Interpretable Phonological–Semantic Dissociation Scoring for Automated svPPA Subtyping

Anonymous ACL submission

Abstract

We developed IPSS-PPA (Interpretable Phonological–Semantic Scoring), a fully automated pipeline for Primary Progressive Aphasia (PPA) subtype classification from picture description speech. The semantic variant (svPPA) is particularly difficult to detect: speech remains fluent and grammatically intact while semantic content degrades, leaving acoustic and fluency-based biomarkers poorly suited to detecting it. We addressed this by routing prediction through ten clinician-specified constructs (anomia, agrammatism, empty speech, and seven others) co-designed with SLPs, making every classification decision directly traceable to named clinical features rather than recovered post-hoc from a black-box model (Rudin, 2019; Koh et al., 2020). Phonological and semantic adequacy scores are computed from these constructs, and semi-automatic speaker diarization reduces the manual annotation bottleneck of prior work. On 254 WAB Picnic recordings, we achieve AUC 0.918 for control vs. svPPA and AUC 0.734 ($p < 10^{-5}$) for svPPA vs. nfvPPA+lvPPA, the first reported result on this clinically critical distinction.

1 Introduction

Primary Progressive Aphasia (PPA) is a neurodegenerative syndrome marked by progressive language deterioration in the absence of generalized cognitive decline (Mesulam, 2013). Gorno-Tempini et al. (2011) define three variants with distinct neuroanatomical profiles: the *semantic* variant (svPPA), driven by ventral temporal degeneration (Hickok and Poeppel, 2007), presents with profound lexical-semantic loss while phonological production and fluency remain largely intact; the *nonfluent* variant (nfvPPA) implicates left posterior frontal and insular cortex, producing effortful, apraxic speech with agrammatism; and the *logopenic* variant (lvPPA), arising from left temporoparietal atrophy, produces word-finding pauses and phonological working-memory deficits with otherwise preserved grammar. Reliable automated subtyping from naturalistic speech would directly support clinical workflows: reducing specialist burden in diagnosis, enabling longitudinal monitoring without repeated clinical visits, and providing objective evidence chains that integrate into existing reporting practice.

Existing automated approaches reach this goal only partially. Acoustic and prosodic features (Nevler et al., 2019; Themistocleous et al., 2018; Fraser et al., 2014; Themistocleous et al., 2021), end-to-end models, and dysfluency-based representations (Rezaii et al., 2024; Peters et al., 2025; Vonk et al., 2025; Lian et al., 2023) reliably characterize nfvPPA and lvPPA, where motor speech errors and dysfluency are salient. Characterizing svPPA, and distinguishing it from the other variants, remains the central unsolved challenge. svPPA manifests primarily as a semantic deficit: speech stays fluent and grammatically intact while content grows markedly impoverished. Patients produce utterances like *"she's holding that kite thing… with the string bits"* in place of precise descriptions. Under the dual-stream model (Hickok and Poeppel, 2007), svPPA selectively impairs the ventral (lexical-semantic) stream while sparing the dorsal (articulatory-phonological) stream, a dissociation that fluency-based and dysfluency-based metrics cannot capture. Acoustic-prosodic approaches reach only 66% svPPA accuracy (Themistocleous et al., 2018) versus 82% for nfvPPA, and the prior state of the art (Peters et al., 2025) does not report svPPA vs. nfvPPA+lvPPA discrimination at all.

To address this gap we developed IPSS-PPA, an end-to-end automated pipeline for PPA subtype classification. The pipeline grounds semantic scoring in the WAB Picnic scene (Nicholas and Brookshire, 1993), whose well-defined entities, spatial relations, and actions admit formal representation

as a scene graph and enable stimulus-grounded semantic evaluation impossible with free speech. On 254 WAB Picnic recordings, IPSS-PPA achieves AUC 0.92 for control vs. svPPA and AUC 0.734 ($p < 10^{-5}$, Cliff's $\delta = 0.469$, medium effect) for svPPA vs. nfvPPA+lvPPA. Three-way subtype accuracy reaches 60.3%, below the prior reported 74.0% on $n = 59$ (Peters et al., 2025) but obtained without manual diarization, on a 4× larger cohort, and with roughly 30 derived features versus 363. A transcript-only LLM baseline scores AUC 0.500 on svPPA vs. nfv+lv. To address this gap, we developed IPSS-PPA, an end-to-end automated pipeline for PPA subtype classification. The pipeline grounds semantic scoring in the WAB Picnic scene (Nicholas and Brookshire, 1993), whose well-defined entities, spatial relations, and actions admit formal representation as a scene graph and enable stimulus-grounded semantic evaluation impossible with free speech.

Our contributions are summarized as follows:

(1) We achieve high-accuracy, automated classification on a clinically critical distinction using a cohort four times larger than prior work. On 254 WAB Picnic recordings, IPSS-PPA achieves AUC 0.92 for control vs. svPPA and AUC 0.734 ($p < 10^{-5}$, Cliff's $\delta = 0.469$, medium effect) for svPPA vs. nfvPPA+lvPPA. Three-way subtype accuracy reaches 60.3% obtained without manual diarization, vastly outperforming a transcript-only LLM baseline (AUC 0.500). Furthermore, group-level construct scores emerge naturally without any supervised optimization on the classification labels.

(2) We introduce a fully interpretable concept bottleneck architecture co-designed and validated with speech-language pathologists (SLPs). Following the framework of Koh et al. (2020), we route predictions through ten clinician-specified constructs (e.g., anomia, agrammatism, empty speech) developed iteratively and validated against QAB diagnostic categories (Wilson et al., 2018). The resulting feature space mirrors the conceptual vocabulary clinicians already use, ensuring every classification is directly traceable to named clinical features without recourse to post-hoc attribution methods (Rudin, 2019).

(3) We deliver a highly scalable, stimulus-agnostic pipeline that dramatically reduces manual annotation barriers and generalizes seamlessly. By integrating pyannote.audio (Bredin et al., 2021) diarization to flag mixed-speaker segments, 174 out of 254 recordings (68.5%) require zero manual intervention, cutting hands-on review time by an estimated 68% to 83%. Additionally, we parameterize the semantic scorer via an exchangeable JSON scene graph, allowing the entire framework to expand to alternative stimuli (such as the Cookie Theft description) with no model retraining or feature re-engineering.

2 Related Work

Early automated PPA characterization relied on task-specific hand-crafted features. Acoustic and prosodic measures reliably capture the articulatory breakdowns characteristic of nfvPPA, with 82% accuracy (Themistocleous et al., 2018), but leave svPPA largely undetected at 66%, since svPPA speech is fluent and prosodically intact (Nevler et al., 2019). Fraser et al. (2014) distinguished svPPA from nfvPPA at 79% accuracy using syntactic and semantic NLP features; Zimmerer et al. (2016) achieved 90% control vs. PPA separation but only 59.4% subgroup accuracy using word-frequency measures. Themistocleous et al. (2021) reached 80% three-way accuracy with a deep neural network over acoustic and morphosyntactic features, correctly identifying 90% of nfvPPA and 95% of lvPPA cases, yet svPPA remained the hardest variant to isolate across every system.

Rezaii et al. (2024) applied unsupervised LLM clustering to 78 PPA patients, achieving 88.5% agreement with clinical diagnoses. Phan et al. (2024) established a transcript-only LLM upper bound for picture-description classification. On the multimodal side, Favaro et al. (2023) automated content-unit scoring for Alzheimer's detection and Ambadi et al. (2021) proposed spatio-semantic graph representations for neurodegenerative speech, though neither has been validated for PPA subtype discrimination. The current state of the art, Peters et al. (2025), integrates acoustic features, ASR-derived linguistics, content-unit scoring, and spatio-semantic graphs to reach 97% binary and 74% three-way accuracy on Cookie Theft recordings. The system requires manual diarization at inference time and does not report svPPA vs. nfvPPA+lvPPA discrimination. Across these methods, predictions are not decomposable into clinically named features, limiting utility for diagnosis support and prospective validation.

Other work frames dysfluency as an interpretable intermediate representation (Lian et al., 2023; Lian and Anumanchipalli, 2024; Lian et al., 2024, 2025),

Table 1: Dataset composition and duration for WAB Picnic recordings.

Group	N	Avg (s)	Var (s^2)	Total (h)
Healthy controls	25	83.89	1432.48	0.58
svPPA	42	90.00	2347.65	1.05
nfvPPA	78	104.29	2638.63	2.26
lvPPA	109	113.60	3633.22	3.44
Total	**254**	103.92	3010.43	7.33

with automated pipelines for detection (Zhou et al., 2024; Zhang et al., 2025), transcription (Guo et al., 2025, 2026), and alignment (Ye et al., 2025) showing strong agreement with clinician judgments on nfvPPA and lvPPA (Vonk et al., 2025, 2026). These pipelines handle motor speech errors well but have limited sensitivity to semantic content, the primary deficit in svPPA. IPSS-PPA addresses this gap by grounding semantic scoring in a stimulus-derived scene graph and routing features through clinical constructs that include anomia and empty speech alongside fluency-based measures.

3 Data

All recordings were drawn from a longitudinal clinical PPA cohort. Participants were diagnosed according to the criteria of Gorno-Tempini et al. (2011) by board-certified neurologists. The dataset comprises 254 WAB Picnic picture-description recordings across four diagnostic groups (Table 1), with one recording per session.

The WAB Picnic scene includes 21 entities, 23 relations, 6 attributes, and 12 content units (Nicholas and Brookshire, 1993). Figure 1 shows entity-level naming rates for healthy controls and svPPA participants (green: $\geq$50% named, orange: 25 to 50%, red: $<$25%). svPPA patients show systematically lower naming rates, consistent with impaired lexical-semantic access in the presence of preserved phonological fluency, and motivating the entity and relation-level scoring in $S_{\text{sem}}^{\text{graph}}$ described in Section 4.

3.1 Preprocessing

To protect participant privacy, all audio recordings and text transcripts were fully de-identified and stripped of any personally identifying information (PII) or protected health information (PHI) prior to analysis.

Audio is naturalistic clinical speech with heterogeneous recording conditions. We apply light denoising and level normalization before diarization to stabilize downstream ASR and phoneme extraction. We use pyannote.audio (Bredin et al., 2021) for first-pass speaker segmentation; flagged mixed-speaker segments are corrected by trained annotators, while high-confidence segments are accepted automatically. We developed the gold scene graph for semantic scoring in collaboration with clinician partners and SLPs, encoded it in JSON, and released it with the supplementary materials. To our knowledge this is the first automated scoring system validated on the WAB Picnic stimulus.

The pyannote.audio pretrained models target meeting and broadcast speech; principal failure modes on clinical recordings are backchannel misattribution, overlap under-detection at turn boundaries, and speaker confusion on dysarthric voices (characteristic of advanced nfvPPA). These errors concentrate at segment boundaries and short ($<$1 s) turns, making them identifiable without a full manual pass.

3.2 Semi-Automation Efficiency

Of 254 recordings, 174 (68.5%) passed through automatically; the remaining 80 required annotator correction of flagged regions. At the recording level, this is a 68.5% reduction in annotation effort. The pyannote.audio errors concentrate in clinician-turn and overlap regions, which together make up 16.9% of total speech, so the hands-on review fraction by duration is closer to 17%, implying a $\sim$83% reduction in diarization time relative to fully manual labelling. We report both numbers (68% and 83%) because they measure different things; the 83% figure is the duration-based estimate consistent with the diarization bottleneck reported by Peters et al. (2025).

4 Methods

IPSS-PPA converts raw picture-description audio into clinically grounded features in two stages, followed by classification. We isolated participant speech via pyannote.audio (v3.1) (Bredin et al., 2021), then ran two parallel branches on the retained segments: Whisper large-v2 (Radford et al., 2023) produced a word-level transcript with timestamps, and HuPER (Guo et al., 2026) produced a lexicon-free phoneme sequence directly from the audio. We derived phonemes acoustically rather than from the ASR hypothesis so that S_{ph} reflects articulatory production quality independently

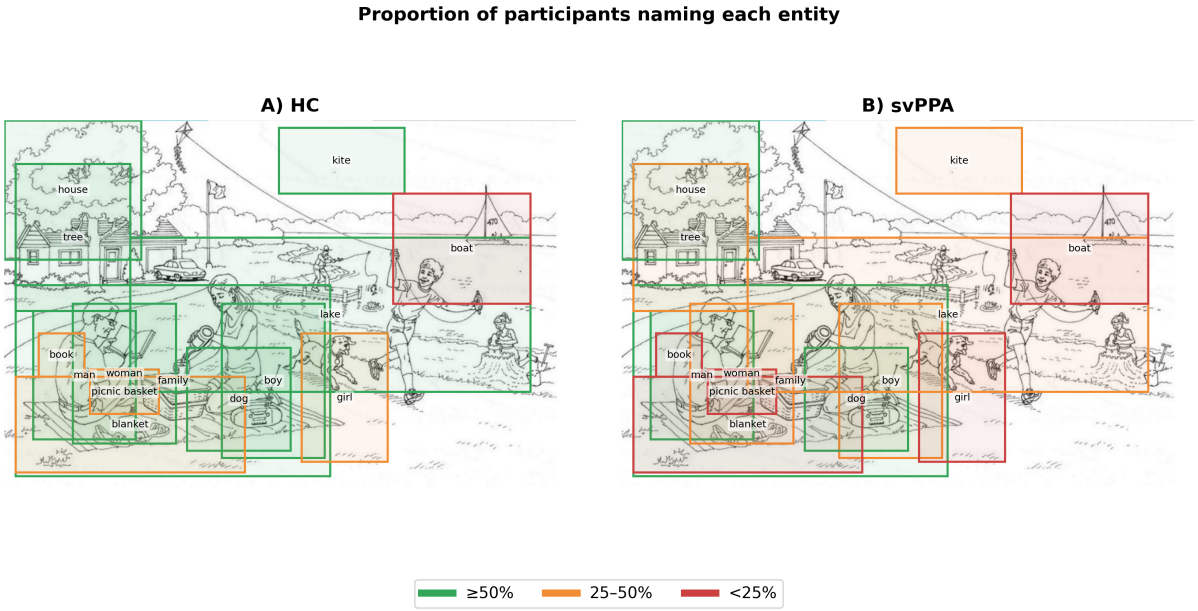

Figure 1: Proportion of participants naming each WAB Picnic entity, for healthy controls (HC, panel A) and svPPA (panel B). Bounding box color indicates naming rate: green ≥50%, orange 25 to 50%, red <25%.

of lexical substitution errors, preserving the separation between the phonological and semantic branches.

4.1 Layer 1: Proxy Features

We developed 20 task-agnostic proxy measures from the transcript and pyannote.audio timing, covering fluency (words per minute, articulation rate, pause duration, pause-to-speech ratio, response latency), utterance structure (mean and median utterance length, clauses per utterance, abandoned utterance rate, utterance count), lexical richness (open/closed-class ratio, content word density, type-token ratio, pronoun usage ratio), morphosyntax (function-word omission rate, morphological error rate, dependency parse completeness), and dysfluency (word-finding pause duration, filled-pause frequency, semantic distance to expected content). We selected these proxies iteratively with SLPs to capture hallmark cues of each PPA variant using only automated tools, with no human reference annotations required. spaCy (Honnibal et al., 2020) provides POS tagging and dependency parsing. Each proxy is z-scored against cohort statistics computed on training folds within each CV split.

4.2 Layer 2: Clinical Constructs

We aggregated Layer 1 z-scores into ten constructs aligned with the Quick Aphasia Battery (QAB) (Wilson et al., 2018): *reduced speech rate, reduced length/complexity, agrammatism, paragrammatism, anomia, empty speech, semantic paraphasias, phonemic paraphasias/neologisms, self-correction,* and *overall communication impairment*. Each construct is a weighted combination of Layer 1 z-scores passed through a bounded nonlinearity:

$$\mathrm{L2}_c = \frac{\tanh(\bar{z}_c) + 1}{2} \tag{1}$$

where $\bar{z}_c$ is the mean of the z-scored proxies assigned to construct c, and the full expression $(\tanh(\bar{z}_c) + 1)/2$ maps into $[0, 1]$. We specified construct definitions and proxy weightings with SLPs to reflect WAB and QAB diagnostic categories. The mappings are pre-specified and theory-driven, a deliberate design choice following Koh et al. (2020): the constructs are the model's intermediate representation, not a post-hoc explanation of it. A summary adequacy score aggregates across constructs: $S_{\text{sem}}^{\text{L2}} = 1 - \text{mean}(\mathrm{L2}_1, \ldots, \mathrm{L2}_{10})$.

4.3 Phonological Adequacy (S_{ph})

We generated expected phonemes from the Whisper transcript via a CMU-based G2P lookup with character-level fallback (Carnegie Mellon University Speech Group, 2014); observed phonemes are read from the HuPER stream after silence tokens are removed. Phonological adequacy is:

$$S_{\text{ph}} = 1 - \frac{\text{Levenshtein}(\hat{P},\ P^*)}{\max(|\hat{P}|,\ |P^*|)} \tag{2}$$

where $\hat{P}$ and P^* are observed and canonical phoneme sequences. $S_{\text{ph}} \in [0, 1]$, with higher values indicating more intact phonological production.

4.4 Semantic Adequacy (S_{sem})

We combine a stimulus-grounded scene-graph score with the Layer 2 adequacy score:

$$S_{\text{sem}} = \alpha\, S_{\text{sem}}^{\text{graph}} + (1 - \alpha)\, S_{\text{sem}}^{\text{L2}}, \quad \alpha = 0.5 \quad (3)$$

The graph score matches the transcript against a normative WAB Picnic scene graph of 21 entities, 23 directed relations, and 12 content units (Nicholas and Brookshire, 1993). Entity mentions are detected via canonical maps with synonym sets; relations use directional rules respecting word order, so *boy holding string* and *string holding boy* score differently, penalizing the semantically anomalous constructions characteristic of svPPA. Graph scoring weights content unit coverage over entity recall, following Nicholas and Brookshire (1993):

$$S_{\text{sem}}^{\text{graph}} = 0.4\, \frac{|\text{matched entities}|}{21} + 0.6\, \frac{|\text{matched CUs}|}{12} \quad (4)$$

Where Peters et al. (2025) encode the temporal order of content-unit mentions weighted by image coordinates, our graph encodes semantic relations between entities, contrasting scene navigation with semantic content coverage.

5 Experiments

5.1 Evaluation Protocol

We report mean ± standard deviation over 10 runs (seeds 0 to 9) with stratified 5-fold cross-validation. Binary tasks report accuracy, F1, and ROC-AUC; imbalanced settings additionally report balanced accuracy, macro-F1, PR-AUC, and confusion matrices.

5.2 Experimental Design

We evaluate across four progressively harder settings.

Experiment 1: Control vs. svPPA ($n = 67$). Binary classification with ablations: (1) S_{sem} only,(2) $S_{\text{ph}}+S_{\text{sem}}$. Tests whether adequacy scores capture the svPPA profile (degraded semantic retrieval with preserved phonological production) against healthy controls.

Table 2: Control vs. svPPA (top models; 10 runs).

Feature	Acc.	F1	AUC
$S_{ph} + S_{sem}$	0.82 ± 0.01	0.85 ± 0.01	$\mathbf{0.918 \pm 0.006}$
$S_{ph} + S_{sem} + \Delta$	0.82 ± 0.01	0.85 ± 0.01	0.917 ± 0.007
Δ + extras	0.82 ± 0.01	0.85 ± 0.01	0.910 ± 0.010

Experiment 2: svPPA vs. nfvPPA+lvPPA ($n = 229$). Binary classification across four feature groups: phonological and semantic adequacy scores, acoustic features, scene-graph scores, and Layer-2 constructs. We evaluate logistic regression (LR) and random forest (RF). This is the clinically critical distinction no prior automated system has reported.

Experiment 3: Balanced svPPA vs. nfvPPA+lvPPA. The full cohort is skewed (svPPA: 42 vs. 187 nfvPPA+lvPPA). We verify results are not an artifact of class imbalance via repeated subsampling (25 vs. 25 per run), reporting balanced accuracy, macro-F1, PR-AUC, and confusion matrices.

Experiment 4: Three-way svPPA vs. nfvPPA vs. lvPPA ($n = 229$). Full subtype classification with macro-F1, per-class F1, and confusion-matrix analysis.

6 Results

6.1 Quantitative Performance

Effect sizes use Cliff's δ with thresholds from Romano et al. (2006): $|\delta| < 0.147$ negligible, 0.147 to 0.330 small, 0.330 to 0.474 medium, ≥ 0.474 large. Group differences use two-sided Mann–Whitney U.

Experiment 1: Control vs. svPPA ($n = 67$). Table 2 reports results. $S_{\text{ph}} + S_{\text{sem}}$ achieves AUC 0.918 ± 0.006. Adding Δ to this pair yields no meaningful gain (0.917 ± 0.007).

To understand why Δ adds nothing on top of the adequacy pair, we trained classifiers on Δ alone (Table 3). For control vs. svPPA, Δ-only achieves AUC 0.562 ± 0.016, near chance, while $S_{\text{ph}} + S_{\text{sem}}$ reaches 0.918 ± 0.006. Among PPA subtypes (svPPA vs. nfv+lv, $n = 229$), Δ-only AUC is 0.498 ± 0.041, providing no subtype-discriminative information on its own. We therefore interpret Δ as an expected directional summary index rather than a standalone biomarker: the raw difference between phonological and semantic adequacy is too noisy to separate subtypes without the Layer 1

Table 3: Δ-only ablation vs. full adequacy pair (LR; 10 runs).

Task	Features	n	AUC
Ctrl vs. svPPA	Δ only	67	0.562 ± 0.016
	$S_{\text{ph}} + S_{\text{sem}}$	67	$\mathbf{0.918 \pm 0.006}$
	$S_{\text{ph}} + S_{\text{sem}} + \Delta$	67	0.917 ± 0.007
svPPA vs. nfv+lv	Δ only	229	0.498 ± 0.041

proxy extraction and Layer 2 construct aggregation that ground both terms.

Experiments 2 to 4: subtype discrimination. Table 4 consolidates results across the three PPA-only settings. On the imbalanced cohort ($n = 229$), Full+L2+G (RF) achieves Acc 0.82 and AUC 0.75 (F1 0.33 minority-class; see Table 7 for full F1). The balanced 25/25 subsample confirms this holds under equal class sizes (BalAcc 0.67, PR-AUC 0.72). Three-way classification yields macro-F1 0.560 with per-class F1 of 0.489 (sv), 0.641 (nfv), and 0.549 (lv).

To attribute the full-model gain to specific components, we trained classifiers on isolated feature subsets ($n = 229$, 10 runs; Table 5). $S_{\text{sem}}^{\text{graph}}$ alone (LR AUC 0.546 ± 0.009) is the strongest single semantic component, exceeding S_{ph} alone (0.447 ± 0.056) and $S_{\text{sem}}^{\text{L2}}$ alone (0.421 ± 0.052). The composite pair $S_{\text{ph}}+S_{\text{sem}}$ remains near chance under LR (0.491 ± 0.021), while RF on the same pair reaches 0.626 ± 0.026 by exploiting nonlinear interactions; global adequacy scores under LR are insufficient for within-PPA subtype separation. The full 15-feature subset ($S_{\text{ph}}, S_{\text{sem}}, S_{\text{sem}}^{\text{graph}}, \Delta$, two acoustic proxies, ten Layer 2 constructs) reaches AUC 0.748 ± 0.009 (LR) and 0.785 ± 0.010 (RF), confirming that Layer 2 constructs are necessary for subtype discrimination beyond global adequacy scores.

6.2 Longitudinal Analysis

Beyond cross-sectional classification, we examined whether the adequacy scores track disease progression in a subtype-specific way. We compared mean percent change from first visit for svPPA ($n = 10$) versus pooled lvPPA+nfvPPA ($n = 37$) over 0, 1, and 2 years, tracking S_{ph}, S_{sem}, and $S_{\text{overall}} = (S_{\text{ph}} + S_{\text{sem}})/2$. Figure 2 shows the expected dissociation: svPPA declines faster on S_{sem} than on S_{ph}, while lvPPA+nfvPPA shows the reverse pattern. The trajectory mirrors the cross-sectional adequacy profile and supports modeling both channels explicitly.

6.3 Baseline Comparison

To test whether the discriminative signal is recoverable from text alone, we evaluated qwen3.5-plus on Whisper transcripts with few-shot prompting (two examples per class, see Appendix C.3 for prompts). For tractability we drew a stratified subsample of 25 recordings per class (50 per task) from the main cohort. AUC 0.50 on svPPA vs. nfv+lvPPA on this subsample (Table 6) confirms that transcript-only models cannot recover the discriminative signal that IPSS-PPA captures from multimodal features.

Table 7 compares IPSS-PPA with LFTK on identical splits. Direct accuracy comparisons are limited by dataset and stimulus differences (Cookie Theft, $n = 59$ vs. WAB Picnic, $n = 254$). IPSS-PPA leads on the clinically critical svPPA vs. nfv+lv task (AUC 0.75 vs. 0.66) and on balanced evaluation (macro-F1 0.67 vs. 0.56), using roughly 30 derived features versus 363 and no manual annotation.

6.4 Construct Profile Analysis

To probe what the model has learned, we examined the Layer 2 construct scores it produces. Figure 3 shows mean adequacy and Layer 2 construct scores by diagnostic group. svPPA exhibits the sharpest drop in scene-graph coverage ($S_{\text{sem}}^{\text{graph}}$: HC 0.27 vs. svPPA 0.12) and the highest anomia (0.58); nfvPPA shows the highest reduced speech rate (0.56) and disproportionate agrammatism. These patterns mirror the Gorno-Tempini et al. (2011) criteria and emerge from clinician-specified construct weights without supervised optimization on classification labels, consistent with concept bottleneck models (Koh et al., 2020).

As a minimal functionally grounded interpretability check (Doshi-Velez and Kim, 2017), we computed Spearman correlations between Layer 2 construct scores and speech-derived clinical proxies ($n = 254$; Table 8). Anomia correlated negatively with scene-graph coverage ($\rho = -0.328$, $p < 10^{-7}$) and composite semantic adequacy ($\rho = -0.507$, $p < 10^{-15}$), consistent with the intended mapping: patients with higher anomia scores produce fewer identifiable picnic entities and content units. These correlations are not a substitute for application-grounded evaluation, which would require independent SLP ratings, but confirm that the anomia construct behaves as clinically

Table 4: Results across evaluation conditions (top models; 10 runs). *Real*: imbalanced cohort ($n = 229$); minority-class F1 reported separately in Table 7. *Balanced*: 25 svPPA vs. 25 nfv+lv per run. *3-way*: sv/nfv/lv ($n = 229$).

Condition	Model	Acc.	AUC	Macro-F1
Real ($n = 229$)	Full+L2 (RF)	0.81 ± 0.01	0.74 ± 0.01	—
	Acoustic+graph (RF)	0.78 ± 0.02	0.65 ± 0.02	—
	Full+L2+G (RF)	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$	—
Balanced (25/25)	Full+L2 (RF)	$\mathbf{0.67 \pm 0.06}$	0.72 ± 0.10	$\mathbf{0.67 \pm 0.06}$
	Full+L2+G (LR)	0.66 ± 0.03	0.70 ± 0.09	0.66 ± 0.03
	Full+L2+G (RF)	0.66 ± 0.05	0.71 ± 0.08	0.66 ± 0.05
3-way (sv/nfv/lv)	RF (best Acc.)	$\mathbf{0.60 \pm 0.02}$	—	0.55 ± 0.02
	LR (best Macro-F1)	0.57 ± 0.02	—	$\mathbf{0.56 \pm 0.02}$

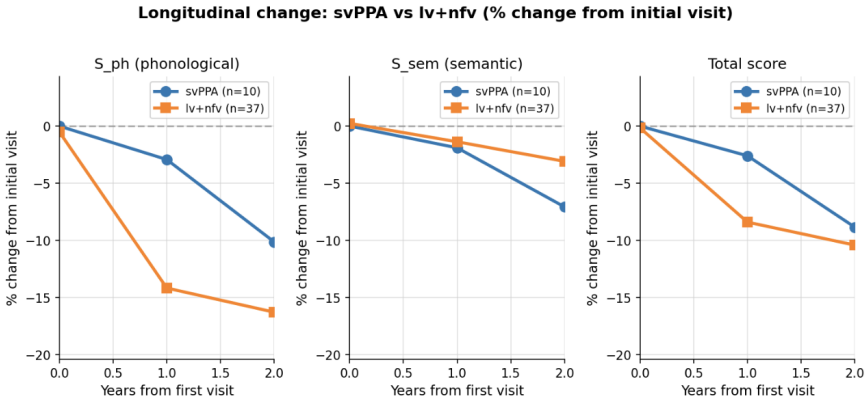

Figure 2: Longitudinal change by subtype (percent change from first visit, years 0 to 2). Blue: svPPA ($n = 10$); orange: lvPPA+nfvPPA ($n = 37$). svPPA declines faster on S_{sem}; lvPPA+nfvPPA declines faster on S_{ph}.

Table 5: Feature ablation, svPPA vs. nfv+lv ($n = 229$; 10 runs, 5-fold CV). The 15-feature ablation set is a focused subset of the ∼30 features used in the main pipeline.

Feature set	# feat.	AUC (LR)	AUC (RF)
S_{ph} only	1	0.447 ± 0.056	0.552 ± 0.023
$S_{\text{sem}}^{\text{graph}}$ only	1	0.546 ± 0.009	0.602 ± 0.025
$S_{\text{sem}}^{\text{L2}}$ only	1	0.421 ± 0.052	0.530 ± 0.029
$S_{\text{ph}} + S_{\text{sem}}$	2	0.491 ± 0.021	0.626 ± 0.026
Full + L2 + graph	15	$\mathbf{0.748 \pm 0.009}$	$\mathbf{0.785 \pm 0.010}$

Table 6: LLM baseline results (qwen3.5-plus) on stratified subsamples of 25 recordings per class drawn from the main cohort.

Task	Acc	F1	AUC
Control vs. svPPA ($n = 50$)	0.56	0.69	0.56
svPPA vs. nfv+lvPPA ($n = 50$)	0.50	0.44	0.50

intended without supervision on classification labels.

To probe how the construct profile behaves on individual recordings, we examined one correctly classified and one misclassified control (Table 9). Case A's negative $\Delta = -0.055$ and low anomia (0.219) anchor the correct prediction. Case B's fragmented output, likely ASR degradation on non-English tokens, depressed scene coverage (CU coverage 0.083) and inflated anomia (0.649) and phonemic paraphasias (0.689), producing a profile superficially resembling svPPA. The failure illustrates a known mode of stimulus-grounded scoring: when ASR quality degrades, spurious tokens reduce entity-match recall without any underlying semantic impairment.

7 Discussion and Conclusion

IPSS-PPA achieves competitive PPA subtype classification without manual preprocessing, using roughly 30 clinician-grounded features versus 363 in the prior state of the art. Three-way accuracy reaches 60.3%, below the 74.0% reported by Peters et al. (2025) on a manually annotated cohort of $n = 59$, but obtained on a 4× larger cohort and without manual diarization. The binary accuracy gap (0.85 vs. 0.97) is consistent with the same automation and stimulus differences. The svPPA vs. nfvPPA+lvPPA result (AUC 0.734, medium effect under Romano et al. (2006)) has no prior automated

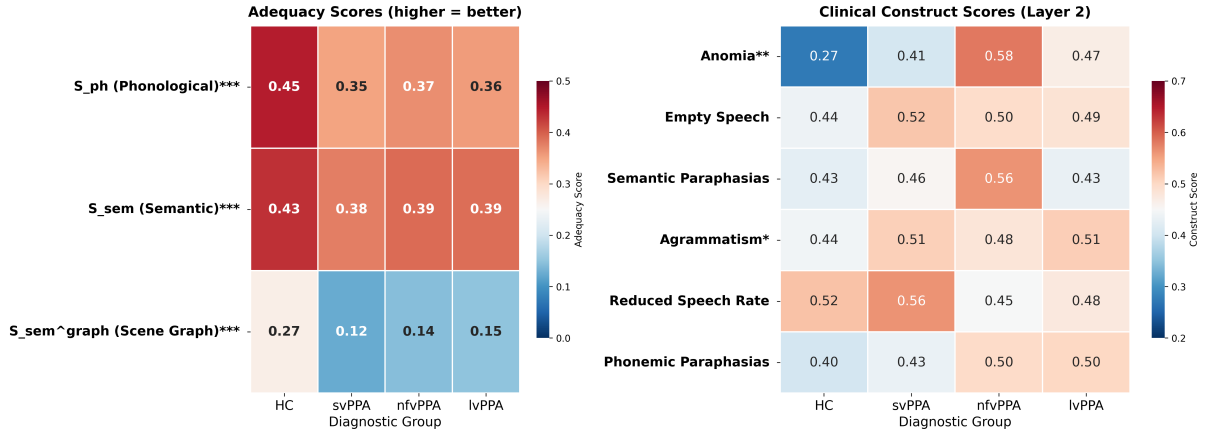

Figure 3: Adequacy and Layer 2 construct scores by group (HC, svPPA, nfvPPA, lvPPA). *Top:* S_{ph}, S_{sem}, $S_{\text{sem}}^{\text{graph}}$; *** $p < 0.001$ vs. HC. *Bottom:* Layer 2 construct scores; ** $p < 0.01$, * $p < 0.05$.

Table 7: IPSS-PPA vs. LFTK (mean over 10 runs). Best per task bolded.

Task	Model	Acc	F1	AUC
HC vs svPPA	IPSS-PPA	0.82	–	0.92
	LFTK	**0.89**	–	**0.96**
svPPA vs nfv+lv	IPSS-PPA	**0.82**	**0.33**	**0.75**
	LFTK	0.81	0.14	0.66
Balanced 25+25	IPSS-PPA	**0.67**	**0.67**	**0.72**
	LFTK	0.57	0.56	0.62
Three-way	IPSS-PPA	0.60	**0.56**	–
	LFTK	**0.66**	0.52	–

Table 8: Construct validity: Spearman ρ between Layer 2 anomia and speech-derived semantic proxies ($n = 254$).

Construct	Proxy	ρ	p
Anomia	$S_{\text{sem}}^{\text{graph}}$	-0.328	$< 10^{-7}$
Anomia	S_{sem}	-0.507	$< 10^{-15}$
Anomia	$S_{\text{sem}}^{\text{L2}}$	-0.450	$< 10^{-12}$

Table 9: Score profiles, Cases A and B (control vs. svPPA). Higher adequacy = more intact; higher L2 = greater impairment.

Score	Case A	Case B
Adequacy & graph scores		
S_{ph}	0.422	0.413
S_{sem}	0.477	0.406
Δ	-0.055	$+0.007$
Entity recall	0.571	0.381
CU coverage	0.250	0.083
Layer 2 constructs		
Reduced speech rate	0.681	0.611
Agrammatism	0.244	0.501
Anomia	0.219	0.649
Empty speech	0.373	0.289
Semantic paraphasias	0.289	0.460
Phonemic paraph./neol.	0.353	0.689
True label	control	control
Predicted	**control**	**svPPA**

benchmark.

The ablations clarify where this performance comes from. The raw dissociation index Δ alone scores near chance on both tasks (AUC 0.562 on control vs. svPPA, 0.498 on svPPA vs. nfv+lv), so Δ is best understood as an expected directional summary index rather than a standalone biomarker. The stimulus-grounded scene graph carries the strongest single semantic signal (LR AUC 0.546 alone), and Layer 2 constructs are necessary to push the full model from near-chance LR pair performance to AUC 0.748 (LR) and 0.785 (RF) on subtype discrimination.

Because predictions route through clinician-specified constructs rather than a black-box feature space (Rudin, 2019; Koh et al., 2020), the system is interrogable using the same conceptual categories clinicians use in routine assessment. The exchangeable scene graph extends this to new picture stimuli without retraining. We treat interpretability here as an architectural property rather than a formally evaluated claim; application-grounded validation with SLP raters remains future work (Doshi-Velez and Kim, 2017).

AUC 0.50 for the LLM baseline on svPPA vs. nfv+lv confirms the discriminative signal is not in the transcript alone. A compact, theory-grounded feature set can match or exceed high-dimensional systems on PPA subtyping, with implications for clinical speech biomarker design more broadly: deployability and transparency matter.

Limitations

Interpretability evaluation. We treat interpretability as an architectural property: construct scores are inspectable by design rather than recovered post-hoc. Formal evaluation in the sense of Doshi-Velez and Kim (2017), including application-grounded studies with SLP raters and perturbation testing, remains future work.

Class imbalance. The svPPA vs. nfvPPA+lvPPA task is imbalanced (42 vs. 187 recordings), yielding modest minority-class F1 despite competitive AUC. Future work should explore oversampling and cost-sensitive training.

Construct weights. Layer 1 to Layer 2 mappings are pre-specified and theory-driven; supervised optimization of these weights is a near-term priority.

Single cohort. All results are on one longitudinal cohort from one clinical site; replication on an independent cohort is required.

Diarization quality. pyannote.audio errors concentrate in overlap and clinician-turn regions; validation against manual segmentation on a held-out subset remains outstanding.

Stimulus specificity. Evaluation is limited to the WAB Picnic scene; transfer to Cookie Theft or other BDAE stimuli requires empirical validation.

Potential Risks and Ethical Considerations

While IPSS-PPA provides highly interpretable construct scores, it is designed strictly as a diagnostic support tool for qualified speech-language pathologists and neurologists, rather than a standalone diagnostic system. Misclassifications or transcription errors stemming from severe acoustic dysfluencies carry the risk of clinical misdirection if deployed without expert human-in-the-loop oversight.

To protect participant privacy, all audio recordings and text transcripts were fully de-identified and stripped of any personally identifying information (PII) or protected health information (PHI) prior to analysis.

References

Dhinakaran Ambadi, John R. Hodges, Muireann Irish, Samrah Ahmed, and Olivier Piguet. 2021. A spatio-semantic framework for characterizing picture description in neurodegenerative disorders. *Frontiers in Human Neuroscience*, 15:626780.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2021. Pyannote.audio: neural building blocks for speaker diarization. In *Proceedings of ICASSP 2021*, pages 8210–8214.

Carnegie Mellon University Speech Group. 2014. The CMU pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. Version 0.7b, accessed 2026-03-04.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Anna Favaro, Mia Cao, Tewodros Gessesse, Sanjana Ghosh, Alejandro Luzardo, Laureano Moro-Velazquez, Ankur Butala, Jesus Villalba, and Najim Dehak. 2023. Automatic detection of Alzheimer's disease from speech and natural language: a systematic literature review. *Frontiers in Aging Neuroscience*, 15:1210894.

Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60.

Maria Luisa Gorno-Tempini, Argye E. Hillis, Sandra Weintraub, Andrew Kertesz, Mario Mendez, Stefano F. Cappa, Jennifer M. Ogar, Jonathan D. Rohrer, Sandra Black, Bradley F. Boeve, Facundo Manes, Nina F. Dronkers, Rik Vandenberghe, Katya Rascovsky, Karalyn Patterson, Bruce L. Miller, David S. Knopman, John R. Hodges, M.-Marsel Mesulam, and Murray Grossman. 2011. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014.

Chenxu Guo, Jiachen Lian, Yisi Liu, Baihe Huang, Shriyaa Narayanan, Cheol Jun Cho, and Gopala Anumanchipalli. 2026. HuPER: A human-inspired framework for phonetic perception. arXiv preprint arXiv:2602.01634. *Preprint*, arXiv:2602.01634.

Chenxu Guo, Jiachen Lian, Xuanru Zhou, Zoe Ezzes, Jet Vonk, and Gopala Anumanchipalli. 2025. Dysfluent WFST: A framework for zero-shot speech dysfluency transcription and detection. In *Proceedings of Interspeech 2025*.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. `https://spacy.io`.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5338–5348.

Jiachen Lian and Gopala Anumanchipalli. 2024. Towards hierarchical spoken language disfluency modeling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 539–551. Association for Computational Linguistics.

Jiachen Lian, Cong Feng, Najim Farooqi, Sherry Li, Aparna Kashyap, Christina J. Cho, Peter Wu, Reid Netzorg, Tian Li, and Gopala K. Anumanchipalli. 2023. Unconstrained dysfluency modeling for dysfluent speech transcription and detection. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jiachen Lian, Xuanru Zhou, Zoe Ezzes, Jet Vonk, Benjamin Morin, Diane P. Baquirin, Zachary Miller, Maria Luisa Gorno-Tempini, and Gopala Anumanchipalli. 2024. SSDM: Scalable speech dysfluency modeling. In *Advances in Neural Information Processing Systems*, volume 37.

Jiachen Lian, Xuanru Zhou, Chenxu Guo, Zhaoheng Ye, Zoe Ezzes, Jet Vonk, Benjamin Morin, Diane Baquirin, Zachary Miller, Maria Luisa Gorno-Tempini, and Gopala K. Anumanchipalli. 2025. Automatic detection of articulatory-based disfluencies in primary progressive aphasia.

M.-Marsel Mesulam. 2013. Primary progressive aphasia and the language network: the 2013 H. Houston Merritt lecture. *Neurology*, 81(5):456–462.

Naomi Nevler, Sharon Ash, Chelsea Jester, David J. Irwin, Mark Liberman, and Murray Grossman. 2019. Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology*, 6(8):1518–1529.

Linda E. Nicholas and Robert H. Brookshire. 1993. A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech, Language, and Hearing Research*, 36(2):338–350.

Franziska Peters, William Richard Bevan-Jones, Gemma Threlfall, Jennifer M. Harris, Julie S. Snowden, Matthew Jones, Jennifer C. Thompson, Daniel J. Blackburn, and Heidi Christensen. 2025. Automatic detection and sub-typing of primary progressive aphasia from speech: Integrating task-specific features and spatio-semantic graphs. In *Proceedings of Interspeech 2025*.

Yen-Khang Phan, Trung-Nghia Le, Nhien-An Le-Khac, and Yi-Hsuan Yang. 2024. Automated content assessment and feedback for Finnish L2 learners in a picture description task. In *Proceedings of Interspeech 2024*, pages 317–321.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Neguine Rezaii, David Hochberg, Megan Quimby, Bonnie Wong, Michael Brickhouse, Alexandra Tourouto-glou, Bradford C. Dickerson, and Phillip Wolff. 2024. Artificial intelligence classifies primary progressive aphasia from connected speech. *Brain*, 147(9):3070–3082.

Jeanine Romano, Jeffrey D. Kromrey, Jesse Coraggio, and Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In *Annual Meeting of the Florida Association of Institutional Research*, pages 1–33.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Charalambos Themistocleous, Marie Eckerström, and Dimitrios Kokkinakis. 2018. Identification of PPA and its variants using machine learning. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3399–3404, Miyazaki, Japan.

Charalambos Themistocleous, Bronte Ficek, Kimberly Webster, Dirk-Bart den Ouden, Argye E. Hillis, and Kyrana Tsapkini. 2021. Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease*, 79(3):1185–1194.

Jet M. J. Vonk, Jiachen Lian, Christina J. Cho, Giada Antonicelli, Zoe Ezzes, Liesbeth D. Wauters, Willa Keegan-Rodewald, Stephen Lynn Kurteff, Diego A. Rodriguez, Nina Dronkers, and Maria Luisa Gorno-Tempini. 2026. AI-based speech error detection to differentiate primary progressive aphasia variants. *medRxiv*. Preprint, February 2026.

Jet M. J. Vonk, Jiachen Lian, Zoe Ezzes, Christina J. Cho, Benjamin T. Morin, Rian Bogley, Zachary Miller, Maria Luisa Mandelli, Gopala Anumanchipalli, and Maria Luisa Gorno-Tempini. 2025. Automated lexical dysfluency analysis to differentiate primary progressive aphasia variants. In *Alzheimer's Association International Conference*.

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

Zhaoheng Ye, Jiachen Lian, Xuanru Zhou, Jiarui Zhang, Han Li, Sherry Li, Chenxu Guo, Ayan Das, Peter Park, Zoe Ezzes, Maria Luisa Gorno-Tempini, and Gopala Anumanchipalli. 2025. Seamless dysfluent speech text alignment for disordered speech analysis. In *Proceedings of Interspeech 2025*.

Jiarui Zhang, Xuanru Zhou, Jiachen Lian, Sherry Li, Wen Li, Zoe Ezzes, Rian Bogley, Liesbeth Wauters, Zachary Miller, Jet Vonk, Maria Luisa Gorno-Tempini, and Gopala Anumanchipalli. 2025.

Analysis and evaluation of synthetic data generation in speech dysfluency detection. In *Proceedings of Interspeech 2025*.

Xuanru Zhou, Aparna Kashyap, Sherry Li, Anshita Sharma, Benjamin Morin, Diane Baquirin, Jet Vonk, Zoe Ezzes, Zachary Miller, Maria Luisa Gorno-Tempini, Jiachen Lian, and Gopala Anumanchipalli. 2024. YOLO-stutter: End-to-end region-wise speech dysfluency detection. In *Proceedings of Interspeech 2024*, pages 937–941.

Vitor C. Zimmerer, Mark Wibrow, and Rosemary A. Varley. 2016. Formulaic language in people with probable Alzheimer's disease: a frequency-based approach. *Journal of Alzheimer's Disease*, 53(3):1145–1160.

A Gold Scene Graph

The gold scene graph encodes the canonical semantic content of the WAB Picnic stimulus as a JSON structure with 21 nodes, 23 directed edges, 6 attributes, and 12 content units. Entity matching uses canonical labels and synonym sets; node IDs are stable across versions.

Listing 1: Gold scene graph (JSON).


```
{
 "scene": "picnic", "schema_version": "v0",
 "nodes": [
  {"id":"family_1","type":"group", "label":"family"},
  {"id":"man_1",  "type":"person","label":"man"},
  {"id":"woman_1", "type":"person","label":"woman"},
  {"id":"boy_1",  "type":"person","label":"boy"},
  {"id":"girl_1", "type":"person","label":"girl"},
  {"id":"dog_1",  "type":"animal","label":"dog"},
  {"id":"kite_1", "type":"object","label":"kite"},
  {"id":"string_1","type":"object","label":"kite string"},
  {"id":"blanket_1","type":"object","label":"blanket"},
  {"id":"basket_1","type":"object","label":"picnic basket"},
  {"id":"food_1", "type":"object","label":"food"},
  {"id":"drink_1", "type":"object","label":"drink"},
  {"id":"book_1", "type":"object","label":"book"},
  {"id":"ball_1", "type":"object","label":"ball"},
  {"id":"tree_1", "type":"object","label":"tree"},
  {"id":"grass_1", "type":"place", "label":"grass"},
  {"id":"sky_1",  "type":"place", "label":"sky"},
  {"id":"lake_1", "type":"place", "label":"lake"},
  {"id":"boat_1", "type":"object","label":"boat"},
  {"id":"house_1", "type":"object","label":"house"},
  {"id":"path_1", "type":"place", "label":"path"}
 ],
 "edges": [
  {"subj":"man_1",  "rel":"member_of", "obj":"family_1"},
  {"subj":"woman_1", "rel":"member_of", "obj":"family_1"},
  {"subj":"boy_1",  "rel":"member_of", "obj":"family_1"},
  {"subj":"girl_1", "rel":"member_of", "obj":"family_1"},
  {"subj":"blanket_1","rel":"on",      "obj":"grass_1"},
  {"subj":"basket_1", "rel":"on",      "obj":"blanket_1"},
  {"subj":"food_1", "rel":"on",       "obj":"blanket_1"},
  {"subj":"drink_1", "rel":"on",      "obj":"blanket_1"},
  {"subj":"woman_1", "rel":"sitting_on", "obj":"blanket_1"},
  {"subj":"man_1",  "rel":"sitting_on", "obj":"blanket_1"},
  {"subj":"woman_1", "rel":"holding", "obj":"book_1"},
  {"subj":"woman_1", "rel":"reading", "obj":"book_1"},
  {"subj":"boy_1",  "rel":"holding",  "obj":"string_1"},
  {"subj":"string_1", "rel":"attached_to","obj":"kite_1"},
  {"subj":"kite_1", "rel":"in",       "obj":"sky_1"},
  {"subj":"dog_1",  "rel":"on",       "obj":"grass_1"},
  {"subj":"dog_1",  "rel":"chasing",  "obj":"ball_1"},
  {"subj":"ball_1", "rel":"on",       "obj":"grass_1"},
  {"subj":"tree_1", "rel":"on",       "obj":"grass_1"},
  {"subj":"lake_1", "rel":"near",     "obj":"grass_1"},
  {"subj":"boat_1", "rel":"on",       "obj":"lake_1"},
  {"subj":"house_1", "rel":"near",    "obj":"path_1"},
  {"subj":"path_1", "rel":"leads_to", "obj":"grass_1"}
 ],
 "attributes": [
  {"node":"kite_1", "attr":"state", "value":"flying"},
  {"node":"dog_1",  "attr":"activity","value":"running"},
  {"node":"family_1", "attr":"activity","value":"picnicking
       "},
  {"node":"blanket_1","attr":"location","value":"foreground
       "},
  {"node":"kite_1", "attr":"location","value":"upper_sky"},
  {"node":"lake_1", "attr":"location","value":"background"}
 ]
}
```


B Content Unit Definitions

The 12 CUs follow Nicholas and Brookshire (1993): nine edge CUs and three attribute CUs.

Table 10: Content units.

ID	Subj	Relation	Obj/Value
CU_E01	blanket	on	grass
CU_E02	woman	sitting_on	blanket
CU_E03	man	sitting_on	blanket
CU_E04	woman	reading	book
CU_E05	boy	holding	string
CU_E06	string	attached_to	kite
CU_E07	kite	in	sky
CU_E08	dog	chasing	ball
CU_E09	boat	on	lake
CU_A01	kite	state	flying
CU_A02	dog	activity	running
CU_A03	family	activity	picnicking

C LLM Baseline

All tasks used few-shot prompting (two examples per class for the binary tasks; zero-shot for the four-way task) with `qwen3.5-plus`; transcripts were truncated to 4,000 characters. The label `nflv` in the binary prompt below denotes the pooled nfvPPA+lvPPA class.

C.1 System Prompts

Listing 2: Control vs. svPPA.

```
You are a clinical linguist. Below is a speech transcript
from a picture description task (picnic scene). Classify
into exactly ONE of:
- control (healthy, normal picture description)
- svPPA (word-finding difficulty, empty speech, semantic
  errors)
Reply with only one token: control or svPPA
```

Listing 3: svPPA vs. nfv+lv (binary). nflv = pooled nfvPPA + lvPPA.

```
You are a clinical linguist. Below is a speech transcript
from a picture description task (picnic scene). Classify
into exactly ONE of:
- svPPA (semantic variant: word-finding difficulty, empty
  speech)
- nflv (nfvPPA or lvPPA pooled: grammar / speech-rate /
  phonological issues)
Reply with only one token: svPPA or nflv
```

C.2 Few-Shot Examples (svPPA vs. nflv)

Listing 4: Few-shot examples.

```
Example 1 (svPPA):
Transcript: They're having a... the outdoor thing. With the
      ... I forget
the word. The thing you put food on.
Classification: svPPA

Example 2 (nflv):
Transcript: The fam... the family is at the picnic. He is
      put- putting
the basket. She is spread the blanket. The child are play.
Classification: nflv
```

C.3 API Call and Response Parsing

Listing 5: API call.

```
import openai
client = openai.OpenAI(api_key=api_key)
r = client.chat.completions.create(
    model="qwen3.5-plus",
    messages=[
        {"role": "system", "content": prompt_template},
        {"role": "user",
         "content": f"Transcript:\n\n{text[:4000]}"},
    ],
    max_tokens=20,
)
raw = (r.choices[0].message.content or "").strip()
```

Listing 6: Response parsing.

```
import re

def parse_llm_response(raw: str, task: str) -> str:
    raw_up = raw.upper()
    if task == "binary_sv_nflv":
        if "SVPPA" in raw_up or "SEMANTIC" in raw_up:
            return "svPPA"
        if "NFLV" in raw_up or "NON-SEMANTIC" in raw_up:
            return "nflv"
        return "unknown"
    if task == "binary_ctrl_sv":
        if "SVPPA" in raw_up:
            return "svPPA"
        if "CONTROL" in raw_up:
            return "control"
        return "unknown"
    for src, tgt in [("CONTROL","control"),("SVPPA","svPPA"),

                     ("NFVPPA","nfvPPA"),("LVPPA","lvPPA")]:
        if src in raw_up:
            return tgt
    return "unknown"
```

D Layer 1 Feature Definitions

Layer 1 extracts 20 task-agnostic proxy measures from the Whisper transcript and pyannote timing stream. Features are z-scored against training-fold statistics within each cross-validation split.

Listing 7: Extraction constants.

```
FILLED_PAUSES = {
    "um","uh","er","ah","eh","hm","hmm","uhm","umhm"
}
WORD_FINDING_PAUSE_THRESHOLD = 1.0 # seconds
MIN_PAUSE_GAP = 0.3              # seconds
```

Table 11: Layer 1 proxy features.

Idx	Name	Description
0	wpm	Words per minute
1	articulation_rate	Phonemes per second
2	mean_pause_dur	Mean pause duration (s)
3	pause_speech_ratio	Pause time / speech time
4	init_latency	Time to first word (s)
5	mlu	Mean length of utterance
6	median_utt_len	Median utterance length
7	clause_per_utt	Clauses per utterance
8	abandoned_utt_rate	Abandoned utterance proportion
9	utterance_count	Total utterance count
10	open_closed_ratio	Open- / closed-class ratio
11	content_word_density	Content word proportion
12	ttr	Type–token ratio
13	pronoun_ratio	Pronoun proportion
14	fn_word_omission	Function-word omission rate
15	morph_error_rate	Morphological error rate
16	dep_completeness	Dependency parse completeness
17	wf_pause_dur	Mean duration of pauses $\geq$1 s
18	filled_pause_freq	Filled pauses per word
19	semantic_distance	Distance to expected entities

E Layer 2 Construct Definitions and Weights

Layer 2 maps Layer 1 z-scores to ten clinician-aligned constructs via $\mathrm{L2}_c = (\tanh(\bar{z}_c) + 1)/2 \in [0, 1]$, where higher values indicate greater impairment. All weights are uniform (1.0).

Table 12: Layer 2 constructs and Layer 1 feature indices.

Construct	Layer 1 indices
Reduced speech rate	0, 1, 2, 3, 4
Reduced length/complexity	5, 6, 7, 9
Agrammatism	7, 14, 15, 16
Paragrammatism	12, 15, 16
Anomia	2, 8, 17, 18
Empty speech	10, 11, 12, 13
Semantic paraphasias	10, 12, 19
Phonemic paraph./neol.	15, 18
Self-correction	4, 8, 18
Overall impairment	0, 5, 10, 11, 19

Listing 8: Layer 2 computation.

```
import numpy as np

zvec = (vec - population_mean) / population_std

def compute_l2(zvec, weights):
    z = (sum(zvec[i]*w for i,w in weights)
         / sum(w for _,w in weights))
    return (np.tanh(z) + 1) / 2

s_sem_l2 = 1.0 - np.mean(
    [compute_l2(zvec, w) for w in all_weights]
)
```

F Phonological Adequacy (S_{ph})

Expected phonemes are from CMU dictionary G2P of the Whisper transcript (character-level fallback for OOV items); observed phonemes are from the HuPER acoustic stream with silence tokens removed.

$$S_{\text{ph}} = \max\left(0,\ 1 - \frac{\text{Lev}(\hat{P},\ P^*)}{\max(|\hat{P}|,\ |P^*|)}\right) \tag{5}$$

Listing 9: S_{ph} computation.

```
def compute_s_ph(transcript, huper_phonemes):
    expected = g2p_transcript(transcript)
    observed = [p for p in huper_phonemes if p != "SIL"]
    edit_dist = levenshtein(expected, observed)
    max_len  = max(len(expected), len(observed), 1)
    per  = edit_dist / max_len
    s_ph = max(0.0, min(1.0, 1.0 - per))
    return s_ph, {
        "per": per, "edit_dist": edit_dist,
        "n_exp": len(expected), "n_obs": len(observed),
    }
```

G Semantic Adequacy and Dissociation Index

G.1 Scene-graph score ($S_{\text{sem}}^{\text{graph}}$)

CU coverage is weighted above entity recall following Nicholas and Brookshire (1993):

$$S_{\text{sem}}^{\text{graph}} = 0.4 \cdot \frac{|\text{matched nodes}|}{21} + 0.6 \cdot \frac{|\text{matched CUs}|}{12} \tag{6}$$

G.2 Composite S_{sem} and dissociation index

$$S_{\text{sem}} = \alpha \cdot S_{\text{sem}}^{\text{graph}} + (1-\alpha) \cdot S_{\text{sem}}^{\text{L2}}, \quad \alpha = 0.5 \tag{7}$$

where $S_{\text{sem}}^{\text{L2}} = 1 - \text{mean}(\text{L2}_1, \ldots, \text{L2}_{10})$. $\Delta = S_{\text{ph}} - S_{\text{sem}}$ is an expected directional summary: positive values tend to occur for svPPA (preserved phonology, degraded semantics), negative values for nfvPPA, and values near zero for controls and lvPPA. The Δ-only ablation in Section 6.1 shows that this directional tendency is too noisy to classify subtypes on its own.

H Classifier Hyperparameters

Listing 10: Logistic regression.

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(
    penalty="l2", C=1.0, class_weight="balanced",
    max_iter=1000, solver="lbfgs", random_state=42,
)
```

Listing 11: Random forest.

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(
    n_estimators=100, max_depth=6,
    class_weight="balanced", random_state=42,
)
```

Listing 12: Cross-validation protocol.

```
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.impute        import SimpleImputer

cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=
     seed)
# Scaler and imputer fit on train fold only.
# Results averaged over 10 seeds (0--9).
```

H.1 Computational Infrastructure and Budget

All classical machine learning models (Logistic Regression and Random Forest) possess a negligible parameter footprint ($< 10^5$ parameters). Heavy workloads, specifically speech-to-text transcription via Whisper (Radford et al., 2023) and speaker diarization via pyannote.audio (Bredin et al., 2021) were executed using an institutional Linux cluster equipped with NVIDIA A100 GPUs. The total computational budget for processing the 254 recordings across all cross-validation folds and baseline iterations was less than 5 GPU hours and 10 CPU hours.

I Statistical Conventions

Effect sizes use Cliff's δ:

$$\delta = \frac{\#(x_i > y_j) - \#(x_i < y_j)}{n_1 \cdot n_2} \tag{8}$$

Thresholds from Romano et al. (2006): $|\delta| < 0.147$ negligible, 0.147 to 0.330 small, 0.330 to 0.474 medium, ≥ 0.474 large. Group differences use two-sided Mann–Whitney U; significance: $*\,p < 0.05$, $**\ p < 0.01$, $***\,p < 0.001$.

J Data Schema

Manifest CSV: `subject_id`, `session`, `subtype`, `label` (0–3), `audio_path`, `task`.

Pipeline output: per recording, `{stem}_feature_extraction.txt` with columns *Timestamp*, *Speaker*, *ASR Transcription*, *Phonemic Encoding*. Only participant-labelled segments are used.

Feature table (`ppa_dissociation_fixed_l2.csv`): `subject_id`, `session`, `subtype`, `label`, `s_ph`, `s_sem`, `s_sem_graph`, `s_sem_layer2_adequacy`,

delta, phoneme_error_rate, silence_ratio, and l2_* for each of the ten Layer 2 construct scores.

K Simulated svPPA Degradation

Control transcripts are degraded by replacing specific lexical items with underspecified hypernyms and inserting filled pauses, mimicking the empty-speech signature of svPPA. Default: replace_ratio = 0.70, insert_filler_ratio = 0.12–0.15; reproducibility via seed + hash(subject_id, session).

Listing 13: Degradation function.

```python
import re, random

REPLACEMENTS = [
    (r"\b(boy|girl|kid|child|son|daughter)\b","person"),
    (r"\b(man|woman|father|mother|dad|mom"
     r"|lady|guy)\b",                    "person"),
    (r"\b(dog|puppy)\b",                 "thing"),
    (r"\b(kite|string|line)\b",          "thing"),
    (r"\b(blanket|rug|mat)\b",           "thing"),
    (r"\b(basket|picnic basket)\b",      "thing"),
    (r"\b(sandwich|food|lunch)\b",       "thing"),
    (r"\b(juice|soda|drink)\b",          "thing"),
    (r"\b(book|magazine)\b",             "thing"),
    (r"\b(ball|tree|grass|sky|lake|water"
     r"|pond|boat|house|home|path|road"
     r"|trail)\b",                       "thing"),
]
FILLERS = ["um", "uh", "you know"]

def degrade_transcript(
        text, replace_ratio=0.70,
        insert_filler_ratio=0.13, seed=None):
    if seed is not None:
        random.seed(seed)
    for pat, repl in REPLACEMENTS:
        text = re.sub(
            pat,
            lambda m: (repl if random.random()
                       < replace_ratio else m.group(0)),
            text, flags=re.IGNORECASE,
        )
    tokens, out = re.findall(r"[a-zA-Z]+", text), []
    for i, t in enumerate(tokens):
        out.append(t)
        if (i < len(tokens)-1
                and random.random() < insert_filler_ratio):
            out.append(random.choice(FILLERS))
    return " ".join(out)
```

L Evaluation Metrics

Accuracy is the fraction of correct predictions. Precision, recall, and F1 are binary (positive = svPPA) for binary tasks and macro-averaged for multiclass tasks. ROC-AUC uses predicted $P(\text{svPPA})$. All results are mean $\pm$ std over 10 runs (seeds 0–9).