

Unlocking the Spoken Style: Mechanistic Localization and Lightweight Control of Orality in LLMs

Zhigen Li^{1,2}, Yanmeng Wang², Jianxiang Peng¹, Yong Zhang²

Dan Shi¹, Ning Cheng², Shaojun Wang², Deyi Xiong^{1,3,*}

¹TJUNLP Lab, School of Computer Science and Technology, Tianjin University, China

²Ping An Technology, Shenzhen, China

³Shenzhen Loop Area Institute

{zhigenli, dyxiong}@tju.edu.cn

Abstract

Voice agents typically use text-first pipeline for safety, but large language model (LLM) outputs are often overly written, making even high-quality text-to-speech (TTS) sound robotic. This is especially problematic in domains such as healthcare and sales, where different audiences require graded orality control. Text-level rewriting can control spoken style, but resources mostly study the reverse direction (spoken-to-written transfer), and its internal locus in LLMs remains unclear. We study English written-to-spoken rewriting with graded orality control. We introduce WSPairs, 4,709 written-spoken pairs from spontaneous conversations, annotated with 0–5 Orality Levels using seven sociolinguistic features. Through causal activation patching and head-level attribution, we find that orality concentrates in late layers and a sparse set of high-influence attention heads. We then propose CAS, a no-weight-update calibrated activation-steering method, and LAS, a learned activation-steering method. CAS improves exact level targeting and reduces bias relative to few-shot prompting, although target ordering remains imperfect at the highest levels. LAS uses 0.004%–0.006% trainable parameters, obtains higher Orality Level Accuracy (L-Acc) than full SFT on Qwen-3B and Mistral-7B, remains the strongest lightweight method on Llama-1B, and leads SFT by up to 3.9 L-Acc points out of domain. These results support localized, efficient control of text-level spoken style. Dataset and code are available.¹

1 Introduction

Many voice-agent systems use a text-first pipeline for safety: a large language model (LLM) drafts a reply, and a text-to-speech (TTS) system speaks it. This separation can simplify text-side review and control in regulated settings (Li et al., 2025). The problem is that written and spoken language

serve different goals (Biber, 1991; Akinnaso, 1985). Written text is compact and information-dense, while spoken conversation uses discourse markers, hedges, and fragmented clauses to support real-time interaction. This mismatch makes voice agents sound overly written and robotic.

Real applications such as healthcare, sales, and customer service also need graded orality control: the same content should be rendered with different levels of orality for different audiences. For example, the written prompt “Are you breathing faster?” might be rendered as “Well, do you feel like you’re breathing a bit faster?” for a typical adult, and as “Okay, um, is your breathing fast? Like, are you breathing faster?” for an older adult with higher support needs (Fig. 1, left). Three gaps separate current methods from graded oral control. **(L1)** Style-transfer benchmarks mostly use binary spoken or written labels in the spoken-to-written direction (Rao and Tetreault, 2018; Briakou et al., 2021; Logacheva et al., 2022), so graded written-to-spoken supervision is missing. **(L2)** Mechanistic interpretability has localized factual recall (Meng et al., 2022), refusal (Wang et al., 2025), and emotion concepts (Zou et al., 2025), but not surface-style attributes like orality. **(L3)** Full or PEFT fine-tuning per target level is expensive and transfers poorly across domains, and existing activation-side controllers (Turner et al., 2024; Wu et al., 2024) are not designed for graded stylistic attributes.

These gaps converge on a single question: where in an LLM is spoken style rendered, and can it be controlled by intervening on a small set of internal components? Three challenges make this hard. **(C1)** No parallel written-spoken corpus with graded intensity exists. **(C2)** Stylistic rendering may be more diffuse than the discrete lookup behaviors localized in prior work. **(C3)** Orality is an ordinal, relative attribute (Stevens, 1946; Thurstone, 2017), so a controller must track levels monotonically rather than flip a binary switch. We make

*Corresponding author.

¹<https://github.com/tjunlp-lab/llm-spoken>

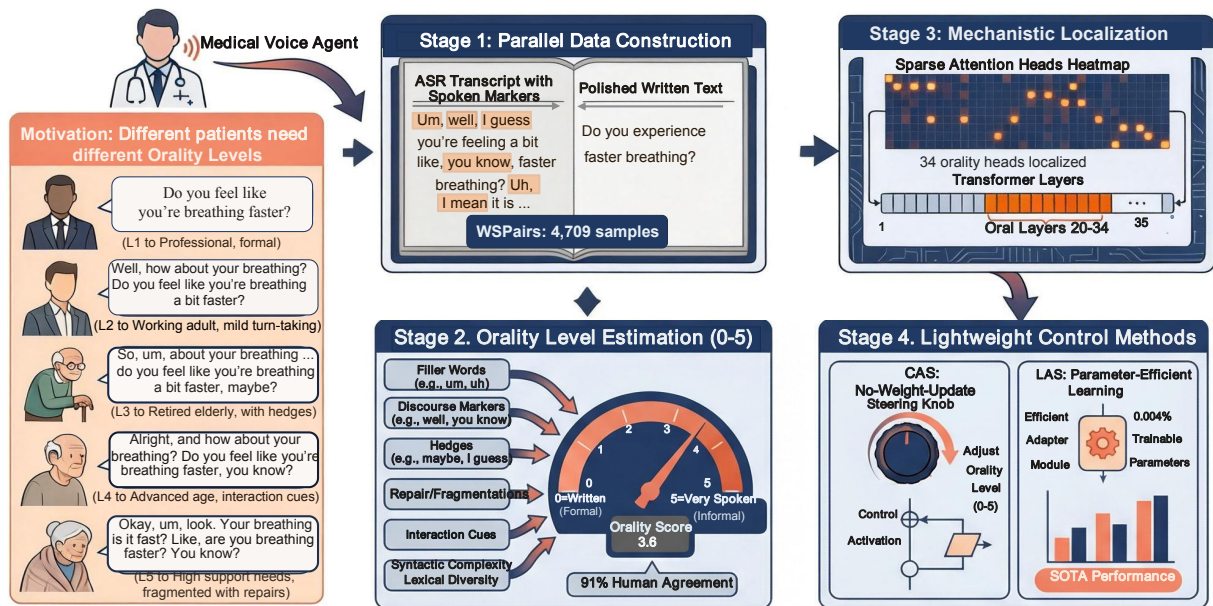


Figure 1: Our work overview. **Left**: Examples of patient-adaptive orality levels (L1-L5). **Middle**: we construct a parallel dataset (WSPairs) using LLMs for initial annotation followed by manual verification; **Right**: to achieve precise control we localize orality-specific heads and layers then implement lightweight steering methods (CAS/LAS).

three contributions. (1) We build **WSPairs**, a 4,709-pair English corpus spanning customer service, medical, and negotiation dialogue, paired with a seven-feature Orality-Level estimator that agrees with human pairwise judgments 91% of the time. (2) Using causal activation patching and head-level contrastive attribution, we find that orality concentrates in a sparse high-influence set of late layers and attention heads, replicating on three backbones (Llama-1B, Qwen-3B, Mistral-7B).² (3) Building on this localization, we propose two lightweight controllers: **CAS**, which calibrates activation steering without weight updates, and **LAS**, a learned activation-steering method. CAS improves exact target-level accuracy over few-shot prompting and prior single-layer steering, while LAS obtains higher L-Acc than full SFT on two of three backbones with only 0.004%–0.006% trainable parameters, remains the strongest lightweight method on the third, and leads three leave-one-domain-out folds by up to 3.9 points over SFT.

2 Related Work

Style and formality transfer. Sociolinguistic studies have long characterized written–spoken differences via features such as hedges, discourse markers, and fragmentation (Akinnaso, 1982; Biber,

²For brevity, we refer to Llama-3.2-1B-Instruct, Qwen-2.5-3B-Instruct, and Mistral-7B-Instruct-v0.3 as Llama-1B, Qwen-3B, and Mistral-7B throughout.

1991). NLP work on this register axis falls into three lines. Informal-to-formal rewriting uses parallel benchmarks (Rao and Tetreault, 2018; Briakou et al., 2021; Logacheva et al., 2022) with binary formal/informal labels. Spoken-to-written resources (Guo et al., 2023) still use categorical labels and mostly cover the reverse direction. End-to-end speech-side modeling (Zhang et al., 2023) operates on acoustic features rather than text. None of these supports graded ordinal control in the deployment-aligned written-to-spoken direction, which is the gap WSPairs and our controllers fill.

Mechanistic localization of behaviors. Mechanistic interpretability (Elhage et al., 2021; Wu et al., 2023; Shi et al., 2026b,a) uses causal mediation and activation patching to attribute behaviors to specific components (Hicks and Tingley, 2011; Meng et al., 2022). Prior work has localized factual recall and editing (Meng et al., 2022; Tamayo et al., 2024; Wang et al., 2025), induction heads and IOI (Olsson et al., 2022; Wang et al., 2023), task vectors (Todd et al., 2024), and emotion concepts (Zou et al., 2025). These behaviors concentrate in sparse middle-to-late-layer components. The prior mechanistic studies cited here do not examine graded orality. Our results place orality in the same late-layer pattern, extending mechanistic interpretability from discrete lookup targets to a continuous surface-style attribute.

Controllable generation and activation steer-

ing. Controllable generation falls into three families. Decoding-time controllers (Dathathri et al., 2020; Krause et al., 2021; Yang and Klein, 2021) guide generation with an auxiliary classifier per token, but are designed for binary or categorical attributes and require an extra pass per token. Weight-side methods (full SFT, PEFT adapters (Hu et al., 2022; Liu et al., 2022; Li et al., 2022)) keep millions of parameters per setting and are hard to calibrate to an intensity. Activation-side methods intervene directly on hidden states: Activation Addition (Turner et al., 2024) adds a difference-of-means vector at one grid-searched layer with a fixed scalar, and ReFT (Wu et al., 2024) learns a rank- r read-write loop on the residual stream; parameter-space LoRA interpolation (Wortsman et al., 2022; Huang et al., 2024) mixes globally trained adapters. CAS extends Activation Addition along two axes: it injects into the full set of localized oral layers and calibrates the coefficient as a polynomial of the target level. LAS replaces ReFT’s rank- r loop with an element-wise affine residual on causally-identified layers and heads. A design-level comparison of all baselines is in Appendix K (Table 14).

3 Preliminaries

This section describes how we construct parallel data (§3.1) and how we derive and validate our Orality Level estimators (§3.2).

3.1 WSPairs Construction

Dataset format. Each instance in WSPairs is a tuple

$$\langle x^w, x^s, \ell^s \rangle,$$

where x^w is a written-style text, x^s is its spoken-style counterpart conveying the same content, and $\ell^s \in \{0, 1, \dots, 5\}$ is the Orality Level assigned to x^s . The 7-dimensional feature vector $\mathbf{f}^s$ is deterministically computed from x^s (§3.2) rather than stored as an independent annotation.

Three-step construction. (i) **Collect** x^s . We aggregated 217,000 human agent utterances from *SpokenWoZ* (Si et al., 2023) (task-oriented), *MediToD* (Saley et al., 2024) (medical), and *Bargaining* (Heddaya et al., 2023) (negotiation), while preserving fillers and interaction markers typical of ASR transcripts. We removed utterances shorter than 15 words before constructing candidate pairs. (ii) **Annotate** x^w . We use five LLMs³ to con-

³We used Gemini-2.5-Flash, Gemini-3-Pro-Preview, GPT-5 mini, DeepSeek Reasoner, and Llama-3.1-8B Instruct to

Domain	Train	Valid	Test	L1	L2	L3	L4	L5
All	3183	471	1055	17.9	<u>20.1</u>	37.5	5.1	19.5
MediTOD	1382	215	456	6.9	<u>28.0</u>	37.1	4.0	24.1
SpokenWoZ	1145	163	386	31.0	19.2	37.0	4.2	8.7
Bargaining	656	93	213	18.5	4.9	39.3	8.8	<u>28.5</u>

Table 1: WSPAIRS splits and spoken-side Orality Level distribution by domain (%). Level 3 is modal and Level 4 is rarest; SpokenWoZ skews lower, Bargaining higher, and MediTOD lies between. Level 0 denotes written inputs and is omitted.

vert x^s into a written counterpart x^w by removing spoken-language markers while preserving named entities and factual content (see Appendix A for the prompt). We then filtered the pairs using semantic duplication, and stylistic criteria. Specifically, we removed pairs with $\text{BERTSCORE}(x^w, x^s) < 0.85$, pairs whose x^s had Jaccard similarity greater than 0.7 with another pair, pairs with $\text{level}^{\text{written}} > 0$, and pairs with $\text{level}^{\text{spoken}} = 0$. Finally, five human annotators reviewed and corrected the written text x^w produced by LLMs. A random subset of 1,247 pairs received a second independent annotation; exact-match agreement had Cohen’s $\kappa = 0.92$ (see Appendix A for annotation protocol details). (iii) **Assign ℓ^s and $\mathbf{f}^s$.** We compute $\mathbf{f}^s$ from x^s and obtain $\ell^s = g(\mathbf{f}^s)$ via the feature-based estimator $g(\cdot)$ (§3.2). We use $g(\cdot)$ rather than direct single-text human annotation because orality is psychometrically relative: rating one text on an absolute 0–5 scale is noisy and inconsistent across raters; see Section 3.2 for more details.

We finally collected 4,709 pairs. Table 1 reports the per-domain splits and level distribution. Appendix Table 8 in §C shows pairs across Levels 1–5, and Appendix Table 10 shows an example with its seven feature annotations. Compared with prior corpora (Rao and Tetreault, 2018; Briakou et al., 2021; Lyu et al., 2021; Logacheva et al., 2022; Guo et al., 2023), WSPairs uniquely combines the written-to-spoken direction, ordinal multi-level labels, and real ASR transcripts on the spoken side; Appendix D and Table 9 give the full comparison.

3.2 The Orality Level Estimator

Seven sociolinguistic features. We quantify oral-ity with seven features grounded in sociolin-guistic theory (Akinaso, 1982, 1985; van Mil-tenburg et al., 2018). Five are spoken-oriented and higher values mean more spoken: (1) fillers (*um*,
cover a broad range of written styles.

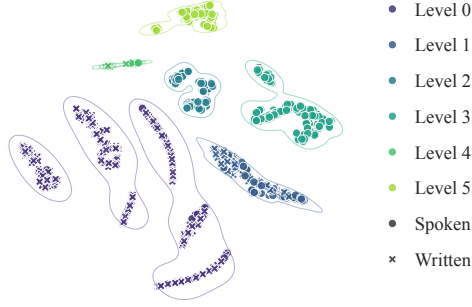


Figure 2: The six feature-derived Orality Levels form separated regions in a t-SNE projection of 2,000 random WSPairs samples, providing a qualitative view of the estimator’s ordinal partition.

uh); (2) interaction markers (*you know, right?*); (3) stance/hedges (*I guess, maybe*); (4) lexical reduction (*wanna*); (5) repetition (*any any*). Two are written-oriented: (6) propositional idea density (PID) and (7) noun–verb ratio (NVR). Let $\mathbf{f}(x) \in \mathbb{R}^7$ be the normalized feature vector: features 1–5 measure the frequency of spoken-language phenomena (Biber, 1991); features 6–7 use standard algorithmic definitions; all are normalized to comparable scales (Appendix E). Unlike features 1–5, PID and NVR are not triggered by specific lexical items: they are POS-/dependency-based structural statistics that capture how information is packaged syntactically, so the estimator is not driven by surface cue words alone (§6). Appendix Table 10 gives a worked example, including the formula, trigger counts, and final score for each of the seven features.

Levels via clustering. The Orality Level maps any text to one of six discrete levels. We combine the seven normalized features via entropy weighting (EWM, Appendix E) into a single orality score, apply weighted PCA, and run k-means with $k=6$ chosen by silhouette score (Appendix F). Clusters are sorted by orality score and mapped to Levels 0–5, giving the reproducible level estimator $g(x) \rightarrow \{0, \dots, 5\}$. Before stylistic filtering, $g(\cdot)$ already separates WSPairs: 89% of x^{written} land at Level 0 and 82% of x^{spoken} at Levels 1–5. In t-SNE space, the six levels form well-separated clusters with clear boundaries (Fig. 2); across Levels 0–5, the seven feature distributions show the intended aggregate direction but overlap and are not uniformly monotonic (Appendix Fig. 8).

Pairwise model and human evaluation. Orality is **ordinal and relative** (Stevens, 1946; Thurstone,

Evaluator	Train signal	Pair Acc
$g(\cdot)$ (EWM + KMeans, ours)	7-feature + clustering	0.910
<i>Family 1: LLM zero-shot judges (protocol = Lst. 2)</i>		
DeepSeek-v4-Pro	N/A (API)	0.850
Gemini-3.5-Flash	N/A (API)	0.825
GPT-5.4-Nano	N/A (API)	0.810
LLM ensemble (3-way majority)	N/A (vote)	0.885
<i>Family 2: Bradley-Terry reward models (10k LLM-ens. pairs)</i>		
RoBERTa-base (125M)	BT pairwise	0.795
Llama-1B	BT pairwise	0.440
Qwen-3B	BT pairwise	0.805
Mistral-7B	BT pairwise	0.855

Table 2: On 200 held-out human pairwise judgments, $g(\cdot)$ has the highest pair accuracy among the evaluated proxies. Its accuracy is 2.5 points above the LLM ensemble and 5.5 points above the best Bradley–Terry reward model. All LLM judges receive Listing 2 verbatim with only the two texts substituted.

2017): annotators can reliably compare two texts, but assigning an absolute label on a single text is noisy. We therefore define Orality Levels by **clustering linguistically motivated features**, so that levels emerge from data rather than from rater scores.

For evaluation, we adopt two ways. First, $g(\cdot)$ assigns a level to the model output and the ground-truth spoken text, and Level-Accuracy (L-Acc) measures whether the two levels match. Second, we train four Bradley–Terry reward models (Bradley and Terry, 1952) on 10k pairwise preferences sampled from WSPairs and labeled by an LLM-judge ensemble using the prompt in Appendix Listing 2; training details are given in Appendix B. We then report the **reward-score MAE** between the model output and the ground-truth spoken text. To benchmark the evaluators, Appendix A describes 200 human anchor pairs (Cohen’s $\kappa = 0.82$) sampled from 50 held-out items. Table 2 compares all proxies on these anchors: $g(\cdot)$ reaches 0.910 pair accuracy, above the LLM ensemble (0.885) and the strongest reward model (0.855). We therefore use L-Acc as a validated feature-based proxy for graded orality, report the reward model as an independent check, and use human ratings for claims about perceived naturalness.

4 Orality Attribution and Intervention

We now probe whether orality is spread across the network or localized to a rendering circuit. We first identify the responsible components (§4.1), analyze their behavior (§4.1), and design two modular interventions on the identified circuit (§4.2). We conduct all analyses on Qwen-3B and replicate key

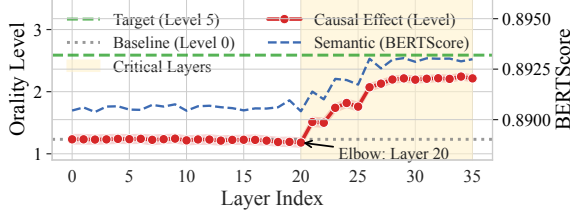


Figure 3: Layer-wise causal patching on Qwen-3B ($n=500$, 95% CI shaded). **Red, left axis:** mean predicted Orality Level after patching layer ℓ . **Blue dashed, right axis:** BERTScore F1 between patched output and the written reference. The yellow band marks $\mathcal{L}_{\text{oral}} = \{20, \dots, 35\}$. See App. Fig. 9 for Mistral-7B.

results on Llama-1B and Mistral-7B.

4.1 Tracing Orality to Model Components

Layer-level localization. We use activation patching (Meng et al., 2022; Hicks and Tingley, 2011) on contrastive prompts that ask the model to produce written (Level 0) or oral (Level 5) outputs for the same input x^w (Appendix Listing 4). For a model $\mathcal{M}$ with L layers, let h_ℓ^{oral} be the hidden state from the oral run. We patch it into the written run, $\hat{y}_\ell = \mathcal{M}(x_w \mid h_\ell \leftarrow h_\ell^{\text{oral}})$, and measure the average indirect effect (AIE) on the predicted Orality Level: $\text{AIE}_\ell = \mathbb{E}[g(\hat{y}_\ell) - g(y_{\text{base}})]$. Since y_{base} is at Level 0, AIE_ℓ reduces to $\mathbb{E}[g(\hat{y}_\ell)]$.

Orality rendering concentrates in late layers. As Figure 3 shows, the patched Orality Level stays near zero up to layer 19 on Qwen-3B (36 layers) and then rises sharply, saturating near the top. Importantly, semantic similarity BERTScore shows only a slight upward drift, suggesting that the patch changes style rather than content. We define $\mathcal{L}_{\text{oral}} = \{20, \dots, 35\}$ based on the elbow at layer 20. The same hockey-stick generalizes across backbones: the elbow shifts to layer 5 on Llama-1B and layer 15 on Mistral-7B (Appendix Fig. 9).

Head attribution. With orality localized to a band of late layers, we zoom in to the head level. Prior work shows that attention heads can implement specific linguistic behaviors (Voita et al., 2019; Olsson et al., 2022). We combine zero ablation, which replaces one head’s output with $\mathbf{0}$, with a contrastive influence score (CIS) targeted at spoken-marker vocabulary. For head a , let $\mathcal{M}_{\setminus a}$ denote the ablated model. The token-level influence is $\mathcal{I}_a(y_t) = P_{\mathcal{M}}(y_t \mid x, y_{<t}) - P_{\mathcal{M}_{\setminus a}}(y_t \mid x, y_{<t})$. To separate orality from general language modeling capacity, we compute a contrastive score between

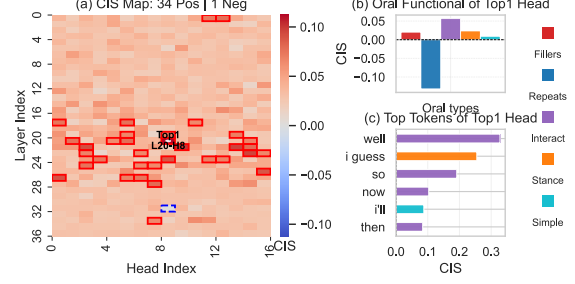


Figure 4: Head-level Contrastive Influence Score (CIS, Eq. 1) on Qwen-3B. **(a)** CIS heatmap across layers $\times$ heads; red/blue boxes mark the 34 positive and 1 negative orality heads ($|\text{CIS}_a - \mu| > 2\sigma$); the top-CIS head is L20-H8. **(b)** Per-marker-type contribution of L20-H8. **(c)** Top tokens L20-H8 promotes in the orality logit direction. See App. Fig. 10 for Mistral-7B.

a small lexicon of oral markers, $\mathcal{V}_{\text{oral}}$ (e.g., fillers and hedges), and a non-oral lexicon, $\mathcal{V}_{\text{non-oral}}$:

$$\text{CIS}_a = \mathbb{E}_{y \in \mathcal{V}_{\text{oral}}} [\mathcal{I}_a(y)] - \mathbb{E}_{y \in \mathcal{V}_{\text{non-oral}}} [\mathcal{I}_a(y)], \quad (1)$$

We then define the orality heads as $\mathcal{A}_{\text{oral}} = \{a : |\text{CIS}_a - \mu| > 2\sigma\}$ where μ and σ are the mean and standard deviation of CIS_a over all heads.

To test whether head attribution depends on the choice of oral markers, we perturb $\mathcal{V}_{\text{oral}}$ and rerun head attribution. The resulting orality-head sets remain consistent with the original one, with Jaccard overlap at least 0.828 throughout (Appendix H, Table 11). The evidence also extends beyond marker emission. Level-Accuracy depends in part on propositional idea density (PID) and noun-verb ratio (NVR; §3.2), which a lexicon-only manipulation would leave unchanged, while LAS on $\mathcal{A}_{\text{oral}}$ remains the strongest lightweight method on all three backbones (§5.2). This result is consistent with the localized heads influencing structural orality, although the lexical CIS does not by itself establish a complete structural mechanism.

Orality heads are sparse and concentrate in $\mathcal{L}_{\text{oral}}$. Of 448 heads on Qwen-3B, only 35 fall outside the two-sigma band (34 positive, 1 negative). Almost all lie in layers 20–30 within $\mathcal{L}_{\text{oral}}$ (Fig. 4a), so the layer-level and head-level localization land on the same band. The same sparse, late-layer pattern replicates on Llama-1B (25 positive, 7 negative, layers 5–15) and Mistral-7B (38 positive, 11 negative, layers 15–31), co-located with each backbone’s $\mathcal{L}_{\text{oral}}$ (Appendix Fig. 10).

The sparse set is robust to localization granularity. Because both controllers use the two-sigma rule, we vary its CIS threshold and head count (Ta-

Selection rule	Heads	L-Acc↑	BERT↑
$ CIS - \mu > 1.0\sigma$	76	0.578	0.908
$ CIS - \mu > 1.5\sigma$	49	0.591	0.909
$ CIS - \mu > 2.0\sigma$	35	0.584	0.911
Top-20 CIS heads	20	0.580	0.913
Top-10 CIS heads	10	0.535	0.912
Matched random non-oral	35	0.503	0.910

Table 3: LAS is stable to CIS thresholds and moderate head counts, supporting a sparse high-influence set.

ble 3). Across 20–76 selected heads, LAS varies by only 1.3 L-Acc points; Top-10 drops 4.9 points, and matched random performs worse than every comparable localized set. Thus, localization identifies a robust sparse high-influence set, not a unique minimal circuit.

Top heads promote discourse markers and hedges. The top-1 CIS head L20–H8 routes most of its contribution into interaction markers and hedges (Fig. 4b), promoting tokens such as *well*, *I guess*, *so*, and *you know* (Fig. 4c). This pattern is consistent with a *discourse-rendering* head rather than a generic style head.

4.2 Modular Intervention on the Orality Circuit

The localization and the behavior analysis in §4.1 suggest a simple design rule: intervene only on $\mathcal{L}_{\text{oral}} \cup \mathcal{A}_{\text{oral}}$ and leave the rest of the model untouched. We instantiate the rule as two methods at different training-cost points.

CAS: no-weight-update calibrated activation steering. CAS steers generation toward a user-specified Orality Level t at inference time without gradient-based weight updates. It does use a held-out validation set to fit a scalar calibration curve. CAS extends Activation Addition (ActAdd) (Turner et al., 2024) in two ways: instead of injecting at a single grid-searched layer, it applies steering to the full localized layer set $\mathcal{L}_{\text{oral}}$; and instead of using a fixed coefficient, it calibrates the steering strength to the target level. Concretely, for each $\ell \in \mathcal{L}_{\text{oral}}$, we compute a steering vector as the L2-normalized mean hidden-state difference between Level 5 and Level 0 inputs at the $\langle \text{assistant} \rangle$ chat-template token (Zou et al., 2025),

$$v_\ell = \mathbb{E}_{t=5}[h_\ell] - \mathbb{E}_{t=0}[h_\ell], \hat{v}_\ell = v_\ell / \|v_\ell\|_2. \quad (2)$$

and inject it during decoding as $h'_\ell = h_\ell + \alpha \hat{v}_\ell$ at every position. Because a fixed α does not reliably

reach a desired ordinal level, we calibrate α on held-out validation samples: for each target $t \in \{1, \dots, 5\}$, we sweep $\alpha \in [0, 4]$ and choose $\alpha_t^* = \arg \min_\alpha |\hat{\ell}(\alpha) - t|$, then fit a quadratic control function through $\{(0, 0)\} \cup \{(t, \alpha_t^*)\}_{t=1}^5$, $\alpha^*(t) \approx w_2 t^2 + w_1 t + w_0$. This produces a single calibrated mapping over all six levels. We use a quadratic because the empirical $\alpha \mapsto \hat{\ell}(\alpha)$ curve shows a rise-then-plateau trend.

We also evaluate isotonic regression (Fielding, 1974) in the ablation study; as shown in Table 6, the quadratic fit improves Level-Accuracy by about 10 points under our validation grid.

LAS: learned activation steering. LAS replaces CAS’s fixed direction and global scalar with a per-component element-wise affine, learned on parallel data with the base model frozen. LAS intervenes on (i) the hidden states of oral layers $\mathcal{L}_{\text{oral}}$ and (ii) the outputs of orality heads $\mathcal{A}_{\text{oral}}$. For each component c with activation $\mathbf{h}^c \in \mathbb{R}^d$, LAS learns $\mathbf{d}_c, \mathbf{b}_c \in \mathbb{R}^d$ and applies

$$\mathbf{h}^{c'} = \mathbf{h}^c + (\mathbf{d}_c \odot \mathbf{h}^c + \mathbf{b}_c). \quad (3)$$

We initialize the scale and bias in Eq. 3 to zero so training starts from the identity map, and optimize these parameters with a cross-entropy loss on WSPairs, totaling $|\phi| = 2d_{\text{model}}(|\mathcal{L}_{\text{oral}}| + |\mathcal{L}_{\text{oral}}^{\text{head}}|)$, which is only 0.004%–0.006% of the backbone parameters (Appendix J, Table 13 for training details).

5 Experiments

We structure the experiments around three questions. (i) Can CAS improve gradient-free target-level control across backbones without updating model weights? (ii) Can LAS provide a better quality–efficiency trade-off than classical PEFT baselines for graded orality control? (iii) Can the orality capability learned by LAS transfer effectively across domains?

5.1 Experiment Setup

Baselines. We compare four baselines across the prompt-to-fine-tune setting: **Few-shot** prompt with level definitions (Appendix Listing 3), full **SFT**, **LoRA** (Hu et al., 2022), and **IA**³ (Liu et al., 2022). LoRA and IA³ are scope-matched to LAS ($r=4$ on the same FFN down-projection layers) to isolate the effect of localization rather than adaptation capacity. In practice, PEFT methods such as LoRA can

Table 4: No-weight-update graded orality control on the WSPairs test set. Token/s = decoding throughput; BERT = BERTScore F1; L-Acc = level accuracy, with subscript denoting signed level bias; ΔR_r = reward-score MAE between the predicted output and the ground-truth spoken text (subscript: Pearson r).

Model	Method	Token/s $\uparrow$	BERT $\uparrow$	L-Acc $\uparrow$ _{bias}	ΔR_r $\downarrow$
<i>Reference: Closed-source models</i>					
DeepSeek-R1	Few-shot	58.7	0.899	0.532 $_{-0.47}$	0.35 $_{0.72}$
Gemini-3-Flash	Few-shot	45.1	0.899	0.593 $_{-0.30}$	0.31 $_{0.76}$
GPT-5.2	Few-shot	45.2	0.903	0.364 $_{-0.98}$	0.42 $_{0.65}$
<i>Open-Source Models</i>					
Qwen-3B	Few-shot	420.5	0.900	0.246 $_{-0.98}$	0.52 $_{0.43}$
	ActAdd	418.1	0.893	0.149 $_{-1.87}$	0.46 $_{0.51}$
	CAS _{ours}	415.8	0.894	0.390 $_{-0.26}$	0.30 $_{0.81}$
Mistral-7B	Few-shot	185.1	0.892	0.210 $_{-1.36}$	0.49 $_{0.46}$
	ActAdd	183.6	0.875	0.324 $_{-0.17}$	0.39 $_{0.62}$
	CAS _{ours}	181.5	0.852	0.378 $_{-0.49}$	0.32 $_{0.74}$
Qwen-7B	Few-shot	175.4	0.902	0.169 $_{-1.63}$	0.58 $_{0.40}$
	ActAdd	173.6	0.899	0.053 $_{-2.56}$	0.51 $_{0.49}$
	CAS _{ours}	171.2	0.878	0.245 $_{-0.63}$	0.33 $_{0.78}$

approach full-SFT performance when given unconstrained capacity, so we report full SFT as an upper bound for PEFT. Decoding-time controllers such as PPLM, GeDi, and FUDGE are omitted because they are designed for per-token classifier-guided control rather than ordinal rewriting. Appendix K and Table 14 summarize the baselines, including where the intervention is applied, which components are updated, what parameterization is used, and how the control signal is calibrated or learned. Full hyperparameters are listed in Appendix Table 13.

Metrics. **L-Acc** is the discrete-level hit rate $\frac{1}{N} \sum_i \mathbf{1}[g(\hat{y}_i) = t_i]$, with signed bias $\frac{1}{N} \sum_i (g(\hat{y}_i) - t_i)$ reported as L-Acc_{bias}; it is a validated feature-based proxy rather than a complete measure of perceived naturalness. ΔR_r is the mean absolute error (MAE) between an output’s and the ground-truth spoken text’s reward score under the independent Mistral-7B Bradley–Terry reward model (§3.2), with subscript r giving the Pearson correlation across samples. **BERTScore F1** measures content preservation. **Tokens/s** is averaged over 1,000 prompts (length 542) at batch 1, greedy decoding, max new tokens 500 on A800 80GB.

5.2 Main Results

CAS improves exact level targeting, but is not a strictly monotonic dial. Table 4 compares Few-shot, single-layer fixed- α ActAdd (Turner et al., 2024), and CAS on three open-source backbones. CAS consistently improves exact target matching over the other no-weight-update baselines, increas-

Table 5: Supervised orality control on WSPairs test. Params (subscript: % of backbone); Tokens/s: decoding throughput; BERT: BERTScore F1. L-Acc: level hit rate (subscript: signed level bias); ΔR_r = reward-score MAE between the predicted output and the ground-truth spoken text (subscript: Pearson r).

Model	Method	Params $\downarrow$	Tokens/s $\uparrow$	BERT $\uparrow$	L-Acc $\uparrow$ _{bias}	ΔR_r $\downarrow$
Llama-1B	SFT _{ref}	1.2B _{100%}	1150.2	0.904	0.514 $_{-0.46}$	0.27 $_{0.81}$
	LoRA	0.6M _{0.053%}	1145.5	0.902	0.378 $_{-0.08}$	0.46 $_{0.66}$
	IA ³	0.1M _{0.011%}	1152.1	0.899	0.363 $_{-0.52}$	0.49 $_{0.61}$
	LAS _{ours}	78K _{0.006%}	1151.8	0.901	0.413 $_{-0.50}$	0.31 $_{0.78}$
Qwen-3B	SFT _{ref}	3.0B _{100%}	416.6	0.913	0.573 $_{-0.42}$	0.24 $_{0.83}$
	LoRA	1.9M _{0.061%}	410.5	0.910	0.464 $_{-0.41}$	0.42 $_{0.71}$
	IA ³	0.4M _{0.013%}	417.2	0.907	0.451 $_{-0.39}$	0.46 $_{0.65}$
	LAS _{ours}	119K _{0.004%}	418.5	0.911	0.584 $_{-0.34}$	0.21 $_{0.86}$
Mistral-7B	SFT _{ref}	7.2B _{100%}	182.4	0.915	0.572 $_{-0.29}$	0.23 $_{0.84}$
	LoRA	2.4M _{0.033%}	178.9	0.918	0.575 $_{-0.46}$	0.28 $_{0.79}$
	IA ³	0.5M _{0.006%}	183.5	0.905	0.479 $_{-0.31}$	0.45 $_{0.66}$
	LAS _{ours}	295K _{0.004%}	175.2	0.915	0.592 $_{-0.30}$	0.19 $_{0.88}$

ing L-Acc over ActAdd by 24.1 points on Qwen-3B (0.390 vs. 0.149), 5.4 on Mistral-7B, and 19.2 on Qwen-7B. The signed bias shrinks at the same time (e.g., $-1.87 \rightarrow -0.26$ on Qwen-3B), showing that the gain comes from better ordinal control rather than simply stronger activation steering.

Direct target-level diagnostics qualify this result (Appendix Table 12). On Qwen-3B, CAS’s mean predicted level increases from 1.39 to 2.37 over targets 1–4 but falls to 2.27 at target 5; its Spearman correlation is 0.236, below Few-shot’s 0.399. Thus CAS improves calibrated target-level accuracy and bias, but does not establish strict monotonic control at the top of the scale.

The independent ΔR_r evaluator confirms the same trend: relative to ActAdd, CAS reduces ΔR by 17.9%–35.3% and increases Pearson r by 19.4%–59.2% across the three backbones, indicating closer alignment with the target spoken intensity under a metric independent of the estimator $g(\cdot)$. The trade-off is mild: BERTScore decreases by at most 0.04, and throughput drops by less than 1.5% on all backbones.

LAS delivers the best quality–efficiency trade-off with 0.004%–0.006% trainable parameters.

Table 5 shows that LAS is the strongest lightweight supervised method across all three backbones. On Qwen-3B, LAS reaches the best L-Acc in the table (0.584) with only 119K trainable parameters, outperforming LoRA and IA³ by 12.0 and 13.3 points (0.584 vs. 0.573 for SFT_{ref}). On Mistral-7B, LAS again ranks first at 0.592, beating LoRA by 1.7 points and SFT_{ref} by 2.0 points with only 295K parameters. On Llama-1B, LAS remains the best PEFT-style method, improving over LoRA and IA³

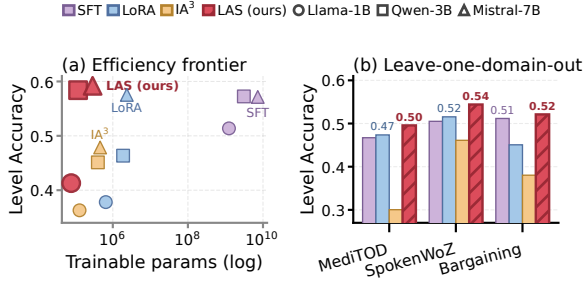


Figure 5: LAS lies on the strongest lightweight parameter–accuracy frontier and leads all three leave-one-domain-out folds. (a) WSPairs in-domain results; method labels are shown on Mistral-7B only. (b) Mistral-7B transfer, with the oral layer-and-head set trained on two domains and reused on the third.

by 3.5 and 5.0 points respectively.

Unlike CAS, LAS restores content preservation to the high-BERT regime (0.901–0.915 across backbones). The independent ΔR_r evaluator gives the same ranking: LAS achieves the lowest ΔR and highest r on every backbone, supporting the conclusion that its outputs better match the target spoken intensity under a metric independent of the estimator $g(\cdot)$.

Overall, LAS combines the highest lightweight-method L-Acc with the smallest parameter count on the three main backbones (Fig. 5a), tuning only 0.004%–0.006% of backbone parameters. A 14B/72B scale check in Appendix Table 15 shows the same compactness advantage, although LAS is below SFT on Qwen-14B and should not be read as a universal replacement for standard adapters.

The localized circuit transfers across domains.

We test cross-domain transfer with a leave-one-domain-out setting on Mistral-7B. For each 2-of-3 split over MEDITOD, SPOKENWOZ, and BARGAINING: we train the oral layer-and-head set once on the two source domains and reuse it unchanged on the held-out target domain. LAS is best on all three folds, outperforming SFT by 0.9–3.9 L-Acc points and the strongest PEFT baseline by 2.2–7.0 points (Fig. 5b).

These results suggest that the localized oral L+H circuit captures a transferable mechanism of orality rather than a domain-specific shortcut.

5.3 Ablation Studies

CAS design. Table 6 shows that CAS depends on both **multi-layer injection** and level-conditioned **calibration**, with the former contributing more. Removing calibration lowers L-Acc by 12.3 points,

Table 6: Ablation of CAS on Qwen-3B, varying *calibration function family* and *multi-layer injection*. —**Calib**: best fixed scalar on $\mathcal{L}_{\text{oral}}$. —**Multi**: $\alpha^*(t)$ on the single best layer. **Iso**: replace quadratic with isotonic regression (Fielding, 1974) on the same (t, α_t^*) points. **Rnd L**: random non-oral layer set; **Rnd v**: random Gaussian vector in $\mathcal{L}_{\text{oral}}$.

	CAS	Single-axis		Calib. family	Sanity	
	(ours)	—Calib	—Multi	Iso	Rnd L	Rnd v
$\bar{\alpha}$	2.56	3.50	2.49	2.56	1.65	3.31
L-Acc $\uparrow$	0.390	0.267	0.141	0.293	0.204	0.163
BERT $\uparrow$	0.894	0.883	0.893	0.888	0.894	0.893

removing multi-layer injection by 24.9, and removing both by 24.1 points, bringing performance back to the ActAdd level (0.149).

The quadratic calibration is also important: replacing it with isotonic regression drops L-Acc from 0.390 to 0.293 at nearly unchanged BERTScore, with the loss concentrated at higher target levels where isotonic regression saturates but the quadratic continues to extrapolate.

Random-layer and random-direction controls perform much worse than CAS, with L-Acc dropping by 18.6 and 22.7 points respectively. Figure 6 gives the complementary inference-time view. In an $\alpha \in [0, 4]$ sweep, only the learned oral direction applied to $\mathcal{L}_{\text{oral}}$ raises the mean predicted level from ≈ 1 to ≈ 4 , while BERTScore stays above 0.85 until $\alpha \approx 3$; random non-oral layers and random Gaussian directions remain near baseline. This strength response supports intervention at the localized oral circuit, but does not imply perfect monotonicity across requested target levels (Appendix Table 12).

LAS design. Table 7 shows that LAS depends on both **where** it intervenes and **how** it parameterizes the intervention. Replacing the localized oral layers and heads (L+H) set with a matched-count random non-oral set lowers L-Acc by 7.7 points, and removing either the layer or the head intervention costs a further 13–19 points, confirming that both sites contribute.

The affine form is also necessary: removing either the multiplicative scale or the additive bias reduces L-Acc by 13–15 points, so effective activation-side control requires both terms jointly. Under the same localized L+H scope, LAS also outperforms LoRA (0.584 vs. 0.568) while using about $4\times$ fewer parameters, whereas IA^3 and rank-4 ReFT under-control high target levels. Among

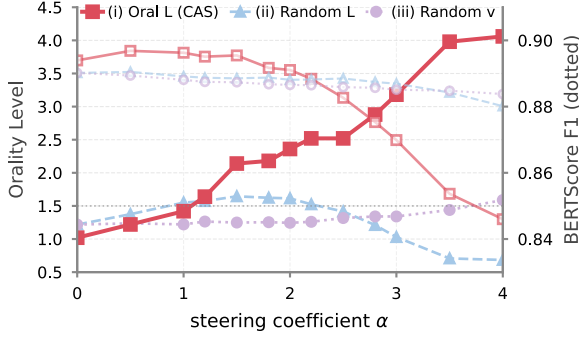


Figure 6: In an α -sweep on Qwen-3B, the learned oral direction applied to $\mathcal{L}_{\text{oral}}$ produces a generally increasing response up to saturation while preserving BERTScore until $\alpha \approx 3$. This response to steering strength does not imply perfect monotonicity across requested target levels (Appendix Table 12). Random non-oral layers and random Gaussian vectors leave the predicted level near baseline.

Table 7: Ablation of LAS on Qwen-3B (WSPairs test). **Localization:** Rnd swaps oral L+H for matched-count random non-oral components; **−Lay./−Hd.** disable layer or head intervention. **Param.:** −Sc./−Bi. drop the multiplicative scale or additive bias. **Adapter:** replace LAS’s affine with rank-4 LoRA, IA³, or rank-4 ReFT on the same L+H.

	LAS	Localization				Param.		Adapter		
	(ours)	Rnd	−Lay.	−Hd.	−Sc.	−Bi.	LoRA	IA ³	ReFT	
Params↓	119K	119K	53K	66K	59K	59K	475K	59K	475K	
L-Acc↑	0.584	0.507	0.397	0.448	0.433	0.457	0.568	0.248	0.252	
BERT↑	0.911	0.912	0.900	0.908	0.906	0.906	0.909	0.900	0.901	

targets 4–5, 25.7% of LAS outputs are predicted at Level 2 or below, compared with 44.9% for IA³ and 61.2% for ReFT. IA³’s gating-only form and ReFT’s low-rank loop may limit their intervention strength in this setting; this architectural explanation is a hypothesis rather than a causal result.

5.4 Human Evaluation

We conduct a blind human evaluation with three NLP experts on 900 outputs (6 methods $\times$ 3 domains $\times$ 5 levels $\times$ 10 items). Inter-annotator agreement is substantial, with Fleiss’ $\kappa=0.64/0.72$ on the two Likert dimensions and Kendall’s $W=0.78$ on method rankings.

LAS is the only method that remains strong on all quality axes simultaneously (Fig. 7): it achieves the highest Spoken Naturalness (3.81/5) and Overall Preference (25.7%), while preserving content at the SFT level on Semantic Preservation (4.82/5) and attaining the best NER F1 (0.851 vs. 0.835 for SFT, 0.819 for LoRA, and 0.773 for IA³; Few-shot

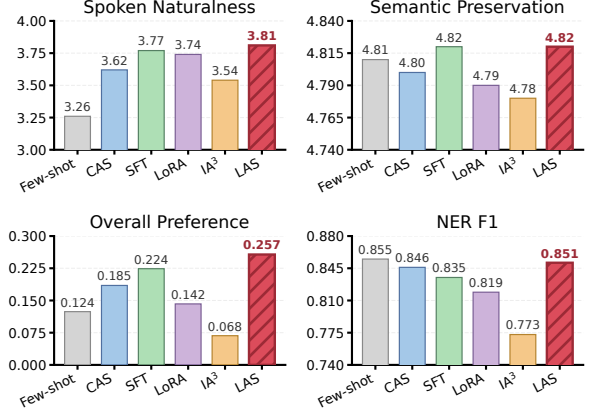


Figure 7: Blind text-level evaluation favors LAS on spoken naturalness and overall preference while tying SFT on semantic and factual preservation. Results are on Qwen-3B; higher is better for Spoken Naturalness (1–5), Semantic & Fact Preservation (1–5), Overall Preference (listwise win-rate), and NER F1 (entity preservation).

reaches 0.855).

We evaluate text directly rather than through TTS-based MOS, since MOS would introduce a confounding TTS model and may penalize natural spoken markers when the synthesizer handles them poorly. Instead, annotators read the rewritten text aloud and judge whether it sounds natural and matches the intended degree of orality. The annotation protocol is given in Appendix N.

Appendix M (Table 16) provides side-by-side outputs from different methods on the same rewriting case. The example illustrates failure modes that aggregate metrics can hide: SFT enters a degenerate loop, CAS over-expands the response, and LAS preserves the two numeric values while adding spoken markers. It is qualitative evidence, not a frequency estimate of these errors.

6 Conclusion

Across the backbones studied, spoken-style rendering concentrates in late layers and a sparse high-influence set of attention heads. Intervening on only 0.004%–0.006% of backbone parameters within this set enables efficient and transferable control: LAS obtains higher L-Acc than full SFT on Qwen-3B and Mistral-7B, remains the strongest lightweight method on Llama-1B, and leads leave-one-domain-out transfer. CAS requires no weight updates and improves exact level targeting over Few-shot and ActAdd, but is not strictly monotonic at the highest levels. This motivates testing localize-then-steer control on other surface styles.

Limitations

The Orality Level estimator $g(\cdot)$ is an English-specific, feature-based discretization of a continuous and relative attribute. Because $g(\cdot)$ assigns WSPairs levels and also computes L-Acc, L-Acc is not an independent measure of human-perceived naturalness. We interpret it as a validated proxy for graded orality and support it with an independently trained Bradley–Terry reward metric and blind human evaluation. The feature set, normalization, and choice of $k=6$ may still affect the resulting levels, and the seven features may miss pragmatic or emerging spoken-language cues.

WSPairs is a selective 4,709-pair sample from 217,000 source utterances. Its spoken side inherits ASR errors, while its written side may retain artifacts from LLM drafting despite human correction. The annotators were English-speaking NLP practitioners recruited in China, not native speakers of English; their judgments may miss dialectal and sociocultural nuances. The estimator may likewise mischaracterize non-mainstream English varieties, and neither cross-lingual transfer nor native-speaker validation is studied.

Our experiments evaluate text-level spoken-style rewriting for text-first voice-agent pipelines, not a complete voice system. We do not evaluate TTS prosody, latency, dialogue-state behavior, end-to-end user experience, or whether a higher Orality Level is preferable in a particular setting. The human evaluation used NLP experts who read outputs aloud rather than healthcare or other domain experts. Clinical or persuasion-sensitive deployment would require domain-expert review, safety testing, and user studies. CAS and LAS also require access to internal activations and custom forward hooks; prompting or standard LoRA/QLoRA may be simpler in black-box APIs or existing serving stacks, and fewer trainable parameters do not by themselves imply lower end-to-end deployment cost.

Finally, the late-layer pattern is evaluated on a limited set of transformer families and may shift under substantially different architectures. The threshold sensitivity results support a sparse high-influence set, not an exact minimal circuit. Head attribution uses zero ablation; mean or resample ablation (Wang et al., 2023) and attribution patching could reduce variance in future CIS rankings.

AI Assistance Statement

Generative AI tools, including ChatGPT and Gemini, were used to assist with language proofreading during the preparation of this work. All generated content was reviewed and verified by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Ethics Statement

This work helps LLMs deliver a more natural spoken-style response when serving humans, with the goal of friendlier voice-agent interactions. Operators in sensitive contexts should disclose that the conversational tone is algorithmic.

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000). We would like to thank the anonymous reviewers for their insightful comments.

References

- F Niyi Akinnaso. 1982. On the differences between spoken and written language. *Language and speech*, 25(2):97–125.
- F Niyi Akinnaso. 1985. On the similarities between spoken and written language. *Language and speech*, 28(4):323–359.
- Douglas Biber. 1991. *Variation across speech and writing*. Cambridge university press.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. *Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. *Plug and play language models: A simple approach to controlled text generation*. Preprint, arXiv:1912.02164.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom

- Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. [A mathematical framework for transformer circuits](https://transformer-circuits.pub/2021/framework/index.html). Transformer Circuits Thread. <https://transformer-circuits.pub/2021/framework/index.html>.
- A. Fielding. 1974. [Statistical inference under order restrictions. the theory and application of isotonic regression](#). Royal Statistical Society. Journal. Series A: General, 137(1):92–93.
- Zishan Guo, Linhao Yu, Minghui Xu, Renren Jin, and Deyi Xiong. 2023. [CS2W: A Chinese spoken-to-written style conversion dataset with multiple conversion types](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3962–3979, Singapore. Association for Computational Linguistics.
- Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2023. [Language of bargaining](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13161–13185, Toronto, Canada. Association for Computational Linguistics.
- Raymond Hicks and Dustin Tingley. 2011. Causal mediation analysis. *The Stata Journal*, 11(4):605–619.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2024. [Lorahub: Efficient cross-task generalization via dynamic lora composition](#). Preprint, arXiv:2307.13269.
- Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. [GeDi: Generative discriminator guided sequence generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhigen Li, Jianxiang Peng, Yanmeng Wang, Yong Cao, Tianhao Shen, Minghui Zhang, Linxi Su, Shang Wu, Yihang Wu, YuQian Wang, Ye Wang, Wei Hu, Jianfeng Li, Shaojun Wang, Jing Xiao, and Deyi Xiong. 2025. [ChatSOP: An SOP-guided MCTS planning framework for controllable LLM dialogue agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17637–17659, Vienna, Austria. Association for Computational Linguistics.
- Zhigen Li, Yanmeng Wang, Rizhao Fan, Ye Wang, Jianfeng Li, and Shaojun Wang. 2022. [Learning to adapt to low-resource paraphrase generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1014–1022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. [Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.
- Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. [ParaDetox: Detoxification with parallel data](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.
- Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. [StylePTB: A compositional benchmark for fine-grained controllable text style transfer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2116–2138, Online. Association for Computational Linguistics.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in gpt](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. [In-context learning and induction heads](#). Preprint, arXiv:2209.11895.
- Sudha Rao and Joel Tetreault. 2018. [Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.
- Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam. 2024. [MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations](#). In *Proceedings of the 2024 Conference on Empirical Methods in*

- Natural Language Processing, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.
- Dan Shi, Zhuowen Han, Simon Ostermann, Renren Jin, Josef van Genabith, and Deyi Xiong. 2026a. [Why does reinforcement learning generalize? a feature-level mechanistic study of post-training in large language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 38979–39000, San Diego, California, United States. Association for Computational Linguistics.
- Dan Shi, Renren Jin, Zhuowen Han, Yuqi Ren, Xinwei Wu, Zhigen Li, and Deyi Xiong. 2026b. [Neuronal insights into LLM attacks: Targeted neuron tuning for precise and robust vulnerability patching](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 34414–34435, San Diego, California, United States. Association for Computational Linguistics.
- Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. [Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 39088–39118. Curran Associates, Inc.
- Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. [Does knowledge distillation really work?](#) In *Advances in Neural Information Processing Systems*, volume 34, pages 6906–6919. Curran Associates, Inc.
- Stanley Smith Stevens. 1946. On the theory of scales of measurement. *Science*, 103(2684):677–680.
- Daniel Tamayo, Aitor Gonzalez-Agirre, Javier Hernandez, and Marta Villegas. 2024. [Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5831–5847, Bangkok, Thailand. Association for Computational Linguistics.
- Louis L Thurstone. 2017. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge.
- Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). Preprint, arXiv:2308.10248.
- Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018. [Varying image description tasks: spoken versus written descriptions](#). In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 88–100, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. [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, Florence, Italy. Association for Computational Linguistics.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Yiqun Wang, Chaoqun Wan, Sile Hu, Yonggang Zhang, Xiang Tian, Yaowu Chen, Xu Shen, and Jieping Ye. 2025. [Tracing and dissecting how LLMs recall factual knowledge for real world questions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23246–23271, Vienna, Austria. Association for Computational Linguistics.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. [Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.
- Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. [DEPN: Detecting and editing privacy neurons in pretrained language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2875–2886, Singapore. Association for Computational Linguistics.
- Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2024. [Left: Representation fine-tuning for language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 63908–63962. Curran Associates, Inc.
- Kevin Yang and Dan Klein. 2021. [FUDGE: Controlled text generation with future discriminators](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pages 3511–3535, Online. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. [SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. [Representation engineering: A top-down approach to ai transparency](#). Preprint, arXiv:2310.01405.

A Annotation Protocol and Human Validation

LLM-assisted written rewrites. For each spoken transcript, we ask an LLM to draft a written rewrite that removes spoken markers and repairs disfluencies while preserving entities and facts. Annotators then correct the draft to ensure semantic equivalence, standardize the written style, and remove residual spoken markers.

Annotator selection. We recruited candidates from a Chinese university and industry through a three-stage selection process; five annotators completed the final dataset review. The process included three steps:

- **Qualification.** Candidates must hold at least a bachelor’s degree and have background knowledge in NLP; preference is given to candidates with annotation experience.
- **Trial annotation.** Candidates who pass the qualification stage receive standardized training. They then annotate a small trial set, and we assess their outputs against the annotation rules.
- **Diversity.** From those who pass the trial annotation, we select individuals from different university majors and professional departments to increase diversity.

Measures to ensure quality and consistency. We use a shared training manual, clarification meetings, and worked examples. Pilot annotations identify issues before full annotation begins. Throughout annotation, we discuss feedback, update the

manual when needed, and re-examine affected annotations. For consistency checking, a random subset of 1,247 pairs received a second independent annotation; Cohen’s κ was 0.92 on exact-match judgments.

III. Compensation. On average, the annotation time was 3.8 minutes per sample, and the compensation was \$8 per hour.

Guidelines summary. Annotators are instructed to: (1) preserve all named entities, numbers, constraints, and user intent; (2) avoid adding new information; (3) remove fillers, turn-taking cues, repetitions, and repairs; (4) produce grammatical, concise written sentences.

See Listing 1 for the detailed annotation manual.

Listing 1: Spoken-to-Written Annotation Guidelines

Spoken-to-Written English Conversion: Annotation Guidelines

This guide details the standards for converting informal spoken English transcripts into formal written text and for evaluating the quality of AI-generated drafts.

I. Task Description

Inputs:

1. Source Text: A transcript of spoken English often from dialogues in professional settings (e.g. healthcare sales customer service).
2. Model Output: An AI-generated draft that attempts to convert the source text into formal written English.

Your Tasks:

1. Revise written English: Using the four quality dimensions below as a guide review the model’s output and make minimal localized corrections. If no revision is needed please leave it blank.
2. Model Output: Select all error categories that are present in the model’s rewritten text. If you have corrected the text you must select at least one category to justify the edits.

II. Annotation Principles

These four quality dimensions are the basis for both your editing and your ratings.

- 1. Unintelligible or Unsalvageable Source: The source spoken text is incomprehensible due to severe ASR errors or incompleteness making a faithful rewrite impossible. This flags the sample for deletion.
- 2. Semantic Accuracy & Completeness: Assesses whether the text accurately and completely reflects the meaning of the original spoken source. This dimension focuses on whether facts logic and intent are faithfully preserved.
- 3. Language Correctness & Fluency: Assesses if the text is grammatically correct natural-sounding and easy to read. This dimension is evaluated independently of the source text focusing solely on the quality of the English prose.
- 4. Style & Formality: Assesses if the text meets the standard of formal written English. This includes removing spoken features (e.g. "um" "uh" "you know") and adopting a professional objective tone suitable for the context (e.g. healthcare business).
- 5. Spelling & Punctuation: Assesses the text for spelling errors typos or incorrect punctuation usage. This dimension covers all surface-level textual errors.
- 6. Instruction Following: Assesses if the model output includes any extraneous content like explanations headers or code formatting. If this occurs please copy the clean rewritten text into your annotation.

III. Examples

Example 1

– Source Text: **Spoken English**
all right. statue on the edition on the coffee bars at risk is. walker to your setting.

– Model Output: ****Written English****
Statue on the edition on the coffee bars at risk is walker to your setting.

(Note: The original spoken text itself is an incorrect speech recognition result. The sentence "statue on the edition on the coffee bars at risk is walker to your setting" makes no logical sense. Therefore the written text is also meaningless. It needs to be marked as Category 1 for subsequent deletion of this invalid sample.)

– Your output:
Final Rewritten Text: ""
Identify Error Types: "1"

Example 2

– Source Text: ****Spoken English****
and unfortunately. i'm not finding any that serve north american foods.
– Model Output: ****Written English****
No establishments serving North American cuisine are available.

(Note: There is nothing in the model output that must be revised; following the principle of minimal modification no revision is necessary.)

– Your output:
Final Rewritten Text: ""
Identify Error Types: ""

Example 3

– Source Text: ****Spoken English****
oh. okay so your license plate is capitalized as j s g. 0. e v m.
– Model Output: ****Written English****
Written Language: \n\nThe license plate number is J S G 0 E V M.

(Note: The model's written text output is completely correct but the leading "Written Language: \n\n" is redundant and didn't follow the instructions.)

– Your output:
Final Rewritten Text: "The license plate number is J S G 0 E V M."
Identify Error Types: "6"

Example 3

– Source Text: ****Spoken English****
they unfortunately it opens on sunday from 2:00 p.m. to 5:00 p.m.. so i i guess you can put it at any time.
– Model Output: ****Written English****
****Spoken Language:**** \nthey unfortunately it opens on sunday from 2:00 p.m. to 5:00 p.m.. so i i guess you can put it at any time.\n\n****Rewritten Formal Text:**** \n\nThe facility is currently open on Sundays from 2:00 p.m. to 5:00 p.m. Therefore an appointment can be scheduled during this time or at a time that suits your convenience.

(Note: The model's written text output is completely correct but the leading "Written Language: \n\n" is redundant and didn't follow the instructions. "an appointment can be scheduled during this time" but not "at a time that suits your convenience")

– Your output:
Final Rewritten Text: "The facility is open on Sundays from 2:00 p.m. to 5:00 p.m. Therefore an appointment can be scheduled for any time within this period."
Identify Error Types: "2&6"

Human validation of levels. We provide annotators a short definition of “more spoken” vs. “more written” (interaction cues, hedging, fragmentation, etc.).

On 200 sampled items with pairwise comparisons, annotators’ selections agree with our level ordering 91% of the time.

See Listing 2 for the detailed annotation manual.

Listing 2: Spoken-vs-Written Compare Annotation Guidelines

Spoken-vs-Written Compare Annotation Guidelines

Task

Given two English variants of the same content, choose which one is more spoken. ``More spoken" means more spontaneous, conversational delivery while preserving all facts and intended meaning.

Core signals of ``more spoken"

- Structure and density: Shorter clauses, sentence splitting, looser syntax, lower lexical density. Written style favors complex subordination and dense noun phrases.
- Contractions and simpler wording: I'm we're it's; simpler verbs and phrasing over formal alternatives.
- Fillers and discourse markers: um, uh, hmm; well, so, okay, look, anyway, you know. These organize turns and soften entries.
- Stance and hedging: I think, I guess, maybe, kind of, sort of, probably. Shows speaker stance/uncertainty.
- Interaction cues: you know?, right?, yeah?, direct address (you), brief acknowledgments (yeah, mm-hm). Signals audience engagement or turn-yielding.
- Repetition and repairs: small natural repeats, restarts, self corrections (The price is it's about), echoing words for emphasis. Should feel purposeful, not random noise.
- Fragmentation and rhythm: Sentence fragments, trailing off, incremental delivery (Okay, so the plan). Still coherent at the discourse level.
- Prosody-like punctuation: Ellipses and dashes can mark spoken rhythm when used sparingly.

Reference scale (for intuition, not labeling)

- Level 0 (written): Formal structure, high lexical density, no fillers, no repetition.
- Level 1 (spoken, slight): Light contractions, mild simplification, rare markers.
- Level 2 (spoken, low): Occasional fillers (um/uh/okay), well/so, slight sentence splitting.
- Level 3 (spoken, medium): Clear conversational flow, some hedges (I think/maybe), more fragmentation, small natural repetition.
- Level 4 (spoken, high): Frequent interaction cues (you know/right?), more hedging, noticeable repetition, shorter clauses, still coherent.
- Level 5 (spoken, high): Strongly spontaneous; frequent fillers, discourse markers, turn-yielding cues; more repetition and repairs; facts preserved.

Annotation note

Do not penalize minor punctuation or casing differences unless they affect meaning or readability. Focus on conversational cues and overall spoken feel while ensuring all information remains intact.

Listing 3: Written to Spoken Prompt

Instruction

You are an expert conversationalist. Your task is to convert formal written language into spontaneous spoken language at a specified ****Orality Level (0–5)**** with ****Semantic Preservation****. Please directly output the rewritten spoken text without any explanations or formatting symbols.

Orality Level Definitions

You must adjust the density of orality words and phrases, repetition, and sentence fragmentation based on the target level:

- Level 0 (written): Formal structure. High lexical density. No fillers. No extra interaction. No repetition.
- Level 1 (spoken, slight): Light contraction and mild simplification. Rare discourse markers. No repetition.
- Level 2 (spoken, low): Add occasional fillers (um/uh/okay) and light turn-organization (well/so). Slight sentence splitting.
- Level 3 (spoken, medium): Clear conversational flow. Some stance/hedges (I think/maybe). More fragmentation. Small, natural repetitions allowed.
- Level 4 (spoken, high): Frequent interaction cues (you know/right?). More hedging. Noticeable repetition and shorter clauses, but still coherent.
- Level 5 (spoken, high): Strongly spontaneous. Frequent fillers + discourse markers + turn-yielding cues. More repetition and repairs. Still preserve every fact.

```

### Semantic Preservation

– Preserve Core Information: Do not add, invent, change, or omit
substantive content (e.g., entities, numbers, conditions, stances).
– Maintain Logical Integrity: Ensure the core meaning and intent are
clear. While sentence structure should be simplified, the original logical
order and core meaning must be maintained.

### Example
**Written Language Level 0:**
The asking price is $25,000, which is reasonable considering the four
bedrooms.\

**Spoken Language Level 1:**
The asking price is $25,000. I think it is reasonable considering the four
bedrooms.

**Spoken Language Level 2:**
Well, the asking price is $25,000. I think that's reasonable given the four
bedrooms.

**Spoken Language Level 3:**
So, the asking price is roughly $25,000. Um, I think that's pretty fair,
you know, given it has four bedrooms.

**Spoken Language Level 4:**
um, you know, the asking price is roughly $25,000. I mean, it's probably
fair. you know? you got four bedrooms.

**Spoken Language Level 5:**
Okay, um, look. The price is... it's like $25,000. Yeah. And, uh, I mean it's
fair it's fair because, you know, you got four bedrooms.

### Let's begin
Please directly output the one spoken text at a specified Orality Level.

**Written Language Level 0:**
Properties in the area have appreciated in value over time. This home has
appreciated substantially since its purchase and will continue to
appreciate, making it a strong investment.

**Spoken Language Level 4:**

```

B Evaluator Validation: Benchmarking $g(\cdot)$ Against LLM Judges and Reward Models

To preempt circularity concerns about using the same EWM+KMeans estimator $g(\cdot)$ for both label assignment and downstream evaluation, we benchmark $g(\cdot)$ against two families of alternative orality evaluators on an independent set of **200 human pairwise judgments** sampled from 50 held-out WSPairs items (§3.2, Appendix A).

Why pairwise, not absolute? Orality is psychometrically ordinal and relative (Stevens, 1946; Thurstone, 2017; Bradley and Terry, 1952): human raters reliably compare two texts but assign noisy absolute levels. The 200 anchor judgments therefore use the *paired-comparison* protocol (Listing 2); all evaluators below are scored by their agreement on these pairs (“pair accuracy”), making the comparison invariant to any specific level definition.

Family 1: LLM zero-shot judges. We use three API-based LLM judges: DEEPSEEK-V4-PRO,

GEMINI-3.5-FLASH, and GPT-5.4-NANO. For each pair, the LLM receives the human annotation protocol of Listing 2 *verbatim* with the two input texts substituted into placeholders; no additional instructions, exemplars, or chain-of-thought hints are added. We report each LLM individually and a 3-way majority vote (denoted “LLM ens.”). These evaluator models are separate from the inference baselines in Table 4.

Family 2: learned reward models. Following the Bradley-Terry preference-learning paradigm, we train scalar reward models $r_\theta(x)$ such that more spoken texts receive higher scores. Training data are **10k** pairwise preferences sampled randomly from WSPairs cross-product (with the LLM-ensemble label as the supervisory signal). We hold the architecture fixed (base LM + scalar head, LoRA $r = 8$ on attention projections) and sweep four backbones: ROBERTA-BASE (125M), QWEN-1.5B, QWEN-3B, and LLAMA-7B. Loss is the standard preference NLL $-\log \sigma(r_\theta(x_{\text{more}}) - r_\theta(x_{\text{less}}))$; each model is trained for 3 epochs at lr 1×10^{-4} with batch 32 and early-stopping on a held-out 10% slice.

Results. Table 2 reports pair accuracy on the 200 human-anchor pairs. The best reward model trails the LLM ensemble by 3.0 points (0.855 vs. 0.885), a result consistent with information loss that can arise when distilling a teacher signal (Stanton et al., 2021), although this comparison does not establish a general ceiling. The EWM+KMeans estimator $g(\cdot)$ is 2.5 points above the LLM ensemble and 5.5 points above the best reward model. This evidence reduces, but does not eliminate, concern about evaluating with the same feature-defined scale used for labels. We therefore retain L-Acc as a validated proxy and interpret it alongside the independent reward and human evaluations.

Interpretation. Human agreement is measured independently of the feature construction, so the benchmark provides external validation for $g(\cdot)$. It does not make L-Acc fully independent of the training labels, nor show that feature-based evaluation dominates all possible learned evaluators. The reward metric and final-output human study therefore remain necessary complementary evidence.

C Dataset Example

Examples from the WSPairs dataset are shown in Table 8.

Level	Written	Spoken
1	The restaurant Little Seoul serves Korean food.	i found a restaurant called little seoul that matches your description. they serve korean food.
1	A start time of 2:45 is proposed.	I was hoping that we could go for like 2:45.
2	Is it necessary to receive orders via email?	uh so do you need to receive orders via email.
2	The cost is £17.60.	okay of course it will cost you seventeen point 60 pounds.
3	What is the nature of the nasal discharge?	Okay and the runny nose what what's the discharge like?
3	Reducing the price to that extent is not feasible.	I just don't see how we can go down that low Mm That's gonna be tricky for us.
4	Selling a house is a common activity.	Pretty good Pretty good Just you know selling a house as one does.
4	Have you experienced any leg pain or swelling?	I I see and have you had any I guess any pain in your legs. Or swelling in your leg? One or the other?
5	How frequently is it used?	Okay. Um how how often would you say you use it?
5	Did he seek medical attention?	Um and um did he did he get um did he see a doctor?

Table 8: Representative WSPAIRS pairs across orality Levels 1–5 (Level 0 is by construction the written side). Spoken sides preserve fillers, hedges, repetitions, and discourse markers from the original ASR transcript; written sides remove these while preserving every entity, number, and intent. Two examples per level.

Table 9: Among the compared style/register-transfer corpora, WSPAIRS combines (a) the deployment-aligned written-to-spoken direction, (b) ordinal multi-level labels rather than a binary tag, and (c) per-sample sociolinguistic feature scores; see Table 10 for one annotation example. **Dir.:** transfer direction. **Labels:** per-sample annotation type (binary style class, transfer-type tag, or ordinal level + structured features). **Spoken src.:** authenticity of the spoken side (ASR = real production ASR; Synth. = LLM- or heuristic-synthesized; – = not a spoken-style dataset). **Pairs:** order-of-magnitude paired sample count.

Dataset	Year	Lang.	Dir.	Labels	Domains	Pairs	Spoken src.
GYAFC (Rao and Tetreault, 2018)	2018	En	informal ↔ formal	binary class	2 (Entertainment, Family)	110K	–
XFORMAL (Briakou et al., 2021)	2021	En/Fr/It/Pt	informal ↔ formal	binary class	1	~70K	–
StylePTB (Lyu et al., 2021)	2021	En	multi-style (21 axes)	21 transfer tags	1 (Penn Treebank)	8K	Synth.
ParaDetox (Logacheva et al., 2022)	2022	En	toxic ↔ neutral	binary class	1	19K	–
CS2W (Guo et al., 2023)	2023	Zh	spoken → written	binary class	open-domain	~8K	ASR
WSPairs (ours)	2026	En	written → spoken	6-level ordinal + 7 ling. feats.	3 (TOD, Medical, Negotiation)	4.7K	ASR

D WSPairs vs. Prior Style/Register-Transfer Datasets

Table 9 positions WSPairs along the dimensions that matter for downstream voice-agent training: transfer direction, label granularity, spoken-side authenticity, and domain coverage. Compared to prior style/formality corpora, WSPairs is distinguished by three orthogonal choices.

(1) Direction: written → spoken. Existing parallel corpora on the written–spoken axis either operate in the reverse direction (CS2W on spoken-to-written conversion for downstream NLU) or on the orthogonal formal/informal axis on monolingual web text (GYAFC, XFORMAL, ParaDetox). The text-first voice-agent pipeline that motivates our work requires the opposite direction (a written-style LLM output is the input; spoken-style rewriting is the target), which is not represented among the public datasets compared here.

(2) Granularity: 6-level ordinal, not binary. Most datasets compared in Table 9 use binary labels (formal/informal, toxic/neutral, spoken/written) or

transfer tags. WSPairs annotates every spoken-side sample with a discrete Orality Level 0–5 derived from a sociolinguistically grounded 7-feature score (§3.2). These ordinal labels directly support the level-conditioned control studied throughout this paper.

(3) Spoken side: real ASR transcripts, not synthesized speech-like text. WSPairs sources the spoken side directly from production ASR transcripts of three task-oriented dialog corpora (SpokenWoZ, MediTOD, Bargaining), preserving authentic disfluencies, false starts, and interaction markers. Some existing benchmarks (e.g., StylePTB style transfer) instead simulate informal style by lexical perturbation, which differs from the naturally occurring structural-syntactic signatures summarized by our PID/NVR features (§3.2).

WSPairs is intentionally not the largest of these corpora: by sourcing the spoken side from real ASR transcripts and re-annotating every pair with a 7-feature ordinal score, we trade scale for deployment-aligned direction and fine-grained annotation. The size is sufficient for parameter-

efficient training: even full SFT on the largest backbone we study (Mistral-7B) converges in 10 epochs on the 3,183-sample training split (Table 1).

Per-sample annotation: example. Beyond the level label that Table 8 displays, every WSPairs sample carries a vector of seven sociolinguistic feature scores (§3.2) from which the level is computed via entropy weighting (Appendix E). Table 10 shows one real written–spoken pair from the corpus (MediTOD subset). Features f_1 – f_4 are vocabulary-matched cues (per-category color); f_5 is repeated-token fraction; f_6 – f_7 are written-side structural statistics.

E Entropy Weight Method (EWM)

To determine feature importance without manual annotation, we use the Entropy Weight Method. EWM assigns higher weights to features with greater variability (lower entropy) across the corpus.

First, we calculate the proportion of the j -th feature for sample i :

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^N x'_{ij}}$$

Next, we compute the information entropy E_j for each feature:

$$E_j = -k \sum_{i=1}^N p_{ij} \ln(p_{ij}), \quad \text{where } k = \frac{1}{\ln(N)}$$

For numerical stability, we add a small ϵ within the logarithm. The information divergence $d_j = 1 - E_j$ reflects the feature’s discriminatory power. Finally, the normalized weight w_j is:

$$w_j = \frac{d_j}{\sum_{j=1}^M d_j}$$

where $M = 7$ is the total number of features.

F Cluster-Count Sensitivity ($k=5,6,7$)

We select $k = 6$ for KMeans by maximizing the average silhouette score over $k \in \{4, 5, 6, 7, 8\}$ on the WSPairs training partition; $k = 6$ wins by a comfortable margin and also matches the ordinal granularity desired by deployment; $k = 5$ slightly underseparates the highest-orality cluster while $k = 7$ over-splits the middle, neither of which improves downstream control accuracy.

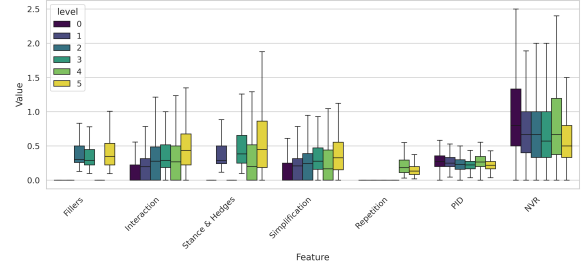


Figure 8: Feature distributions vary across estimator-assigned Orality Levels. Spoken-oriented features generally increase with level, whereas PID and NVR tend to decrease; the overlapping and occasionally non-monotonic distributions show that no single feature determines the level. This plot describes the feature-based estimator’s partition and is not an independent validation of perceived naturalness.

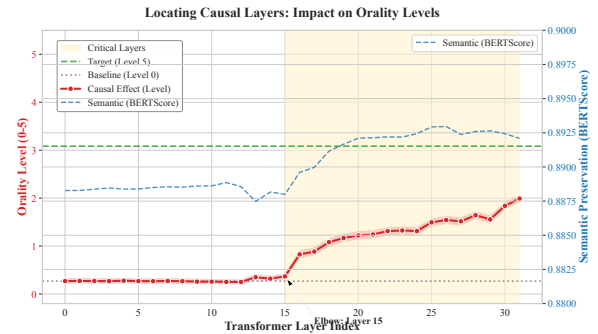


Figure 9: Mistral-7B shows the same hockey-stick pattern as Qwen with the elbow shifted to layer 15; the corresponding set is $\mathcal{L}_{\text{oral}} = \{15, \dots, 31\}$. This replicates late-layer orality rendering on a second backbone. Red curve: mean predicted Orality Level after patching each layer of Mistral-7B from an oral run into a written run; shaded band: 95% CI. Companion to Fig. 3.

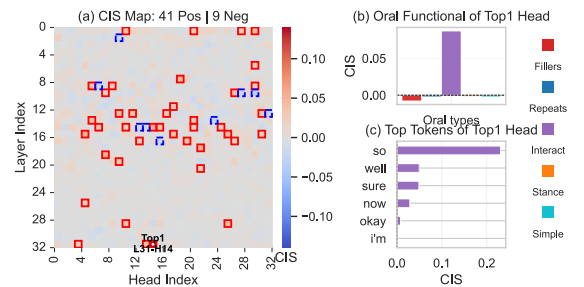


Figure 10: Mistral-7B also concentrates orality in a sparse set of late-layer heads (38 positive, 11 negative, layers 15–31), mirroring the Qwen pattern in Fig. 4. (a) CIS heatmap across heads and layers; red/blue boxes mark positive/negative *orality heads* ($|CIS_a - \mu| > 2\sigma$). (b) Per-marker-type breakdown of the top-CIS head across oral types. (c) Top tokens promoted by the top-CIS head.

Table 10: Every WSPAIRS sample carries seven sociolinguistic feature scores, from which the ordinal Orality Level is derived; in-text triggers are color-coded by feature category. Real MediTOD example.

	f_1 Fillers	f_2 Interaction	f_3 Stance/Hedges	f_4 Lex. reduction	f_5 Repetition	f_6 PID	f_7 NVR
Written:	<i>Do you live alone or with others?</i>						
Spoken:	OK, ok, I see, uhm, alright, and do you live alone or do you live with other people?						
Formula	$(c/\sqrt{N})(1+0.1u)$	$(c/\sqrt{N})(1+0.1u)$	$(c/\sqrt{N})(1+0.1u)$	$(c/\sqrt{N})(1+0.1u)$	r/N	DEPID	$N_n/(N_v+1)$
Triggers	OK, ok, uhm	I see, alright	vague noun	people	light verb	do × 2	trigram
Count c	3	2	1	2	$r=6$	N/A	N/A
Unique u	2	2	1	1	N/A	N/A	N/A
Score	0.71	0.71	0.47	0.47	0.44	0.28	0.33

Orality Level (0–5): 4 (c =in-text matches; u =unique vocab types; r =tokens inside a repeated n -gram, $n \in \{1, 2, 3\}$; N =valid non-punct tokens.)

G Contrastive Prompt for Activation Patching

For layer-level activation patching (§4.1), we run the same written input through two prompts that differ only in the requested target level: Level 0 for the written baseline and Level 5 for the oral baseline. Listing 4 shows the exact template.

Listing 4: Contrastive Prompt for Activation Patching

You are an expert conversationalist. Your task is to convert formal written language into spontaneous spoken language at a specified **Orality Level (0–5)** with **Semantic Preservation**. Please directly output the rewritten spoken text without any explanations or formatting symbols.

Orality Level Definitions

You must adjust the density of orality words and phrases, repetition, and sentence fragmentation based on the target level:

- Level 0 (written): Formal structure. High lexical density. No fillers. No extra interaction. No repetition.
- Level 1 (spoken, slight): Light contraction and mild simplification. Rare discourse markers. No repetition.
- Level 2 (spoken, low): Add occasional fillers (um/uh/okay) and light turn–organization (well/so). Slight sentence splitting.
- Level 3 (spoken, medium): Clear conversational flow. Some stance/hedges (I think/maybe). More fragmentation. Small, natural repetitions allowed.
- Level 4 (spoken, high): Frequent interaction cues (you know/right?). More hedging. Noticeable repetition and shorter clauses, but still coherent.
- Level 5 (spoken, very high): Strongly spontaneous. Frequent fillers + discourse markers + turn–yielding cues. More repetition and repairs. Still preserve every fact.

Semantic Preservation

- Preserve Core Information: Do not add, invent, change, or omit substantive content.
- Maintain Logical Integrity: Ensure the core meaning and intent are clear.

Let's begin

Please directly output the one spoken text at a specified Orality Level.

Written Language Level 0:
 $\backslash$ {written_text\}

Spoken Language Level $\backslash$ {target_level_in_prompt\}:

Lexicon variant	$ \mathcal{V}_{\text{oral}} $	Head Jaccard
Original (paper)	84	1.000
Halved	42	0.828
Doubled	168	0.914
Filler-synonym swap	84	0.942

Table 11: Lexicon perturbations retain 82.8%–94.2% Jaccard overlap with the original Qwen-3B head set. The lowest overlap occurs when the marker lexicon is halved.

H Lexicon Sensitivity for Head Attribution

The Contrastive Influence Score (CIS, Eq. 1) depends on the marker lexicon $\mathcal{V}_{\text{oral}}$. To test sensitivity to this choice, we re-run head attribution on Qwen-3B under three lexicon perturbations: (a) halve $|\mathcal{V}_{\text{oral}}|$ by random sampling, (b) double $|\mathcal{V}_{\text{oral}}|$ by adding spoken markers from independent dialogue corpora, and (c) replace fillers (um, uh) with synonym variants (er, erm). Table 11 shows Jaccard overlap from 0.828 to 0.942 with the original 35-head set. The attribution is therefore reasonably stable but not invariant to lexicon composition.

I CAS Steering Ablation: Direction and Layer Set

We complement the LAS-side component decomposition (Table 7) with a CAS-side ablation that varies the two factors of the additive intervention: the *steering direction* and the *layer set* on which it is applied.

Setup. We sweep $\alpha \in [0, 4]$ on the calibration set and report the mean predicted Orality Level and the mean BERTScore for each α . Three conditions: (i) learned v_ℓ in $\mathcal{L}_{\text{oral}}$ (full CAS); (ii) learned v_ℓ in a matched-count random non-oral layer set; (iii)

Method	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	ρ	Viol.	L-Acc
Few-shot	1.39	1.28	1.40	2.63	2.80	0.399	25.0%	0.246
ActAdd	0.81	0.89	1.00	0.90	1.35	0.144	25.0%	0.149
CAS	1.39	1.89	2.16	2.37	2.27	0.236	25.0%	0.390

Table 12: Mean predicted Orality Level by requested target on Qwen-3B.

random Gaussian vectors in $\mathcal{L}_{\text{oral}}$.

Finding. Only (i) produces a steady rise from Level ≈ 1 at $\alpha=0$ to Level ≈ 4 at $\alpha=4$ with BERT staying ≥ 0.85 until $\alpha \approx 3$. Condition (ii) saturates near Level 1.5 and then *drops* as α grows (steering at the wrong site produces incoherent text whose Level estimate falls). Condition (iii) is essentially flat at Level 1.2–1.5 (a random direction in even the right layer set carries no orality signal). Both ablations are individually necessary, confirming the component-decomposition result of §5.3 from the inference-time dynamics view.

Target-level diagnostics. Table 12 directly compares requested and mean predicted levels. CAS improves exact target matching and reduces signed bias (Table 4), but its target 5 mean is below its target 4 mean and its rank correlation is below Few-shot. Adjacent violation is the fraction of the four successive target pairs whose mean prediction decreases.

J Training Details

LAS parameter accounting. The trainable parameter count of LAS decomposes additively over the two intervention sites:

$$|\phi| = 2 d_{\text{model}} \cdot |\mathcal{L}_{\text{oral}}| + 2 d_{\text{model}} \cdot |\mathcal{L}_{\text{oral}}^{\text{head}}|, \quad (4)$$

In Eq. 4, $\mathcal{L}_{\text{oral}}$ is the set of oral layers receiving residual interventions and $\mathcal{L}_{\text{oral}}^{\text{head}} = \{\ell: \exists h, (\ell, h) \in \mathcal{A}_{\text{oral}}\}$ is the set of layers containing at least one oral head (each contributing one o_{proj} intervention). Per-backbone breakdown: Qwen-3B 119K (16 residual + 13 head layers, $d=2048$, 0.0039%), Llama-1B 78K (11+8, 0.0063%), Mistral-7B 295K (17+19, $d=4096$, 0.0041%). The implementation is identical for standard multi-head and grouped-query attention since both write into o_{proj} along contiguous head_dim blocks.

Training configuration. Table 13 reports the training configuration used for every supervised method on every backbone. All methods are trained for 10 epochs with effective batch size 64,

bf16, and cosine learning-rate scheduling with 10% warm-up; evaluations use greedy decoding with KV cache enabled, `max_new_tokens = 500` on an NVIDIA A800 80GB cluster (4 GPUs per run via `accelerate launch -multi_gpu`). Reported throughput averages over 1,000 prompts at batch size 1, prompt length 542 tokens.

Reproducibility notes. LAS training takes approximately 0.4–1.5 GPU-hours per backbone (full WSPairs). LAS checkpoints store only the affine ($\mathbf{d}_c, \mathbf{b}_c$) parameters and mask buffers, totaling < 1 MB per backbone, making multi-style serving from a single base checkpoint cheap. Code and trained adapter weights will be released alongside WSPairs.

Head intervention under GQA. For backbones with grouped-query attention (Qwen-3B uses 16 query heads / 2 KV heads; Mistral-7B uses 32/8; Llama-1B uses 32/8), per-head intervention happens at the `o_proj` input rather than the Q/K/V projections, since GQA only shares the K/V projections across heads while `o_proj` still consumes a per-head d_{head} block of activations. Concretely, before each `o_proj` call we reshape its input to $[B, T, n_{\text{head}}, d_{\text{head}}]$, multiply each oral head index $h \in \mathcal{A}_{\text{oral}}^{(\ell)}$ by the learned $(1 + \mathbf{d}_{(\ell, h)})$ and add $\mathbf{b}_{(\ell, h)}$ along the d_{head} axis, then flatten back. Non-oral head slots receive identity, so GQA’s shared-K/V structure is preserved and we do not need to materialize per-head Q activations.

CAS direction extraction: token positions. For each layer $\ell \in \mathcal{L}_{\text{oral}}$, the oral direction v_{ℓ} (Eq. 2) is computed from hidden states read at the single chat-template token immediately preceding the model’s first generated token, specifically the assistant role-tag token `<|im_start|>assistant` for Qwen, `[/INST]` for Mistral, and the corresponding assistant header for Llama. We average over $N=512$ Level-0 inputs and $N=512$ Level-5 inputs to form $\mathbb{E}[h_{\ell} \mid t]$. At inference, the layer-level direction $\hat{v}_{\ell}$ is broadcast to every decoding position; we verified that restricting injection to assistant-only positions versus all positions changes Level-Acc by ≤ 0.5 points on Qwen-3B but produces noticeably more degenerate outputs on Mistral, so we report the broadcast variant throughout.

LAS application positions. Unlike CAS, LAS’s affine ($\mathbf{d}_c, \mathbf{b}_c$) is applied to every token at training and inference (no position gating), since the

Table 13: Training details for supervised methods. “Effective BS” = per-device batch size $\times$ gradient accumulation $\times$ number of GPUs. All runs use bf16 and 10 epochs on 4 $\times$ A800 80GB.

Method	Backbone	LR	Per-dev BS	Grad accum	Eff. BS	Optimizer	Target / Notes
SFT	all	1e-5	4	4	64	AdamW (fused)	all weights updated
LoRA	all	5e-5	4	4	64	AdamW (fused)	$r=4$, $\alpha=8$, dropout=0, target=down_proj (all layers)
IA ³	all	1e-3	4	4	64	AdamW (fused)	target=down_proj (feed-forward modules)
LAS	Qwen-3B	5e-3	2	8	64	AdamW (HF default)	oral layers 20–35 (16) + oral heads via qwen.json (13 layers)
LAS	Llama-1B	1e-3	4	4	64	AdamW (HF default)	oral layers 5–15 (11) + oral heads via llama.json (8 layers)
LAS	Mistral-7B	5e-3	2	8	64	AdamW (HF default)	oral layers 15–31 (17) + oral heads via mistral.json (19 layers)

Table 14: Design comparison of activation-steering and PEFT baselines against CAS/LAS. Columns describe the operating space, per-component map, trainable parameters (d = hidden dimension, r = adapter rank), target components, use of gradient training, and calibration mechanism.

Method	Operates on	Per-component map	Params/comp.	Target components	Gradient training	Calibration
Few-shot	input prompt	demonstrations in prompt	0	input prompt	$\times$	none
ActAdd (Turner et al., 2024)	activation	$h \leftarrow h + \alpha v$	0	single grid-searched layer	$\times$	scalar α tuned per setting
CAS (ours)	activation	$h\ell \leftarrow h\ell + \alpha^*(h_{\text{level}}) v\ell$	0	all oral layers $\mathcal{L}_{\text{oral}}$	$\times$	validation-fitted $\alpha^*(t)$
SFT (full)	weights	$W \leftarrow W'$ (all weights updated)	$ W $	all weights	$\checkmark$	end-to-end
LoRA (Hu et al., 2022)	weights	$W \leftarrow W + BA$, $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times d}$	$2dr$	FFN down-projections	$\checkmark$	end-to-end
IA ³ (Liu et al., 2022)	weights	$y \leftarrow \ell \odot (Wh)$	d	K, V, FFN projections	$\checkmark$	end-to-end
ReFT (Wu et al., 2024)	activation	$h \leftarrow h + R^T(WRh + b - Rh)$, $R \in \mathbb{R}^{r \times d}$	$2dr + r$	residual stream of selected layers	$\checkmark$	end-to-end
LAS (ours)	activation	$h \leftarrow h + \mathbf{d} \odot h + \mathbf{b}$	$2d$	oral L+H ($\mathcal{L}_{\text{oral}} \cup \mathcal{A}_{\text{oral}}$)	$\checkmark$	end-to-end

residual is initialized to identity and the network learns position-conditional behavior via the affine’s interaction with the underlying hidden state. We experimented with restricting LAS to post-assistant-tag positions only; the restricted variant underperforms by 1.8–2.4 Level-Acc points across backbones because it loses the smoothing effect of all-position training, so we adopt the unrestricted form.

Decoding setup for evaluation. All quantitative results in Tables 5–4 use greedy decoding (temperature 0, top- p disabled) with max_new_tokens=500 and KV cache enabled. Each prompt is run once per random seed and we report the mean over three seeds {0, 1, 2}; standard deviations on L-Acc are ≤ 0.012 across seeds, well below the inter-method gaps in the main tables. For human evaluation (§5.4) we additionally sample with temperature 0.7, top- $p=0.9$ to introduce surface variation while preserving level control.

K Design Comparison of Baselines

We summarize the design choices of all baselines and our methods in Table 14. The table makes three axes of differentiation explicit: **where** the intervention happens (input/activation/weight), **what** the per-component map looks like (parameterized form), and **how** it is calibrated (none / scalar / polynomial / end-to-end training).

L Scale Check on Larger Backbones

Table 15 extends the supervised comparison to Qwen-14B and Qwen-72B under the same WS-

Model	Method	Params $\downarrow$	Tokens/s $\uparrow$	BERT $\uparrow$	L-Acc $\uparrow$	$\Delta R_r\downarrow$
Qwen-14B	SFT _{ref.}	14.0B	96.4	0.903	0.492	0.34
	LoRA	3.6M	95.7	0.902	0.453	0.40
	IA ³	0.7M	97.6	0.900	0.431	0.45
	LAS _{ours}	453K	94.8	0.904	0.463	0.37
Qwen-72B	SFT _{ref.}	72.0B	27.6	0.918	0.581	0.27
	LoRA	12.1M	27.1	0.911	0.527	0.29
	IA ³	2.4M	28.4	0.907	0.489	0.38
	LAS _{ours}	1.4M	26.8	0.919	0.598	0.25

Table 15: Supervised orality control on Qwen-14B/72B.

Pairs test setup and metrics as Table 5. LAS remains compact and competitive: it is 2.9 L-Acc points below SFT but 1.0 point above LoRA on Qwen-14B, and it leads SFT by 1.7 points and LoRA by 7.1 points on Qwen-72B. Throughput is similar across the lightweight methods. We treat this as a scale check for open-weight models with activation access, not evidence that LAS is universally simpler or cheaper to deploy than standard adapters.

M Qualitative Comparison Across Methods

To complement the aggregate metrics, we provide one side-by-side qualitative view of every method on the same written input. All outputs are produced by Qwen-3B with the prompt template in Listing 3; LAS and CAS additionally activate their respective steering modules. Outputs are taken verbatim from the released prediction files; the only edits are truncation indicators (“...”) for clearly degenerate loops and removal of trailing whitespace.

