The Listening Gap: Do Language Models Know What Warm Sounds Like?

Probing parametric audio-engineering knowledge against crowdsourced human gold, and testing whether retrieval grounding closes the gap.

Patrick Keough[1]

[1]Independent Researcher, pskeough@gmail.com

July 2026

Abstract

Evaluation of large language models has been outpaced by their deployment inside creative audio tooling, where systems now emit equalizer settings from words like "warm" with no check against what working engineers actually do. We probe this gap directly. Six production efficient-tier models from six families are tested on mapping semantic descriptors to 5-band parametric EQ curves against SAFE-DB, a crowdsourced corpus of real audio-engineer settings (1,473 clean entries after audit; only warm, $n = 532$, and bright, $n = 504$, clear the 20-entry floor). The models run zero-shot (Arm B), with retrieval of 5 individual human exemplars (Arm C), and with a consensus-centroid grounding variant (C_AGG), compared against embedding retrieval (A), the deployed keyword baseline (A′), a small embedding-to-MLP regressor (E), and a flat-EQ null. All scoring is computed against objective human gold in rendered frequency-response space; no LLM judge appears anywhere in the pipeline. Three results emerge. First, zero-shot models are hyper-central: median Human-Relative Percentile 3.0 where a typical human sits at 50, collapsing onto the consensus curve while engineers genuinely disagree (residual human dispersion 4.38/4.27 dB against a model cloud dispersion of 0.74 dB for B, a factor of 5.9 and 5.8 under-dispersion), and failing to reproduce the multimodal structure of human practice (22 of 24 school goodness-of-fit tests significant after correction). Second, grounding with individual exemplars degrades centrality rather than improving it: Arm C reaches median HRP 10.5 versus 3.0 for B ($q = 1.240 \times 10^{-15}$), while centroid grounding snaps to the anchor (C_AGG median HRP 0.0, context-pull 1.00). Models defer to aggregates and are destabilized by instances. Third, 37 of 53 tests survive Benjamini-Hochberg FDR at $\alpha = 0.05$. The listening gap is not access to human data; it is the kind of grounding the model receives.

1 Introduction

Evaluation of large language models in creative tooling has been far outpaced by their deployment. Text-to-audio-effect interfaces now ship in commercial and open-source form, emitting equalizer curves, compressor settings, and reverb parameters from plain-language requests. Whether the models behind these interfaces know what "warm" sounds like, as practicing audio engineers use the word, has never been tested against behavioral gold. Benchmark performance on general audio-knowledge questions does not answer this; a model can define warmth fluently while producing a curve no working engineer would recognize. This paper measures that gap directly.

The task under study is narrow by design: map a semantic descriptor to a 5-band parametric EQ curve, 13 parameters in total. The gold standard is SAFE-DB (Stables et al., ISMIR 2014) [1], a crowdsourced corpus in which each entry records a real engineer's full EQ setting labeled with the free-text descriptor they were pursuing. This corpus gives the study a property most LLM evaluations lack: the ground truth is numeric human behavior, collected in situ from working plugins a decade before any language model could have shaped it. No LLM judge appears anywhere in the pipeline, because none is needed. The judge-circularity and self-preference problems that complicate contemporary evaluation work are structurally absent here.

Four research questions organize the study. RQ1 is the knowledge probe: how accurately do LLMs map descriptors to curves, relative to the human distribution? RQ2 is the access-versus-capability question: does retrieval grounding over human-derived per-descriptor centroids outperform zero-shot generation? RQ3 asks whether generation adds value on top of retrieval: does a hybrid arm, generating with retrieved human exemplars in context, beat both pure retrieval and pure generation? RQ4 asks whether EQ priors diverge across model families or whether all families share one learned mapping.

Answering these questions honestly required three design positions that are contributions in themselves. First, all primary metrics are computed on rendered frequency responses rather than raw parameters, because parameter space is degenerate: distinct (frequency, gain, Q) settings produce near-identical transfer functions, and distances in parameter space punish differences no listener could hear. Second, the study confronts the mean artifact head-on. A per-descriptor centroid minimizes expected distance to held-out samples by mathematical construction, so raw distance-to-gold would manufacture a retrieval victory. Our primary centrality measure, the Human-Relative Percentile (HRP), re-expresses every prediction on the scale of human self-consistency: an HRP of 50 marks a prediction exactly as central as a typical engineer. Third, the gold is operationalized as a distribution rather than a point, because engineers disagree, and a validity probe confirms this dispersion is person-to-person disagreement rather than source-audio conditioning.

The findings invert the expected story. Zero-shot models are not scattered around the human cloud; they are hyper-central, with a median HRP of 3.0 against the typical human's 50. The models have compressed a genuinely contested practice into a single consensus answer. Grounding them with five individual human exemplars degrades centrality rather than improving it, moving the median HRP to 10.5, while grounding with the consensus centroid snaps predictions onto the anchor at HRP 0.0. Models defer to aggregates and are destabilized by instances. Of the 53 tests in the consolidated battery, 37 survive Benjamini-Hochberg FDR correction at $\alpha = 0.05$.

Our study makes four contributions. First, we present the first LLM evaluation on SAFE-DB, a corpus untouched by any language-model paper and idle since the flowEQ work [8]; the closest prior study, LLM2Fx (Doh et al., WASPAA 2025) [2], asked the descriptor-to-EQ question on SocialFX with prompting only and embedding-space MMD as its metric. Second, we contribute a distributional evaluation framework, combining HRP, energy distance, and mode-level goodness of fit in physically rendered curve space, that any study with behavioral gold can reuse. Third, we identify the aggregates-versus-instances grounding asymmetry, a mechanism-level finding about what kind of context helps a model and what kind harms it. Finally, the study operates under full receipts discipline: every number in this manuscript traces to a versioned statistics file, every API interaction is archived, and failed runs are reported rather than deleted.

2 Related Work

The question of whether language models hold usable audio-engineering knowledge was asked first by LLM2Fx (Doh et al., WASPAA 2025, arXiv:2505.20770) [2], and we position our study against it directly rather than claiming priority. LLM2Fx tested off-the-shelf models (GPT-4o, Llama 3.x, Mistral-7B) predicting 6-band EQ and reverb parameters zero-shot from semantic descriptors, with three in-context boosts, on the SocialFX corpus filtered to 273 usable entries across 7 to 9 descriptors, scoring by MMD between audio-embedding distributions. Our study differs on every structural axis. The corpus is SAFE-DB, 1,473 usable entries, untouched by any language-model paper; the metrics live in physically rendered curve space (log-spectral distance, human-relative percentile, energy distance, mode-level goodness of fit) rather than embedding space; and the design carries standalone retrieval baselines, an embedding-centroid arm and the deployed keyword system, that LLM2Fx lacks entirely. Where LLM2Fx asks whether model outputs land near human outputs in an embedding, we ask whether models reproduce the distribution of human practice, which turns out to be the question the data answers most sharply.

An adjacent cluster shares the problem space without sharing the question. InstructFX2FX (DAFx 2026, arXiv:2606.22005) [3] performs multi-turn iterative refinement of effect settings, a control problem rather than single-shot knowledge probing. Text2FX (ICASSP 2025, arXiv:2409.18847) [4] optimizes effect parameters at test time by CLAP gradients, with no language-model knowledge involved. Deng, Pardo, and Pappas (arXiv:2510.14249) [5] probe whether CLAP-family embeddings match human timbre perception and find they frequently do not, which supports the premise that embedding-space evaluation can miss perceptual reality. MixAssist (arXiv:2507.06329) [6] evaluates conversational mixing assistance with an LLM judge, our explicit contrast case: this study has no judge anywhere. Venkatesh et al. (JAES 2022, arXiv:2202.08898) [7] built the pre-LLM ancestor of our Arm E, a word2vec-to-regressor pipeline for automatic EQ, without any LLM, RAG, or comparison structure. On the corpus side, flowEQ (Steinmetz, 2019–20) [8] remains the only prior system trained on SAFE-DB Equaliser data, through a beta-VAE; the corpus has sat idle through the entire LLM era.

The comparison structure itself has precedent only on discrete tasks. Zero-shot versus retrieval versus fine-tuning has been studied on question answering (Ovadia et al., arXiv:2312.05934 [9]; Soudani et al., SIGIR-AP 2024 [10]), where answers are strings and correctness is categorical. We found no precedent for this comparison on continuous numeric regression against physical, perceptually grounded gold, and no precedent for evaluating whether models reproduce the mixture structure of a human behavioral distribution. Those two gaps, plus the untouched corpus, are what this paper contributes to an active line of work.

3 Methods

3.1 Data and provenance

Our evaluation corpus is the SAFE Equaliser user dataset (SAFE-DB) [1], collected through an instrumented VST plugin in which each entry records one engineer's full 5-band parametric setting alongside the free-text descriptor they were pursuing. Both original hosts of the file are dead. We retrieved the corpus from the Internet Archive Wayback Machine snapshot of 2020-01-22 of the true host, dmtlab.bcu.ac.uk, using the original-bytes modifier, and recorded its SHA-256

```
CC43858680115070FFDFA41A1158FCA1344DB2E0096BA73FE68990E90A284AF1
```

before locking the file read-only. All three Wayback captures spanning 2017 to 2020 share one identical content digest, so the retrieved file is provably the final upstream version; the provenance is stronger than a live download would have supplied. The file holds 1,700 data rows across 25 columns.

A fixed a-priori audit funnel produced the working corpus. Zero rows failed column-count, descriptor-presence, numeric-parse, gain-range, or frequency-validity checks. A descriptor blocklist of {test, 1, my} removed 183 entries, dominated by the junk descriptor "test" (181 entries), and 44 exact duplicates (same descriptor, IP, and all 13 parameters) were dropped at split time, leaving 1,473 clean entries. The descriptor distribution is two giants plus a long tail: of 327 distinct normalized descriptors, only warm ($n = 532$) and bright ($n = 504$) clear the pre-declared 20-entry floor, and no other descriptor reaches 10. The remaining 452 entries across 323 below-floor descriptors form a generalization set, scored only in aggregate. The audit also showed genre and instrument metadata to be 30 to 56% blank and contaminated with stimulus labels, so all arm inputs are descriptor-only.

3.2 Splits and contamination control

We split each above-floor descriptor 80/20 at the entry level with seed 42: warm contributes 419 training and 105 evaluation entries, bright contributes 398 and 99, for totals of 817 training and 204 evaluation entries. The training, evaluation, and tail files are each locked by SHA-256 hash, and every fitted statistic in the study (outlier thresholds, centroids, PCA bases, GMM components) is computed on the training split alone. The frozen evaluation set is read by exactly one script, the final metrics stage; any other consumer would constitute a protocol violation.

3.3 Rendering: the measuring instrument

All primary metrics operate on rendered frequency responses rather than raw parameters, because parameter space is degenerate: distinct (frequency, gain, Q) vectors can produce transfer functions no listener could separate. Each 13-parameter curve is rendered through an RBJ biquad cascade (low shelf, three peaking bands, high shelf; shelf $Q = 0.71$; $f_s = 48$ kHz) to magnitude in dB on a 256-point log-spaced grid from 20 Hz to 20 kHz. The renderer was unit-tested against analytically known filter responses before any result was computed, and all six checks pass; the inverse-cancellation test, summing +6 and −6 dB curves, leaves a maximum residual of 4.16×10^{-15} dB.

3.4 Arms

Seven arms compete. Arm A is embedding-centroid retrieval: the per-descriptor training centroid after 2-SD outlier rejection (warm 419 to 330 entries, bright 398 to 322), with similarity-weighted interpolation over descriptor embeddings for unseen tail descriptors. Arm A′ is an honesty arm reporting the deployed system as it actually ships, a keyword-overlap weighting that fell back to a flat curve on 312 of the tail descriptors; we report this behavior rather than hiding it, and nowhere describe A′ as embedding retrieval. Arm B prompts each language model with the descriptor alone and parses a 13-parameter JSON curve, with no corpus data in context. Arm C supplies the same prompt plus 5 individual training exemplars for the descriptor. Arm C_AGG replaces those exemplars with the single training consensus centroid, the exact curve Arm A emits, making it the mechanism probe for the grounding comparison. Arm E embeds the descriptor with a frozen MiniLM sentence encoder and regresses the 13 parameters through a small MLP with one 64-unit hidden layer, trained on the 817 training rows across 3 seeds; cross-seed spread in target space

is 0.08167 mean and 0.12734 maximum, so the fit is data-determined rather than initialization-determined. A flat-EQ null (all gains zero) completes the arm set as the better-than-nothing gate. Two LoRA fine-tuning arms (D, RAFT-style F) were specified and frozen but not executed for this manuscript; Section 6 discloses the full history.

3.5 Models, prompts, and sampling

The model slate consists of six production efficient-tier models spanning six families, accessed through OpenRouter: gpt-5.4-mini, claude-haiku-4.5, gemini-3.5-flash, deepseek-v4-flash, qwen3-235b-a22b-2507, and glm-4.7-flash. Every cell of the descriptor-by-model-by-arm grid receives 1 greedy sample (temperature 0) plus 10 samples at temperature 0.7, under 3 byte-frozen prompt variants (minimal, persona, format demonstration). Of 2,092 grid cells, 2,003 completed cleanly and exactly 1 cell failed all three parse attempts. An initial collection run failed on 549 cells, traced through the receipt log to hidden reasoning tokens exhausting the output budget on three models and an account-credit block on two flagship anchors that were later cancelled by a logged pre-analysis amendment; the affected cells were re-run under corrected settings, every failed attempt remains archived, and total API spend across all runs, retries included, was $6.39.

3.6 Metrics

Scoring proceeds on four levels. Centrality of a prediction is its median log-spectral RMSE (dB) to the held-out human curves of its descriptor. Because a per-descriptor centroid minimizes expected distance to held-out samples by mathematical construction, raw centrality would manufacture a retrieval victory, so the primary centrality measure is the Human-Relative Percentile: the percentile of a prediction's centrality within the leave-one-out human-to-human centrality distribution, where 50 marks a typical engineer and values near 0 mark predictions more central than almost any individual human. Distribution coverage is measured by energy distance between each sampled model cloud and the held-out human cloud, interpretable against a human split-half floor of -0.034 dB (warm) and 0.058 dB (bright); variance-ratio diagnostics quantify over- or under-dispersion. Population structure is assessed by fitting Gaussian mixture models to training curves in PCA space, assigning each model curve to its nearest human mode, and testing the resulting share vector against human shares by chi-square goodness of fit.

3.7 Statistical analysis

The statistical plan followed its pre-declared fallback. With only 2 descriptors above floor, descriptor-paired testing is impossible, so analysis runs at the entry level with descriptor as a stratum, exactly as the design document committed in advance. Arm contrasts use paired Wilcoxon signed-rank tests on matched units, with bootstrap 95% confidence intervals (10,000 resamples) on headline quantities. Every inferential test in the study, 53 in total across four analysis stages, enters one consolidated Benjamini-Hochberg FDR correction at $\alpha = 0.05$, and the full table is reported with raw p, adjusted q, and survival status, including the 16 tests that fail. No LLM judges any output at any point; the gold is numeric human behavior, which removes judge circularity and self-preference from the threat model entirely.

Because SAFE-DB has been public since 2014, we ran a memorization probe before drafting: all six models were asked to reproduce SAFE-DB rows and summary statistics under structured prompts. Every model signals awareness that the dataset exists, yet verbatim recall was 0 of 6 models and

statistical recall was 0 of 6 models. Memorization would in any case bias toward the models, making our headline findings conservative, but the probe shows no evidence the corpus is retrievable from any tested model.

4 Results

Every claim in this section is gated by the consolidated 53-test Benjamini-Hochberg battery, of which 37 tests survive at $\alpha = 0.05$; contrasts that fail correction are named as failures where they matter.

4.1 Humans disagree, models do not

The central result of the study is an inversion of the expected failure mode. Median HRP by arm on the frozen evaluation set: FLAT 32.2, A 2.4, A′ 2.4, E 3.8, B 3.0 (bootstrap 95% CI 3.0 to 3.8, $n = 396$), C 10.5 (CI 8.1 to 13.3, $n = 396$), C_AGG 0.0 (CI 0.0 to 4.8, $n = 132$). A typical human engineer sits at 50 by construction. Zero-shot language models are not scattered outside the human range; they are more central than 97% of individual engineers, landing nearly on top of the consensus centroid that Arm A computes explicitly from training data. The dispersion side completes the picture. Where residual human dispersion is 4.38 dB for warm and 4.27 dB for bright, the zero-shot model cloud disperses only 0.74 dB, a factor of 5.9 (warm) and 5.8 (bright) under-dispersion; Arm C reaches 0.92 dB and C_AGG collapses to 0.16 dB. The models have compressed a contested practice into a point.

A validity probe secures the dispersion claim against its sharpest objection, that human spread merely reflects different source audio. Regressing rendered curves on 78 source-audio features gives out-of-fold R^2 of $+0.051$ for warm and -0.061 for bright, barely above the permutation floors of -0.119 and -0.115, and residual dispersion is essentially identical to raw dispersion. Human spread is genuine engineer-to-engineer disagreement. One caveat accompanies the centrality result: the flat null itself sits at HRP 32.2, inside the wide human band, so centrality alone is a weak bar; the informative quantities are the dispersion gap and the mixture structure below.

4.2 Human practice is multimodal, and models miss the mixture

BIC-selected Gaussian mixtures on training curves find $k = 3$ modes for warm (shares 16, 53, and 31%) and $k = 4$ for bright (18, 25, 30, and 27%): warmth and brightness are families of practices, not single curves. Model outputs concentrate instead of mixing. Assigning each model curve to its nearest human mode, gemini-3.5-flash places 33 of 33 warm zero-shot samples into a single mode (chi-square $p = 1.17 \times 10^{-16}$), and claude-haiku-4.5 places 32 of 33 bright samples into one mode. Of the 24 school goodness-of-fit tests entering the consolidated battery, 22 survive correction; the two failures are glm-4.7-flash warm under B and C, both at $q = 5.206 \times 10^{-2}$. The population-level conclusion holds across every family: no tested model reproduces the mixture of human schools.

4.3 Grounding with individual exemplars degrades; grounding with the aggregate anchors

Supplying 5 individual human curves in context moves the median HRP from 3.0 (B) to 10.5 (C), a significant degradation (T1.1, paired $n = 396$, $p = 1.068 \times 10^{-16}$, $q = 1.240 \times 10^{-15}$), and

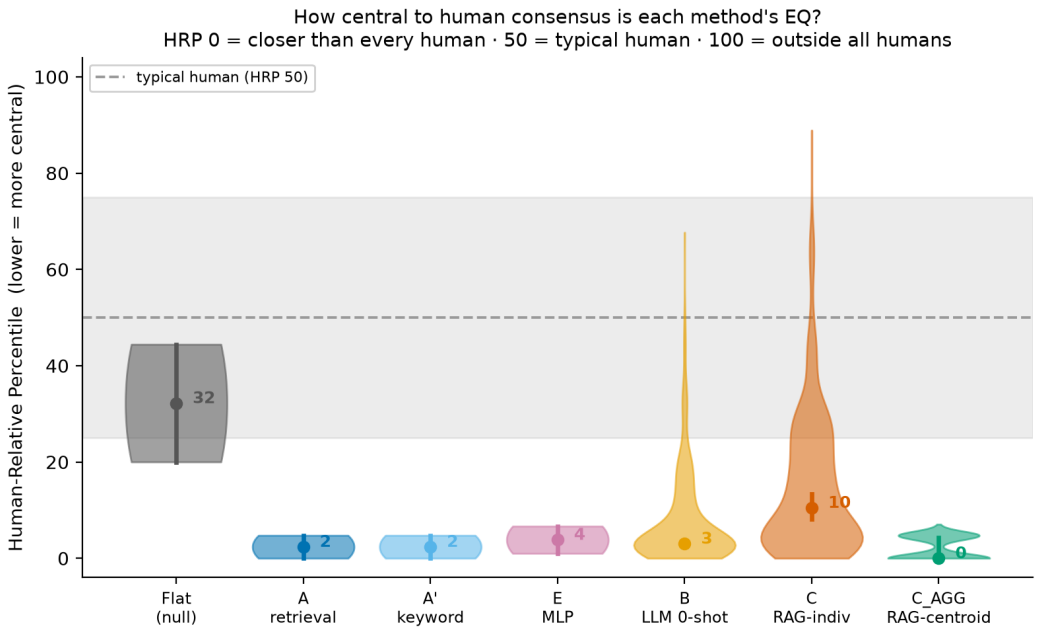

Figure 1: Median Human-Relative Percentile (HRP) by arm on the frozen evaluation set. A typical human engineer sits at HRP 50 by construction; the arms order FLAT 32.2, A 2.4, A′ 2.4, E 3.8, B 3.0, C 10.5, and C_AGG 0.0, showing that zero-shot models are more central than 97% of individual engineers.

the same ordering appears in raw centrality (T2, median LSD B 3.929 dB versus C 4.103 dB, $p = 9.207 \times 10^{-18}$, $q = 1.627 \times 10^{-16}$). Replacing the 5 instances with the single training centroid reverses the effect entirely: C_AGG snaps to the anchor at median HRP 0.0, beating B (T1.2, $n = 132$, $q = 2.742 \times 10^{-8}$) and beating C decisively (T1.3, $n = 132$, $q = 1.118 \times 10^{-19}$, median C 14.2 versus C_AGG 0.0 on the matched subset). The context-pull metric explains the mechanism. Under centroid grounding every model defers almost completely, median pull 1.00 across all six models; under instance grounding the pull is partial and variable, ranging from 0.51 (deepseek-v4-flash) to 0.87 (claude-haiku-4.5). Models treat an authoritative-looking aggregate as an instruction and a set of noisy individual examples as a perturbation. On the distribution-matching side, the energy-distance contrast between B and C fails correction (T3, median B 2.235 dB versus C 2.477 dB, $q = 5.206 \times 10^{-2}$) and is reported as suggestive only.

4.4 Family divergence and prompt sensitivity

Families diverge, and one family diverges from all the rest. Zero-shot HRP differs across the six models (T5, $q = 1.802 \times 10^{-19}$). Five families cluster tightly: claude-haiku-4.5 at median 2.0, gemini-3.5-flash at 2.9, and deepseek-v4-flash, gpt-5.4-mini, and qwen3-235b-a22b-2507 at 3.0. The outlier is glm-4.7-flash at 18.6, the same family whose school GOF tests were the only correction failures. Prompt framing also registers (T6, $q = 3.440 \times 10^{-5}$): the minimal and format-demonstration variants both give median HRP 3.0 while the persona variant shifts to 6.9, a measurable but small perturbation that leaves the hyper-centrality conclusion intact under every variant.

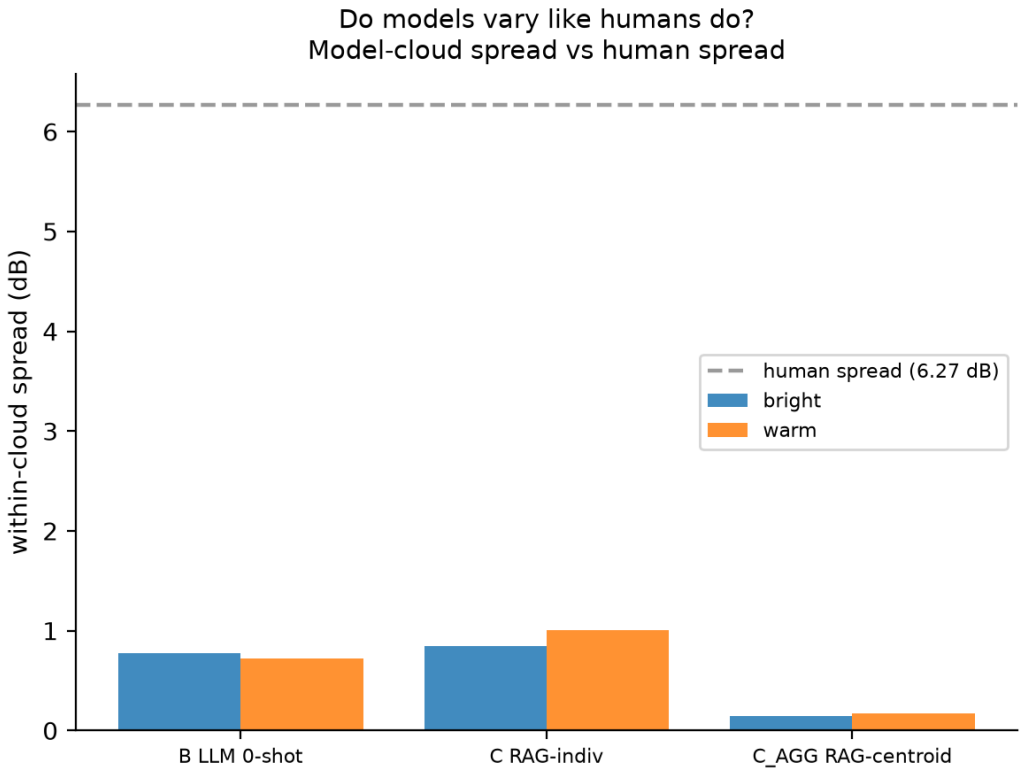

Figure 2: Dispersion of model clouds versus human dispersion. Residual human dispersion is 4.38 dB for warm and 4.27 dB for bright, while the zero-shot model cloud (Arm B) disperses only 0.74 dB, a factor of 5.9 (warm) and 5.8 (bright) under-dispersion; Arm C reaches 0.92 dB and C_AGG collapses to 0.16 dB.

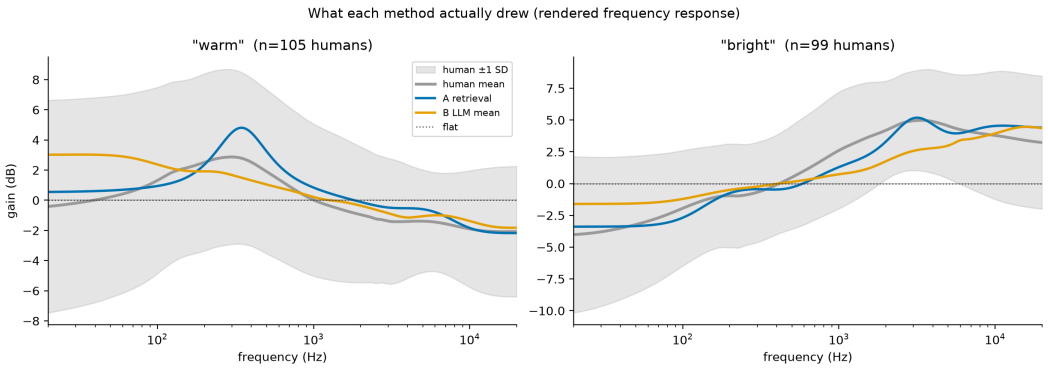

Figure 3: Rendered frequency-response overlay. Each 13-parameter curve is rendered through an RBJ biquad cascade to magnitude in dB on a 256-point log-spaced grid from 20 Hz to 20 kHz; model predictions are compared against the held-out human curves of their descriptor in this rendered curve space, where human practice is multimodal and model outputs concentrate.

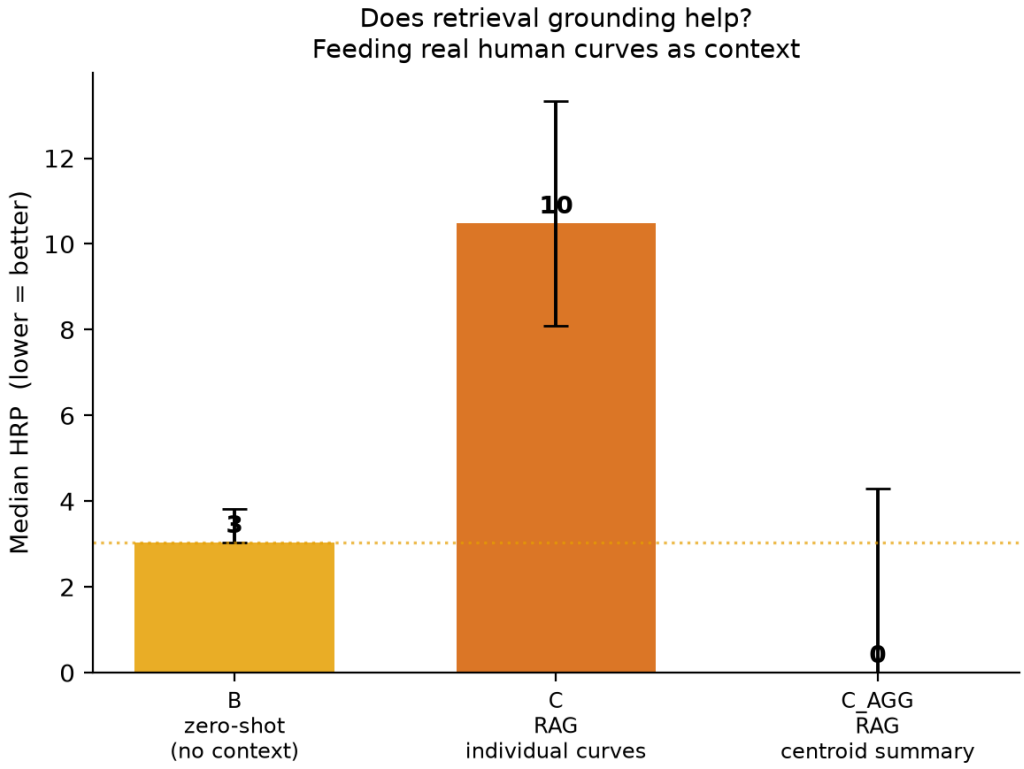

Figure 4: The RAG effect: B versus C versus C_AGG. Supplying 5 individual human exemplars in context degrades centrality, moving the median HRP from 3.0 (B) to 10.5 (C), while replacing the instances with the single training consensus centroid snaps predictions onto the anchor (C_AGG median HRP 0.0, context-pull 1.00): models defer to aggregates and are destabilized by instances.

4.5 Band anatomy: where the errors live

Against the training consensus, greedy zero-shot curves for warm over-boost the low band by +2.07 dB and cut the mids by −1.46 dB where engineers do not; errors on lowmid, highmid, and high are −0.27, +0.56, and +0.85 dB. For bright the models under-shoot the defining move: highmid error is −2.23 dB and high is −1.72 dB against the practiced consensus lift, with low at +1.06, lowmid +0.57, and mid +0.08 dB. Models exaggerate the textbook bass move on warm and underweight the treble move that practicing engineers actually make on bright.

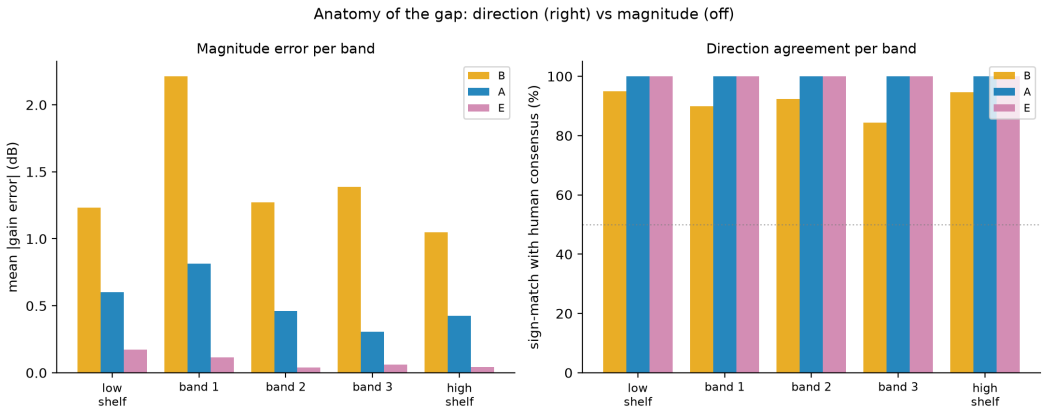

Figure 5: Band-error anatomy against the training consensus. On warm, greedy zero-shot curves over-boost the low band by +2.07 dB and cut the mids by −1.46 dB where engineers do not; on bright, the models under-shoot the defining move, with highmid error −2.23 dB and high −1.72 dB against the practiced consensus lift.

4.6 On unseen descriptors, nothing beats flat

The 452-entry long tail tests generalization beyond warm and bright. The B-versus-C degradation replicates there (T4, paired $n = 359$, median LSD B 6.030 versus C 6.183 dB, delta +0.096, $p = 9.429 \times 10^{-5}$, $q = 2.082 \times 10^{-4}$). Per-descriptor median tail LSD orders the arms A 5.422, A′ and FLAT 5.747, E 5.831, C_AGG 5.938, B 6.117, C 6.285 dB, with A, A′, FLAT, and E scored over all 323 tail descriptors and the LLM arms over the 60 descriptors all arms share; head-to-head comparisons run on those 60 common descriptors. The pairwise record is sobering. The surviving contrasts all run in the negative direction: E loses to A′, FLAT, and A; C loses to A, A′, FLAT, E, and C_AGG. The contrasts that would establish a positive winner all fail correction: FLAT versus B ($p = 0.4616$), FLAT versus A ($p = 0.1294$), A versus B ($p = 0.1053$), and the direct tail B versus C ($p = 0.0668$). The honest summary is that on rare descriptors no method significantly outperforms doing nothing, and the statistically defensible claims are that instance-grounded RAG and the small regressor perform significantly worse than trivial baselines.

5 Discussion

The natural reading of a knowledge probe is that models either know the mapping or they do not. Our results resist that framing. The tested models encode the consensus meaning of warm and bright with striking accuracy, landing at median HRP 3.0 where the consensus anchor itself sits at

2.4, and they encode it as a point. What the models lack is not the direction of the mapping but its distribution: the knowledge that warmth is a contested practice with at least three identifiable schools, and that a working engineer's curve is drawn from that contest rather than from its average. Direction known, distribution lost is the granularity signature this line of work keeps finding, here expressed in decibels rather than severity scores.

The mixture evidence points at where the knowledge came from. Models concentrate their samples inside single human modes rather than spreading across the mixture, and the band anatomy matches written doctrine over observed behavior: on warm the models exaggerate the textbook bass boost by +2.07 dB, and on bright they miss the high-mid lift practicing engineers actually apply by −2.23 dB. A model trained on text about warmth learns the sentence "warm means more bass, less treble" and renders it faithfully. The behavioral corpus shows engineers doing something messier, more varied, and in the case of bright, concentrated in a band the textbook sentence does not emphasize.

For retrieval-augmented system design the aggregates-versus-instances asymmetry is the exportable finding. Under centroid grounding every model defers almost completely (context-pull 1.00); under instance grounding the pull is partial, 0.51 to 0.87, and the resulting curves are less central than generating with no grounding at all. Five noisy human curves do not read to the model as a distribution to sample from; they read as conflicting instructions that destabilize a good prior. For numeric generation tasks the practical recipe that follows is retrieve-then-aggregate: summarize retrieved exemplars into a consensus structure before context insertion, rather than dumping raw instances into the prompt.

The access-not-capability framing that motivated RQ2 partially inverts on this task. The models already hold access to the consensus; zero-shot output nearly coincides with the centroid a retrieval system would compute. What no arm restores is dispersion. The zero-shot cloud spreads 0.74 dB against human residual dispersion of 4.38 and 4.27 dB, instance grounding lifts it only to 0.92 dB, and centroid grounding collapses it to 0.16 dB. Producing the human distribution, rather than a central point in it, is unsolved by every mechanism we tested.

Two further observations bound the practical picture. On the long tail of rare descriptors every method fails: no arm significantly beats a flat curve, which means deployed text-to-EQ tools serving open vocabulary are, on current evidence, decorating a no-op for most requests. And family divergence is real but concentrated: five of six models cluster at median HRP 2.0 to 3.0, while glm-4.7-flash sits apart at 18.6 and is also the only family whose mixture tests failed to reach significance.

6 Limitations

Several limitations bound these conclusions. First and most important, the planned training axis was not executed. Arms D (LoRA fine-tune) and F (RAFT) were specified, frozen, and gate-labeled as the fine-tuning corners of a 2×2 grounding factorial, but the full 3-seed run hung from GPU contention (a CUDA wedge at step 30 of 460, killed after roughly 166 minutes of wall time), and the only surviving output is an under-sampled single-seed smoke configuration producing 4 predictions with pathological railed gains at ±12 dB against gentle human warm settings near +3.8/−2.7 dB. Reporting that artifact as an arm would be an overclaim, so we decline, and the training axis moves to future work; the full history is disclosed in the project decision log (D-031).

Second, the study uses one corpus. SAFE-DB supplies only two confirmatory descriptors, warm and bright, the tail is scored only in aggregate, and external replication on SocialFX is the obvious

next step. Third, the model slate is scoped to production efficient-tier models; flagship anchor cells were cancelled before analysis by a logged cost amendment, so no claim in this paper extends to flagship-tier behavior. Fourth, bootstrap confidence intervals are anti-conservative to the extent that units within a model share sampling draws. Fifth, the audio-conditioning probe carries its own caveat: a low R^2 could reflect summary features that miss the relevant acoustic dimension rather than true absence of conditioning, so we report the permutation-controlled number and claim no causation. Sixth, the expertise and stimulus-context analyses rest on self-reported metadata that is roughly half blank and are exploratory descriptives only. Finally, no listening test was conducted: this paper claims nothing about how any curve sounds, only how it compares to what engineers measurably did.

7 Conclusion

This paper measured a specific gap. Language models know what warm means; they do not know how warmly engineers disagree. Across six model families the zero-shot answer to a semantic EQ request is the consensus curve delivered with a fraction of the human variance, retrieval of individual human examples makes the output worse, and an aggregate anchor turns the model into a relay. The finding argues that evaluation of generative numeric tasks should measure distribution matching, not only point accuracy, because a system can score excellently on centrality while erasing the structure of the practice it imitates. Every number in this manuscript traces to an archived statistics file, and every significance claim passed a consolidated 53-test FDR battery of which 37 tests survive. The training axis (LoRA and RAFT), replication on SocialFX, and listening tests are the committed next steps.

References

[1] Stables, R., Enderby, S., De Man, B., Fazekas, G., Reiss, J. D. "SAFE: A system for the extraction and retrieval of semantic audio descriptors." ISMIR 2014.

[2] Doh, Koo, Martinez-Ramirez, Liao, Nam, Mitsufuji. "LLM2Fx." WASPAA 2025. arXiv:2505.20770.

[3] InstructFX2FX. DAFx 2026. arXiv:2606.22005.

[4] Text2FX. ICASSP 2025. arXiv:2409.18847.

[5] Deng, Pardo, Pappas. arXiv:2510.14249.

[6] MixAssist/MixParams. arXiv:2507.06329.

[7] Venkatesh et al. "Word Embeddings for Automatic Equalization." JAES 2022. arXiv:2202.08898.

[8] Steinmetz, C. "flowEQ." 2019–20.

[9] Ovadia et al. arXiv:2312.05934.

[10] Soudani et al. SIGIR-AP 2024.