivity checks

The main text reports a 5×5 decoding grid (Figure 1). 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

The single-layer Qwen2.5-7B-Instruct [Qwen et al., 2025] probe covers K/Gate/Up at layers 10 and 20, severities 75% and 100%, and 20 prompts. It was not TAB-scored and did not match functional disruption. In a separate multi-layer probe, mean repeated-subword-trigram fraction reaches .406 for Gate at 25% severity while K is .021 (Figure 10). These are deterministic surface metrics, not TAB-label rates. 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.

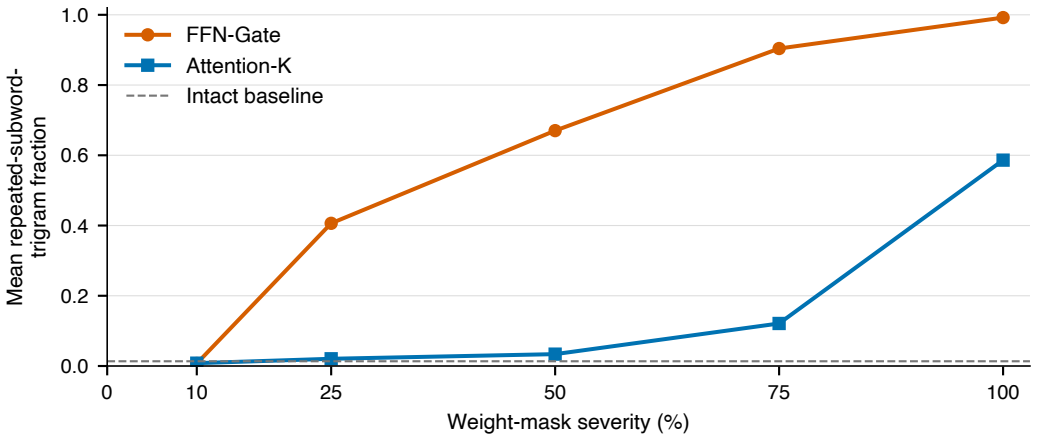


Figure 10: **Exploratory 7B multi-layer surface probe.** The measured quantity is mean repeated-subword-trigram fraction, not TAB symptom prevalence. At 25% severity, Gate is .406 and K is .021. Conditions were not matched for functional disruption.

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	**2.9**	**0.155**	**0.050**
	Random (±std)	27.3 ± 1.0	0.475 ± 0.018	0.000
gate_proj (103.5M)	Targeted	**1.0**	**0.050**	0.000
	Random (±std)	36.3 ± 1.6	0.500 ± 0.073	0.017
Intact baseline		27.4	0.471	0.000

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 ↑	Length ↓	Unique ratio ↓	Other ↑
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

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 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 ↑	Other-like ↑	Off_topic ↑	Persev. ↑
K targeted	**0.20**	**0.100**	**0.100**	**0.100**
K random (±std)	0.167 ± 0.024	0.008 ± 0.012	0.017 ± 0.024	0.000 ± 0.000
Intact baseline	0.20	0.025	0.050	0.000

U Combined human and LM TAB-symptom panel

Figure 11 combines the human and model TAB-symptom profiles from Figures 8 and 2. 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.

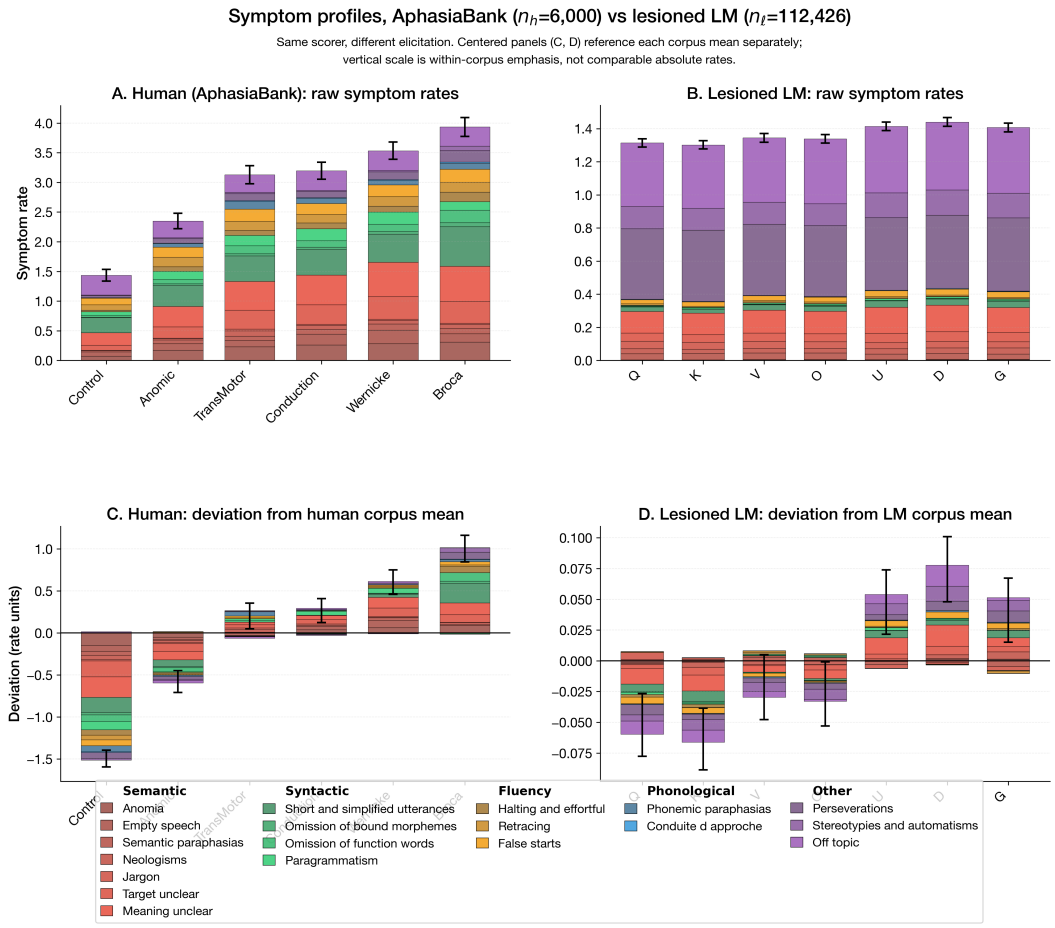


Figure 11: **Combined human and lesioned-LM TAB-symptom panel.** AphasiaBank ($n_h = 6{,}000$) and lesioned LMs ($n_\ell = 112{,}426$). (A, B) Raw symptom rates. (C, D) Deviations from the corresponding corpus mean. Centered panels show within-corpus emphasis only. Colors are shared across all four panels; 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

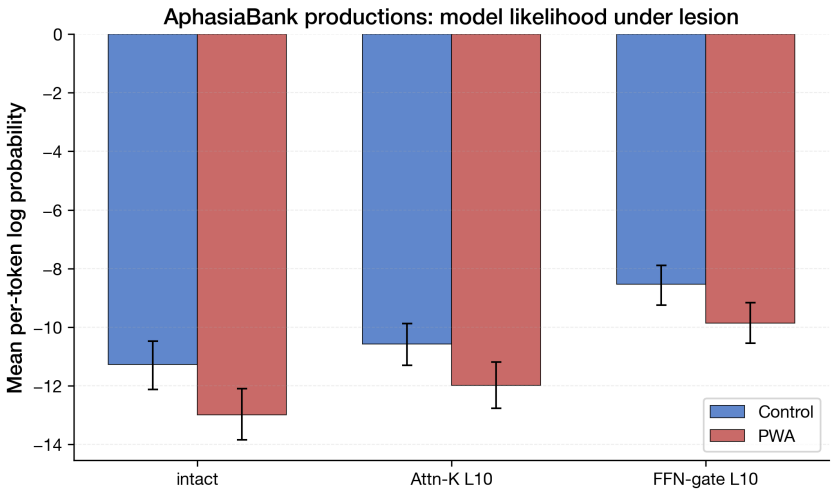


Figure 12: **Lesions shift overall likelihood but preserve the PWA–Control gap.** Mean per-token log-probability assigned by Gemma-3-1B-IT to 50 PWA and 50 Control AphasiaBank responses under three lesion conditions, with bootstrap 95% CIs.

W Activation patching

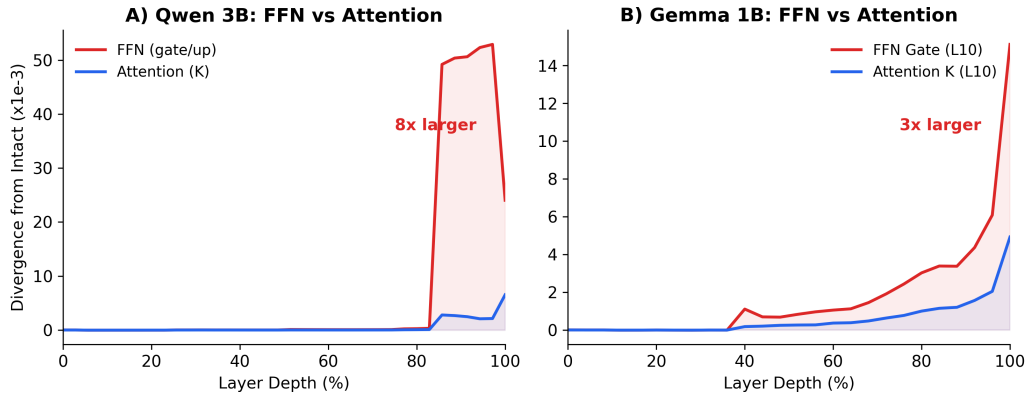


Figure 13: **FFN ablation perturbs the residual stream more than attention-K.** Hidden-state divergence from the intact model after ablation. (A) Qwen 3B bridge probe: about $10\times$ larger peak for FFN (red) than attention-K (blue), $p = 3 \times 10^{-6}$. (B) Gemma 1B bridge probe: about $2\times$ larger.

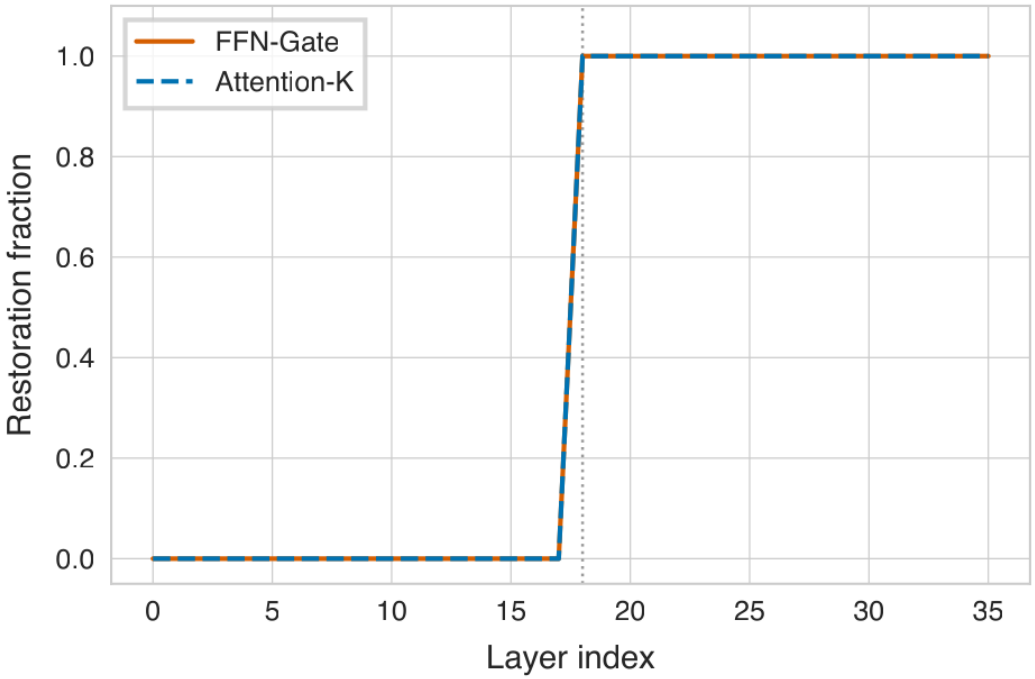


Figure 14: **Gate ablation starts from a larger baseline KL perturbation than attention-K in this small bridge probe.** Activation patching on Qwen 3B (3 prompts; mean $\pm$ std). Gate starts from a higher baseline KL divergence at the inspected layer, while restoration curves are otherwise similar; 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 10-symptom subset	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 final-layer 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.

Dimension	Evidence in this paper	Not claimed
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 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 TalkBank 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 CPU, Python/pandas	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 CPU, Python/pandas	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 unscored single-layer conditions plus 20-item minimal-pair smoke; 3B: 98-condition ablation slice, hidden-state probes, and 70-item minimal-pair check	7B unscored 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 present a behavioral-profile framework, a modest pooled attention/FFN difference under a fixed assay, and a human/LM comparison with partial, decoder-dependent overlap. Effect sizes, weak held-out classification, and scorer dependence qualify the empirical claims. Depth and 3B/7B probes are exploratory; no projection-specific syndrome, general linguistic hierarchy, or implication for human aphasia theory is claimed.

   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: The Limitations section 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 sampled mask per partial-lesion 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 model variants, the Bernoulli weight-zero ablation procedure with a sampled mask per partial-lesion 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 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` 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 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.