MELISSA: Multi-level Evaluation with LLM-based Integrated Self-Scrutiny and Auditing

Amirhossein Afsharrad[1,2*] **Sri Jaladi**[1,3*] **Nima Yazdani**[2]
Ali Ansari[1,2] **Seyed Shahabeddin Mousavi**[1,2] **Sanjay Lall**[1]

[1]Stanford University [2]micro1 [3]Human Behavior

{afsharrad, sjaladi, aliansarinik, ssmousav, lall}@stanford.edu
{nima, ali}@micro1.ai

Abstract

As AI systems increasingly conduct complex multi-turn interactions, reliable evaluation becomes critical yet challenging. Current LLM-as-judge approaches suffer from severe biases and struggle with lengthy conversations, while monolithic evaluation misses quality variations across dialogue segments. We present MELISSA, a framework that hierarchically decomposes conversations and learns bias corrections from human judgments, requiring no model fine-tuning. Evaluating 100 AI-conducted technical interviews with expert annotations reveals surprising insights: bias correction alone reduces error by over 50%, indicating LLMs struggle with scale calibration rather than quality discrimination; properly aligned GPT-4o-mini outperforms unaligned Claude 3.5[2], enabling order-of-magnitude cost reductions; and optional audit mechanisms show mixed results—while potentially providing confidence signals, they often introduce unnecessary edits that degrade performance when models already produce well-calibrated evaluations, highlighting the importance of empirical validation over intuitive design. These findings demonstrate that simple calibration transforms weak models into reliable judges, while reinforcing that high-quality human data remains essential for automated evaluation systems.

1 Introduction

The proliferation of AI systems in complex, multi-turn interactions—from customer service to technical interviews—has created an urgent need for reliable evaluation methods. Traditional human evaluation, while remaining the gold standard, faces scalability constraints: expert annotators are expensive, evaluation is time-consuming, and maintaining consistency across evaluators proves challenging. These limitations have driven the development of automated evaluation approaches, with the LLM-as-judge paradigm emerging as a promising solution. Early studies reported substantial agreement between LLM judges and human evaluators [Zheng et al., 2023], spurring adoption across various applications including dialogue systems [Fu et al., 2023], chatbot responses [Wang et al., 2023], and complex reasoning tasks [Liu et al., 2023].

However, systematic investigations have revealed fundamental reliability issues in LLM-based evaluation. Position bias—where models favor responses based on their position rather than quality—causes consistency rates as low as 22.7% [Shi et al., 2024]. Self-enhancement bias leads models to overrate their own outputs by 8.91% on average [Ye et al., 2024]. Even minor prompt template variations can cause 76-point accuracy swings, suggesting extreme brittleness in evaluation behavior [Sclar et al.,

*Equal contribution
[2]All references to Claude models in this paper refer to the Sonnet variant of each version.

2023]. The CALM framework identifies 12 distinct bias types affecting LLM judges, with robustness rates varying from 0.566 to 0.832 across models and tasks. While architectural innovations like PORTIA demonstrate consistency improvements through specialized designs [Li et al., 2024], these approaches require substantial engineering effort and still leave significant reliability gaps.

The challenges compound dramatically in multi-turn settings, where conversations span dozens of exchanges across diverse topics. Current monolithic approaches attempt global assessment of entire conversations, missing critical quality variations across segments. LLMs exhibit 15-30% performance degradation when evaluating multi-turn versus single-turn interactions [Bai et al., 2024, MT-Eval Team, 2024], struggle with dialogue-level phenomena like consistency and coherence [Chen et al., 2024], and fail to capture professional assessment nuances [Kwon et al., 2024]. As conversations grow longer—our technical interviews often exceed 20 minutes—computational costs become prohibitive when using state-of-the-art models, while smaller models lack the capability for accurate holistic assessment. Existing hierarchical approaches show promise: LLM-Rubric achieves 2× error reduction through multi-dimensional evaluation [Hashemi et al., 2024], HRM proves more stable across granularities [Wang et al., 2025], and ConvBench reveals evaluation failure cascades [Liu et al., 2024a]. Yet none address the fundamental calibration problem we identify: LLMs can distinguish quality levels but systematically miscalibrate their scores relative to human scales.

We introduce MELISSA (Multi-level Evaluation with LLM-based Integrated Scoring and Alignment), a framework combining hierarchical decomposition with learned bias correction that requires no model fine-tuning and can be implemented with standard prompting. Our contributions include: **(1) Novel dataset:** 100 AI-conducted technical interviews by Zara [Zhou et al., 2024, Allbert et al., 2025] with expert human annotations, addressing the critical gap in publicly available benchmarks for extended multi-turn evaluation. **(2) Practical framework:** MELISSA employs L hierarchical levels with relevance weighting, N independent evaluation passes for robustness, and constrained optimization for learning weights and biases. The framework requires only standard LLM API calls—no fine-tuning, specialized training, or reward model modifications. **(3) Critical insights on bias correction:** Through systematic ablation, we demonstrate that bias correction dominates performance improvements—reducing error by over 50% even with uniform weights—while weight optimization provides smaller refinements. This finding suggests LLMs' primary challenge lies in scale calibration rather than quality discrimination, challenging prior emphasis on architectural complexity [Wang et al., 2024, Liu et al., 2024b, He et al., 2024]. **(4) Enabling weak models:** Proper alignment enables dramatically smaller models to match larger ones: calibrated GPT-4o-mini outperforms unaligned Claude 3.5, reducing evaluation costs by orders of magnitude while maintaining quality. **(5) Understanding audit mechanisms:** Our experiments reveal that prompting LLMs to audit their own evaluations often triggers unnecessary edits that degrade performance—some models show 80% deterioration. This finding, supporting literature on self-correction failures [Kamoi et al., 2024], led us to make auditing optional and highlights the importance of empirical validation over intuitive design choices. **(6) Practical evaluation metric:** We propose Threshold Absolute Error (TAE) as a simple complement to MAE, recognizing that predictions within ±0.5 of targets are functionally equivalent in discrete 5-point scoring systems commonly used in practice.

While recent benchmarks expand evaluation scope—Arena Hard's 500 queries [LMSYS Team, 2024], WildBench's 0.98 correlation with humans [Lin et al., 2024], LiveBench's contamination resistance [LiveBench Team, 2024]—MELISSA provides a complementary process framework rather than fixed test sets. Our results reinforce that high-quality human data remains fundamental: the performance ceiling of any learning-based evaluation system is bounded by its training data quality [Northcutt et al., 2021, Plank, 2022]. As automated evaluation becomes prevalent, human judgment paradoxically becomes more, not less, critical for maintaining alignment with human values.

2 The MELISSA Framework

MELISSA evaluates multi-turn interactions through a systematic pipeline that decomposes complex conversations into a hierarchy of L evaluation levels with relevance-aware aggregation.

2.1 Overview and Problem Setup

Let $\mathcal{C}$ be a conversation with T turns between agents (human or artificial). We decompose this conversation into L hierarchical levels, where level $\ell \in \{1, \dots, L\}$ contains units $U_\ell = \{u_{\ell,1}, \dots, u_{\ell,|U_\ell|}\}$.

Each unit represents a conversational segment at that level's granularity—individual turns at Level 1, topical sections at intermediate levels, and the complete conversation at Level L.

For each criterion c in the evaluation criteria set $\mathbb{C}$, MELISSA produces a final score S_c by hierarchically aggregating evaluations across all levels, with relevance weighting and optional bias correction. We use N independent evaluation trials per unit for robustness, with scores denoted $s^{\text{init}}_{c,\ell,u,n}$ for initial evaluations and $s^{\text{audit}}_{c,\ell,u,n}$ when optional auditing is enabled. Relevance weights $r_{c,u}$ determine each unit's contribution, leading to relevance-weighted means $\bar{s}_{c,\ell}$ at each level. The framework learns weights $\mathbf{w}_c$ and bias terms b_c to align with human judgments. See Appendix A for complete notation.

2.2 Hierarchical Decomposition

The framework operates on L hierarchical levels, where $L \geq 2$ can be chosen based on application requirements. The optimal choice of L and unit boundaries depend on both the application domain and specific evaluation criterion.

General level structure. The hierarchical decomposition typically follows this pattern: Level 1 captures individual exchanges between agents (turns), assessing immediate responsiveness and local coherence. Intermediate levels (2 through $L-1$) contain progressively larger conversational units that group related content, such as topical sections or conversation phases. Level L encompasses the complete interaction transcript, evaluating overall trajectory and outcomes. This multi-granularity approach enables MELISSA to capture both local quality variations and global coherence patterns that single-level evaluation would miss.

Criterion-specific adaptation. Different criteria may benefit from different hierarchical structures even within the same conversation. For instance, in a technical interview, "Technical Depth" might decompose along topical boundaries (algorithms, system design, databases), while "Communication Skills" might align better with conversational phases (introduction, technical discussion, closing). This flexibility enables precise evaluation by matching the hierarchical structure to how each quality dimension naturally manifests. The key insight is that optimal decomposition varies not just by conversation type but by what aspect of quality is being measured. See Appendix G for more details.

2.3 Relevance Weighting

Not all conversational units contribute equally to every evaluation criterion. We introduce relevance weighting to focus evaluation on pertinent content:

$$r_{c,u} = f_{\text{rel}}(u, c) \in [0, 1] \text{ or } \{0, 1\}, \tag{1}$$

where f_{rel} is typically implemented via an LLM prompted to assess the relevance of unit u for criterion c. This can be:

- **Binary filtering** ($r_{c,u} \in \{0, 1\}$): Excludes irrelevant units entirely
- **Continuous weighting** ($r_{c,u} \in [0, 1]$): Assigns importance weights

For example, greeting exchanges might receive $r_{\text{TQQ},u} = 0$ for Technical Question Quality while $r_{\text{HLI},u} = 1$ for Human-like Interaction.

2.4 Multi-Pass Evaluation with Optional Audit

To ensure statistical robustness, each evaluation unit is assessed through N independent trials:

$$s^{\text{init}}_{c,\ell,u,n} = f_{\text{eval}}(u, c, \ell) \quad \forall u \in U_\ell,\, n \in \{1, \dots, N\} \tag{2}$$

where f_{eval} is realized through an LLM with appropriate prompting that specifies the evaluation criterion, level context, and scoring scale. Multiple trials improve robustness through averaging and enable variance assessment. When additional confidence is needed, an optional audit stage reviews each initial score:

$$s^{\text{audit}}_{c,\ell,u,n} = f_{\text{audit}}(u, s^{\text{init}}_{c,\ell,u,n}, c) \tag{3}$$

where f_{audit} prompts an LLM to review and potentially revise the initial score. The audit mechanism is particularly useful when initial evaluations show high variance or when deployment requires additional confidence guarantees.

2.5 Hierarchical Aggregation

Level scores are computed using relevance-weighted averaging:

$$\bar{s}_{c,\ell} = \frac{\sum_{u \in U_\ell} r_{c,u} \cdot \left(\frac{1}{N} \sum_{n=1}^{N} s_{c,\ell,u,n} \right)}{\sum_{u \in U_\ell} r_{c,u}} \tag{4}$$

where $s_{c,\ell,u,n} = s^{\text{audit}}_{c,\ell,u,n}$ if audit is enabled, otherwise $s_{c,\ell,u,n} = s^{\text{init}}_{c,\ell,u,n}$. For binary relevance weights, this simplifies to averaging only over units where $r_{c,u} = 1$. Final scores combine level scores with learned or default weights:

$$S_c = \alpha_c \cdot \left(\sum_{\ell=1}^{L} w_{c,\ell} \cdot \bar{s}_{c,\ell} \right) + b_c \tag{5}$$

where $\mathbf{w}_c = (w_{c,1}, \ldots, w_{c,L})$ are level weights, α_c is a scale alignment factor (typically 1 when scales match), and b_c is a bias correction term. Parameters can be set to defaults (uniform weights, $\alpha_c = 1$, $b_c = 0$) or learned from human judgments (Section 3).

Algorithm 1 MELISSA Evaluation Pipeline

```
Require: Conversation 𝒞, Criteria ℂ, Parameters (N, L, audit flag)
Ensure: Final scores {S_c} for all criteria
Decompose 𝒞 into L hierarchical levels {U_1, ..., U_L}
for each criterion c ∈ ℂ do
    Assess relevance r_{c,u} for all units using Eq. 1
    Perform N independent evaluations for relevant units using Eq. 2
    if audit enabled then Review each score using Eq. 3
    Compute relevance-weighted level scores s̄_{c,ℓ} using Eq. 4
    if human scores available then
        Learn optimal w*_c and b*_c via constrained optimization (Section 3)
    else Use uniform weights: w_{c,ℓ} = 1/L, b_c = 0
    Aggregate final score using Eq. 5
end for
return {S_c : c ∈ ℂ}
```

Algorithm 1 presents the simplified MELISSA pipeline. See Algorithm 2 in Appendix B for a more detailed implementation.

3 Parameter Learning

While MELISSA can operate with default uniform weights, learning optimal parameters from human judgments significantly improves alignment and reduces evaluation error. This section describes the optimization procedure for learning level weights and bias terms.

3.1 Optimization Objective

Given m conversations with human scores $\{y_c^{(i)}\}_{i=1}^m$ for criterion c, we minimize the mean squared error between predictions and human judgments:

$$\begin{aligned} \min_{\mathbf{w}_c \in \mathbb{R}^L, b_c \in \mathbb{R}} \quad & \sum_{i=1}^{m} \left(y_c^{(i)} - \alpha_c \cdot \sum_{\ell=1}^{L} w_{c,\ell} \cdot \bar{s}^{(i)}_{c,\ell} + b_c \right)^2 \\ \text{subject to} \quad & \sum_{\ell=1}^{L} w_{c,\ell} = 1, \quad w_{c,\ell} \geq 0 \quad \forall \ell \in \{1, \ldots, L\} \end{aligned} \tag{6}$$

where $S_c^{(i)} = \alpha_c \cdot \sum_{\ell=1}^{L} w_{c,\ell} \cdot \bar{s}^{(i)}_{c,\ell} + b_c$ is our prediction for sample i, and $\bar{s}^{(i)}_{c,\ell}$ denotes the relevance-weighted level score (as defined in Eq. 4) computed for sample i at level ℓ. The constraints ensure weights form a valid probability distribution over levels. We fit this model using MSE loss to ensure strict alignment during optimization, penalizing all deviations from human judgments.

where $\hat{y}_c^{(i)}$ is our prediction for sample i, and $\bar{s}^{(i)}_{c,\ell}$ denotes the relevance-weighted level score (as defined in Eq. 4) computed for sample i at level ℓ. The scale alignment factor α_c is fixed as the ratio between the maximum possible human score and the maximum model-generated score, ensuring proper scale matching when these differ. In our experiments with matching 1-5 scales, $\alpha_c = 1$. The constraints ensure weights form a valid probability distribution over levels.

This is a convex optimization problem (quadratic objective with linear constraints), which we solve using standard convex optimization solvers. While MELISSA can operate with default uniform weights ($\mathbf{w}_c = \mathbf{1}/L$) and zero bias ($b_c = 0$), our experiments (Section 5) demonstrate that learning these parameters from human judgments significantly improves alignment. We conduct ablation studies to analyze the relative importance of weight optimization versus bias correction, revealing insights into how LLM judges differ from human evaluators and how proper alignment enables smaller models to achieve competitive performance.

3.2 Evaluation Metrics

We distinguish between training and evaluation metrics:

Training: Mean Squared Error (MSE) as in Eq. 6 enforces strict alignment during optimization, penalizing all deviations from human judgments regardless of magnitude. This stricter approach during training ensures the model learns accurate calibration across the entire score range. The smooth gradients provided by MSE are an additional benefit for optimization stability.

Evaluation: Before introducing our metrics, it is important to understand the nature of predicted and ground truth values in our framework. MELISSA's raw predictions from individual LLM evaluations are integers on the 1-5 scale when $N = 1$. However, our final predictions $\hat{y}_i^c$ are real-valued due to two factors: (1) averaging across N independent trials, and (2) the weighted linear combination across levels with learned weights and bias. Ground truth human scores y_i^c are inherently integers when provided by a single expert, but become non-integer when averaged across multiple experts. Specifically, with k experts providing integer scores, the averaged ground truth can only take values from the set $\{\frac{j}{k} : j \in \mathbb{Z}, k \leq j \leq 5k\}$. See Section 4 for details on our human evaluation protocol.

We report two complementary metrics:

- **Mean Absolute Error (MAE)**: Standard metric for comparison with prior work

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i| \tag{7}$$

- **Threshold Absolute Error (TAE)**: Acknowledges that predictions within $\pm\tau$ of the target are functionally equivalent:

$$\text{TAE}_\tau(y, \hat{y}) = \max(0, |y - \hat{y}| - \tau) \tag{8}$$

For our 1-5 scale evaluations, we use $\tau = 0.5$, recognizing that differences smaller than half a scale point are practically meaningless. This threshold is particularly appropriate because in real-world applications, scores are typically rounded to the nearest integer for display and decision-making on Likert scales. With TAE, a prediction of 3.4 receives zero loss when the ground truth is 3 (since $|3.4 - 3| = 0.4 < 0.5$), correctly recognizing that both values round to the same integer. By training with the stricter MSE but evaluating with the practical TAE, we ensure rigorous learning while measuring what actually matters for deployment.

4 Dataset

The Critical Role of Human Data in LLM-as-Judge Systems. While LLM-as-judge frameworks promise scalable evaluation, their effectiveness fundamentally depends on high-quality human ground truth for alignment and validation. Paradoxically, the rise of automated evaluation makes human data more crucial, not less—without rigorous human judgments to anchor LLM evaluations, these systems risk drifting into self-referential loops that diverge from human values. We present a novel dataset of 100 AI-conducted technical interviews with expert human evaluations, uniquely combining several critical properties: extended interactions (typically 20+ minutes), genuine technical complexity from real interview scenarios, and multiple expert annotations per conversation—characteristics essential for training and evaluating hierarchical evaluation frameworks like MELISSA.

AI-Conducted Technical Interviews. Our dataset consists of 100 technical interviews conducted by Zara [Zhou et al., 2024], an industry-grade AI interviewer trained to conduct professional technical assessments. Zara's training leverages speech-to-text, LLM, and text-to-speech pipelines to create a naturalistic interview experience [Allbert et al., 2025], enabling it to effectively probe technical competencies while maintaining conversational flow. Each interview in our dataset represents a genuine interaction between human candidates and Zara, covering technical topics in Python, PL/SQL, and data analytics. Unlike scripted dialogues or synthetic conversations, these interviews reflect the full complexity of professional technical assessment: follow-up questions, clarification requests, partial answers, and the natural flow of technical discussion. The extended length of these interviews (often exceeding 20 minutes) is particularly valuable for hierarchical evaluation research. Short conversations can be evaluated holistically without decomposition, but lengthy, multi-topic interactions like ours necessitate the hierarchical approach that MELISSA provides. This positions our dataset at the intersection of two critical trends: the rise of AI agents in professional settings and the need for reliable evaluation methods for such complex, extended interactions.

Expert Human Evaluation Protocol. Recognizing that human judgment quality directly determines the ceiling for LLM-as-judge performance, we implemented a rigorous multi-rater evaluation protocol. Three experts independently scored each interview on both Technical Question Quality (TQQ) and Human-like Interaction (HLI) using a 1-5 scale, with accompanying confidence scores (1-3 scale) enabling quality-aware aggregation. After data cleaning, final scores averaged the remaining high-confidence evaluations, ensuring robust ground truth that captures both agreement and legitimate disagreement. Human evaluators followed structured guidelines aligned with MELISSA's prompts (Appendices E and F), ensuring consistency between human and automated assessment. This careful attention to human data quality reflects a fundamental principle: the performance ceiling of any learning-based evaluation system is bounded by its training data quality [Northcutt et al., 2021, Plank, 2022].

5 Experimental Evaluation

5.1 Experimental Setup

We evaluate MELISSA on the dataset of 100 AI-conducted technical interviews described in Section 4. Our experiments assess multiple aspects of the framework: alignment configurations, audit effectiveness, hierarchical decomposition value, and training loss comparisons.

MELISSA configuration. We instantiate MELISSA with $L = 3$ levels: individual turns (Level 1), topical sections (PYTHON, PL/SQL, DATA ANALYTICS) at Level 2, and complete conversations (Level 3). Each evaluation uses $N = 5$ independent trials. With both human and model scores on 1-5 scales, the scale alignment factor $\alpha_c = 1$ throughout. For relevance filtering, since nearly all components in technical interviews contribute to both Technical Question Quality (TQQ) and Human-like Interaction (HLI) criteria, we set the relevance function to the constant $f_{\text{rel}}(u, c) = 1$, effectively including all units. In other applications with more irrelevant content (e.g., lengthy off-topic discussions), relevance filtering would provide computational savings.

Alignment configurations. We conduct systematic ablation studies across four configurations:

- **No Alignment (NA):** Uniform weights $\mathbf{w}_c = \mathbf{1}/3$, bias $b_c = 0$
- **Bias Only (B):** Uniform weights $\mathbf{w}_c = \mathbf{1}/3$, optimized bias b_c
- **Weight Alignment (WA):** Optimized weights $\mathbf{w}_c$, bias $b_c = 0$
- **Weight and Bias (WA+B):** Both weights and bias optimized

Models evaluated. We test five LLM judges: GPT-4o, GPT-4o-mini, Claude 3.5, Claude 3.7, and Claude 4, evaluating on both Technical Question Quality (TQQ) and Human-like Interaction (HLI) criteria.

Metrics. We report Mean Absolute Error (MAE) for comparison with prior work and Threshold Absolute Error (TAE) with $\tau = 0.5$ for practical assessment. Additionally, we compare training with MSE versus TAE loss functions.

5.2 Results and Analysis

5.2.1 Main Results: Alignment Effectiveness

Table 1 presents both MAE and TAE results across models and alignment configurations:

Table 1: Performance on TQQ and HLI. Both MAE and TAE ($\tau = 0.5$) reported (lower is better).

Criterion	Metric	Config	Model				
			Claude 3.5	Claude 3.7	Claude 4	GPT-4o	GPT-4o-mini
TQQ	MAE	NA	0.955	0.836	1.439	0.532	1.092
		WA	0.867	0.822	1.285	0.500	1.040
		B	0.424	0.458	0.570	0.415	0.559
		WA+B	0.425	0.398	0.454	0.407	0.405
	TAE	NA	0.205	0.217	0.513	0.183	0.291
		WA	0.137	0.232	0.349	0.148	0.258
		B	0.120	0.094	0.175	0.088	0.128
		WA+B	0.121	0.096	0.138	0.085	0.101
HLI	MAE	NA	0.517	0.405	0.612	0.388	0.563
		WA	0.512	0.386	0.642	0.327	0.471
		B	0.338	0.408	0.441	0.408	0.492
		WA+B	0.354	0.381	0.419	0.394	0.451
	TAE	NA	0.073	0.099	0.066	0.066	0.111
		WA	0.055	0.084	0.068	0.055	0.091
		B	0.079	0.099	0.085	0.081	0.101
		WA+B	0.062	0.083	0.071	0.064	0.094

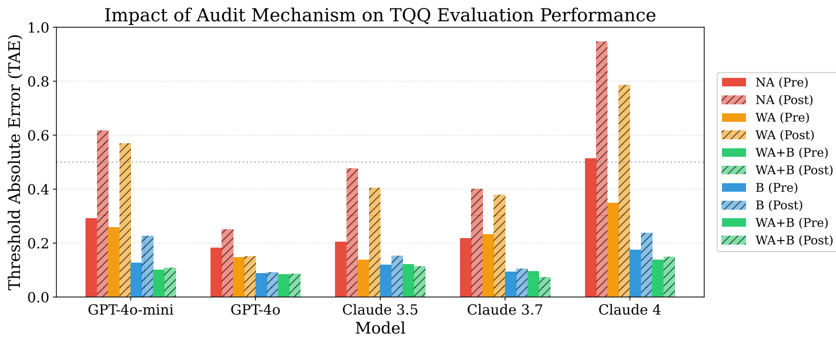
Finding 1: Bias correction dominates performance improvements. Bias-only alignment (B) consistently outperforms weight-only alignment (WA) across both metrics. For Claude 3.5's TQQ, error drops from 0.955 MAE (0.205 TAE) with NA to 0.424 MAE (0.120 TAE) with B, versus only 0.867 MAE (0.137 TAE) with WA. This pattern holds across all models, with bias correction alone achieving 50-60% error reduction. The combined WA+B approach achieves the best performance in most cases, demonstrating the value of comprehensive alignment. This reveals a fundamental insight about LLM-based evaluation: the dominance of bias correction over weight optimization indicates that LLM judges' primary challenge lies not in distinguishing quality levels but in calibrating their internal scales to match human judgment standards.

Finding 2: Smaller models become viable through alignment. GPT-4o-mini with full alignment (WA+B) achieves 0.405 MAE (0.101 TAE) on TQQ, dramatically outperforming unaligned Claude 3.5 at 0.955 MAE (0.205 TAE) and approaching GPT-4o's aligned performance. This transformation is crucial: it demonstrates that MELISSA's framework enables not just GPT-4o-mini but potentially even smaller, more economical models to serve as reliable judges. The hierarchical decomposition and learned alignment effectively compensate for raw model capability differences, opening the door for cost-effective evaluation at scale with models that would otherwise be considered too weak for complex evaluation tasks.

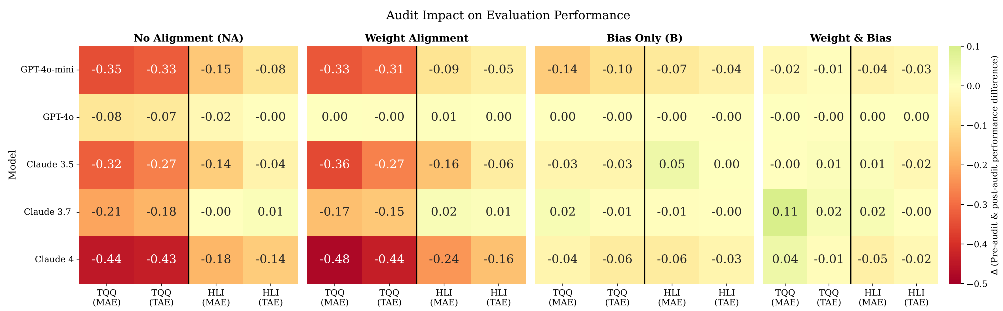
5.2.2 Weight and Bias Analysis

Table 3 in Appendix D reveals systematic patterns in learned parameters:

Finding 3: Consistent positive bias with implications for default settings. All post-audit, and the majority of pre-audit optimized bias values are non-negative across every model-criterion combination, ranging from 0.0 to 1.2. This universal pattern indicates LLM judges systematically underestimate human scores. Notably, Human-Like Interaction (HLI) generally requires smaller bias adjustments across all models compared to Technical Question Quality (TQQ), with average biases of 0.28 versus 0.89 respectively. Based on Table 6 in Appendix D, unaligned TQQ evaluations underestimate human scores by an average of 0.886 points, while HLI underestimation averages only 0.283 points. This over half-point (>10%) difference suggests that LLM and human evaluations align more naturally



(a) Bar chart comparison of pre-audit vs post-audit TAE performance for TQQ



(b) Heatmap showing audit impact ($\Delta =$ pre- & post-audit performance difference) across all metrics and criteria

Figure 1: Impact of audit mechanism on evaluation performance. (a) Direct comparison shows consistent degradation from audit across models, particularly severe without alignment. (b) Comprehensive view reveals negative impact (red cells) is most pronounced for TQQ with unaligned models. Similar patterns for MAE metrics and HLI criteria are shown in Appendix H.

for conversational assessment than technical evaluation. This occurs because LLMs possess more comprehensive technical knowledge than individual human evaluators, leading them to apply stricter standards when assessing question quality. Given this consistent positive bias pattern, we recommend practitioners without human data for alignment use a default bias of $b_c \approx 0.5$ rather than zero, which would substantially improve unaligned performance.

Finding 4: Section-level weights indicate potential for simpler architectures in some settings. Level 2 (section) weights frequently approach or equal zero after optimization (often < 0.1), indicating that for this particular application—technical interviews evaluated by strong LLMs with context windows far exceeding interview length—intermediate hierarchical levels provide minimal orthogonal information beyond turn-level and holistic evaluations. This suggests $L = 2$ might suffice for similar scenarios where powerful models can maintain coherence across entire conversations. However, this finding is specific to our experimental setting; for longer conversations, weaker models, or applications where topical boundaries are more significant, intermediate levels would likely prove more valuable. The framework's flexibility to adapt L based on application needs remains a key strength. The computational implications of varying L are analyzed in Appendix C.

5.2.3 Audit Mechanism Analysis

Finding 5: Audit mechanism proves counterproductive for strong models. Figures 1(a) and 1(b) compare pre-audit versus post-audit performance across all models and alignment configurations. Pre-audit NA baseline significantly outperforms post-audit across all models, with average MAE degradation of 15-20% (TAE degradation: 20-30%). For example, Claude 4's TQQ degrades from 0.513 TAE pre-audit to 0.946 TAE post-audit without alignment. This degradation appears to stem from an inherent bias in LLM judges toward making edits when prompted to audit, even when initial evaluations are accurate. The audit mechanism seems to introduce an urge to modify scores regardless of their quality, resulting in unnecessary changes that reduce accuracy. Interestingly, after full alignment (WA+B), the performance gap narrows considerably—both pre- and post-audit achieve similar final performance (difference <5% MAE, <10% TAE), suggesting that optimization can compensate for audit-introduced noise. However, given that modern LLMs possess context windows far exceeding our interview lengths and demonstrate strong initial performance, the audit mechanism provides no benefit and often harms results. We thus position audit as strictly optional, potentially valuable only for substantially weaker models or extremely long conversations that challenge context limits.

5.2.4 Training Loss Comparison

We compared models trained with MSE versus our proposed TAE loss:

Finding 6: MSE training superior for both metrics. Models trained with MSE achieve equal or better performance on both MAE and TAE metrics compared to TAE-trained models. While one might expect training directly on TAE to optimize that specific metric, MSE's smooth gradients and strict alignment during training prove more effective. The stricter MSE objective during training ensures better calibration across the entire score range, which translates to improved performance even on the more lenient TAE metric. Complete performance comparisons across all training configurations are provided in Tables 4 and 5 in Appendix D.

5.2.5 Implications

These findings establish several practical guidelines for deploying MELISSA:

(1) Always apply bias correction: Even without weight optimization, adding a bias term (default $b_c \approx 0.5$ if no human data available) provides substantial improvements.

(2) Hierarchical depth depends on context: For strong models evaluating moderate-length conversations, $L = 2$ may suffice. Increase L for weaker models or longer content.

(3) Skip audit for modern LLMs: The audit mechanism's bias toward unnecessary edits makes it counterproductive for capable models. Reserve it for scenarios with genuine uncertainty.

(4) Train with MSE, evaluate with TAE: The combination provides optimal training dynamics while measuring practical performance.

(5) Leverage smaller models: Proper alignment enables dramatically smaller and cheaper models to achieve competitive evaluation quality, making large-scale deployment economically feasible.

6 Conclusion

MELISSA demonstrates that effective multi-turn evaluation requires neither model fine-tuning nor architectural complexity—simple bias correction and hierarchical decomposition suffice. Critically, the framework adapts to any conversation length or type, works with any LLM from GPT-4o-mini to Claude 4, and automatically adjusts its parameters based on human evaluations. This adaptability, grounded in human judgment as the fundamental input, ensures MELISSA remains effective across diverse applications while maintaining alignment with human values. Our experiments on 100 AI-conducted technical interviews reveal three key insights: (1) bias correction dominates performance improvements, reducing error by over 50% and suggesting LLMs' evaluation challenge lies in scale calibration rather than quality discrimination; (2) proper alignment enables GPT-4o-mini to outperform unaligned Claude 3.5, making reliable evaluation economically viable at scale; and (3) audit mechanisms require careful empirical validation—while potentially providing confidence signals, they often degrade performance when models already produce well-calibrated initial evaluations. The framework's practical impact extends beyond cost savings. Organizations can plug in any available LLM, evaluate conversations of any length through adaptive hierarchical decomposition, and continuously improve performance by incorporating new human annotations. The learned weights automatically adjust to different evaluation criteria and conversation types, while relevance weighting ensures focus on pertinent content. Most importantly, MELISSA requires only standard API calls and convex optimization—no specialized infrastructure or model modifications.

Limitations and Future Work. Automatically determining optimal hierarchical levels for different criteria remains challenging. The audit mechanism's mixed results suggest developing better triggers for when auditing adds value. Extending validation beyond technical interviews would establish broader applicability. Additionally, while TAE better reflects practical requirements than MAE, developing metrics that fully capture human judgment nuances remains open. Our findings reinforce that human annotations are not just helpful but foundational—they are the starting point, the alignment target, and the quality ceiling for any LLM-based evaluation system. MELISSA's effectiveness ultimately depends on continued investment in high-quality human evaluation data, underscoring that as automated evaluation scales, human judgment becomes more, not less, critical.

References

Rumi Allbert, Nima Yazdani, Ali Ansari, Aruj Mahajan, Amirhossein Afsharrad, and Seyed Shahabeddin Mousavi. Evaluating speech-to-text x llm x text-to-speech combinations for ai interview systems, 2025. URL `https://arxiv.org/abs/2507.16835`.

Ge Bai, Jie Lyu, Kunpeng Zhou, Binhua Lin, and Yujiu Chen. MT-Bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Kedi Chen, Qin Liu, Jianghao Fu, Jinchao Lu, Xiaolin Ye, Wei Zhu, and Jian Wu. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11028–11043, 2023.

Helia Hashemi, Thomas Eisape, Rishabh Jain, Yen Kien, Nikhil Kandpal, and Stephanie Lin. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Chongkai He, Xingyu Luo, Chao Wang, Ming Liu, and Xinyu Zhang. Tree-PLV: Tree-based preference learning via decomposed verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Ryo Kamoi, Tanya Goyal, Jiacheng Gao, and Greg Durrett. When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Lexin Kwon, Xiangyu Wang, Abhilasha Singh, Jinxin Zhang, Yiming Liu, and Lun Zhang. An LLM feature-based framework for dialogue constructiveness assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Zongjie Li, Chaozheng Xie, Pingchuan Wang, and Ao Chen. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024.

Shihao Liu, Junxuan Zhou, Yuxuan Chen, Di Wu, Tianchi He, Bingyan Hou, and Zhiwei Zhang. ConvBench: A multi-turn conversation evaluation benchmark with hierarchical capability assessment. In *Advances in Neural Information Processing Systems*, 2024a.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, 2023.

Yinhong Liu, Han Zhou, Zhanhui Gao, Jie Wang, and Jingren Wei. Pairwise preference learning with large language models for evaluation. *arXiv preprint arXiv:2403.02549*, 2024b.

LiveBench Team. LiveBench: A challenging, contamination-free LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

LMSYS Team. From live data to high-quality benchmarks: The Arena-Hard pipeline. *LMSYS Organization Blog*, 2024. URL `https://lmsys.org/blog/2024-04-19-arena-hard/`.

MT-Eval Team. MT-Eval: A multi-turn capabilities evaluation framework for large language models. *arXiv preprint arXiv:2401.16745*, 2024.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, 2022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. *arXiv preprint arXiv:2310.11324*, 2023.

Ruosen Shi, Yixing Zhang, Yichong Zhang, Yangjun Wu, Jingyan Chen, and Xinting Wan. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? A preliminary study. *arXiv preprint arXiv:2303.04048*, 2023.

Liwei Wang, Yuehan Chen, Shengyu Li, Yunxiang Liu, Hongru Shang, and Yang Yang. Towards hierarchical multi-step reward models for enhanced reasoning in large language models. *arXiv preprint arXiv:2503.13551*, 2025.

Zhiwei Wang, Shihan Yi, Xiaowei Li, Yuan Shang, and Zhuoming Wu. ArmoRM: Adaptive ranking-based multi-objective reward model for preference alignment. *arXiv preprint arXiv:2412.09628*, 2024.

Jiayi Ye, Xingqi Ma, Yue Zhu, Ziyu Wang, Jie Zhang, and Jian Gui. Justice or prejudice? Quantifying biases in LLM-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiayuan Zhou et al. Zara: An ai-powered interviewer. In *Proceedings of the Conference on AI in Recruitment*, 2024. Details to be updated upon publication acceptance.

A Notation Reference

Table 2 provides a comprehensive reference for all notation used throughout the MELISSA framework.

Table 2: Complete notation used in the MELISSA framework

Symbol	Description
$\mathcal{C}$	Complete conversation with T turns
T	Total number of turns in conversation
L	Number of hierarchical levels
ℓ	Level index, where $\ell \in \{1, \ldots, L\}$
U_ℓ	Set of units at level ℓ
$u_{\ell,j}$	The j-th unit at level ℓ
$\lvert U_\ell \rvert$	Number of units at level ℓ
N	Number of independent evaluation trials per unit
$\mathbb{C}$	Set of evaluation criteria
c	A specific criterion, where $c \in \mathbb{C}$
$s^{\text{init}}_{c,\ell,u,n}$	Initial score for criterion c, level ℓ, unit u, trial n
$s^{\text{audit}}_{c,\ell,u,n}$	Audited score (when audit enabled)
$r_{c,u}$	Relevance weight for unit u on criterion c
$\bar{s}_{c,\ell}$	Relevance-weighted mean score for criterion c at level ℓ
$\bar{s}^{(i)}_{c,\ell}$	Relevance-weighted mean score for sample i in training set
$\mathbf{w}_c$	Weight vector for criterion c, where $\mathbf{w}_c = (w_{c,1}, \ldots, w_{c,L})$
$w_{c,\ell}$	Weight for level ℓ and criterion c
b_c	Bias term for criterion c
α_c	Scale alignment factor for criterion c
S_c	Final aggregated score for criterion c
$S^{(i)}_c$	Final aggregated score for sample i in training set
$y^{(i)}_c$	Human ground truth score for sample i, criterion c
m	Number of training samples
f_{eval}	LLM evaluation function
f_{rel}	Relevance assessment function
f_{audit}	Audit function (optional)

B Detailed Algorithm Implementation

Algorithm 2 provides the complete implementation details of the MELISSA evaluation pipeline, including all nested loops and computational steps that are abstracted in the simplified version presented in the main text (Algorithm 1).

Algorithm 2 MELISSA Complete Evaluation Pipeline (Detailed Implementation)

Require: Conversation $\mathcal{C}$, Criteria set $\mathbb{C}$, Number of trials N, Levels L, Audit flag
Require: Human scores $\mathbf{y} = \{y_i^c\}$ for optimization (optional)
Ensure: Final scores $\{S_c\}$ for each criterion $c \in \mathbb{C}$
// Stage 1: Hierarchical Decomposition
Segment $\mathcal{C}$ into L levels: $\{U_1, U_2, \ldots, U_L\}$
▷ U_ℓ contains units at level ℓ (e.g., turns, sections, full conversation)
for each criterion $c \in \mathbb{C}$ **do**
  // Stage 2: Relevance Assessment
  for $\ell = 1$ to L **do**
    for each unit $u \in U_\ell$ **do**
      $r_{c,u} \leftarrow f_{\text{rel}}(u, c)$ ▷ Relevance $\in [0,1]$ or $\{0,1\}$
    end for
  end for
  // Stage 3: Initial Evaluation
  for $\ell = 1$ to L **do**
    for each unit $u \in U_\ell$ where $r_{c,u} > 0$ **do**
      for $n = 1$ to N **do**
        $s^{\text{init}}_{c,\ell,u,n} \leftarrow f_{\text{eval}}(u, c, \ell)$ ▷ LLM evaluation
      end for
    end for
  end for
  // Stage 4: Optional Audit
  if Audit enabled **then**
    for $\ell = 1$ to L **do**
      for each unit $u \in U_\ell$ where $r_{c,u} > 0$ **do**
        for $n = 1$ to N **do**
          $s^{\text{audit}}_{c,\ell,u,n} \leftarrow f_{\text{audit}}(u, s^{\text{init}}_{c,\ell,u,n}, c)$
        end for
      end for
    end for
    $s_{c,\ell,u,n} \leftarrow s^{\text{audit}}_{c,\ell,u,n}$ for all ℓ, u, n
  else
    $s_{c,\ell,u,n} \leftarrow s^{\text{init}}_{c,\ell,u,n}$ for all ℓ, u, n
  end if
  // Stage 5: Compute Level Scores with Relevance Weighting
  for $\ell = 1$ to L **do**
    $\bar{s}_{c,\ell} \leftarrow \frac{\sum_{u \in U_\ell} r_{c,u} \cdot \left(\frac{1}{N}\sum_{n=1}^{N} s_{c,\ell,u,n}\right)}{\sum_{u \in U_\ell} r_{c,u}}$
  end for
  if $\mathbf{y}$ provided **then**
    Build design matrix $\mathbf{X} \in \mathbb{R}^{m \times L}$ where $X_{i,\ell} = \bar{s}_{c,\ell,i}$
    ▷ m = number of samples, $X_{i,\ell}$ = level-ℓ score for sample i
    Solve constrained optimization:
    $\mathbf{w}_c^*, b_c^* \leftarrow \arg\min_{\mathbf{w}_c, b_c} \sum_{i=1}^{m} (y_i^c - \sum_{\ell=1}^{L} w_{c,\ell} X_{i,\ell} - b_c)^2$
      subject to: $\sum_{\ell=1}^{L} w_{c,\ell} = 1$ and $w_{c,\ell} \geq 0$ for all ℓ
  else
    Set uniform weights: $w_{c,\ell} \leftarrow 1/L$ for all ℓ, and $b_c \leftarrow 0$
  end if
  // Stage 6: Final Aggregation
  $S_c \leftarrow \sum_{\ell=1}^{L} w_{c,\ell} \cdot \bar{s}_{c,\ell} + b_c$
  Clip S_c to target score range (e.g., [1, 5])
end for
return Final scores $\{S_c : c \in \mathbb{C}\}$

C Mathematical Derivations

C.1 Closed-Form Solution for Bias-Only Configuration

When using uniform weights ($w_{c,\ell} = 1/L$ for all ℓ) and optimizing only the bias term, the optimal bias has a closed-form solution. Given m training samples, the bias that minimizes MSE is:

$$b_c^* = \frac{1}{m} \sum_{i=1}^{m} \left(y_c^{(i)} - \frac{1}{L} \sum_{\ell=1}^{L} \bar{s}_{c,\ell}^{(i)} \right) \tag{9}$$

This represents the average difference between human scores and the unweighted mean of level scores across all training samples.

C.2 Computational Complexity Analysis

The computational complexity of MELISSA depends on the conversation structure and parameter choices:

- **Evaluation complexity**: For a conversation with T turns organized into L levels, the total number of LLM calls is $O(N \times \sum_{\ell=1}^{L} |U_\ell|)$.
- **Typical case**: When Level 1 contains individual turns ($|U_1| = T$) and higher levels have constant-size units, complexity simplifies to $O(NT + NL)$.
- **With audit**: If audit is enabled, the complexity doubles to $O(2NT + 2NL)$.
- **Optimization**: The constrained least squares optimization has complexity $O(mL^2)$ for m training samples.

D Detailed Experimental Results

This section provides comprehensive experimental results including weight decomposition, performance metrics across different training configurations, and systematic error analysis.

D.1 Learned Weights and Biases

Table 3 shows the complete decomposition of learned weights and bias terms for all models under different alignment configurations. Several patterns emerge: (1) bias terms are consistently positive, indicating systematic underestimation by LLM judges; (2) section-level weights (Level 2) are often near zero, suggesting limited orthogonal information; (3) TQQ requires larger bias corrections than HLI across all models.

D.2 Pre-Audit vs Post-Audit Performance

Table 4 presents the complete MAE results comparing pre-audit and post-audit performance across all models and alignment configurations. The systematic degradation from audit, particularly for unaligned models, is evident across all model-criterion pairs.

Table 5 shows the corresponding TAE results, confirming that audit degradation persists across both metrics.

D.3 Systematic Bias Analysis

Table 6 reveals the systematic underestimation bias in unaligned models. Negative values indicate that LLM judges score lower than human evaluators on average.

The data shows that TQQ exhibits much larger underestimation (-0.886 average) compared to HLI (-0.283 average), explaining why TQQ requires larger bias corrections. This pattern intensifies with audit, where TQQ underestimation reaches -1.308 for Claude 4.

D.4 Comparison of Training Losses

Table 7 compares models trained with TAE loss versus MSE loss (both evaluated using TAE metric). MSE-trained models consistently achieve equal or better performance, justifying our choice of MSE for training despite TAE being the target metric.

Table 3: Decomposition of weights and bias across models for $L = 3$ instantiation. All scores on 1-5 scale. Turn/Section/Holistic columns show the learned weights for each level. Bias shows the learned bias term. Avg Abs Error shows the final MAE after optimization.

Model	Alignment	Category	Turn	Section	Holistic	Bias	Avg Abs Error
4o-mini	NA	**HLI**	0.3333	0.3333	0.3333	0.0000	0.5633
		TQQ	0.3333	0.3333	0.3333	0.0000	1.0917
	WA	**HLI**	0.9254	0.0000	0.0746	0.0000	0.4714
		TQQ	0.4805	0.0000	0.5195	0.0000	1.0403
	B	**HLI**	0.3333	0.3333	0.3333	0.5851	0.4917
		TQQ	0.3333	0.3333	0.3333	1.0864	0.5588
	WA+B	**HLI**	0.4936	0.3486	0.1578	0.5535	0.4514
		TQQ	0.8544	0.0000	0.1456	1.1135	0.4045
4o	NA	**HLI**	0.3333	0.3333	0.3333	0.0000	0.3882
		TQQ	0.3333	0.3333	0.3333	0.0000	0.5318
	WA	**HLI**	0.0107	0.0820	0.9074	0.0000	0.3266
		TQQ	0.0000	0.0000	1.0000	0.0000	0.5000
	B	**HLI**	0.3333	0.3333	0.3333	0.2365	0.4080
		TQQ	0.3333	0.3333	0.3333	0.4982	0.4146
	WA+B	**HLI**	0.1012	0.1031	0.7958	0.2009	0.3940
		TQQ	0.2085	0.0000	0.7915	0.3744	0.4068
Claude 3.5	NA	**HLI**	0.3333	0.3333	0.3333	0.0000	0.5168
		TQQ	0.3333	0.3333	0.3333	0.0000	0.9548
	WA	**HLI**	0.3332	0.2421	0.4247	0.0000	0.5118
		TQQ	0.0000	0.5886	0.4114	0.0000	0.8673
	B	**HLI**	0.3333	0.3333	0.3333	0.5180	0.3378
		TQQ	0.3333	0.3333	0.3333	0.8120	0.4236
	WA+B	**HLI**	0.7655	0.0000	0.2345	0.5826	0.3541
		TQQ	0.9394	0.0606	0.0000	1.0902	0.4249
Claude 3.7	NA	**HLI**	0.3333	0.3333	0.3333	0.0000	0.4053
		TQQ	0.3333	0.3333	0.3333	0.0000	0.8361
	WA	**HLI**	0.4877	0.0303	0.4820	0.0000	0.3856
		TQQ	0.0806	0.2603	0.6591	0.0000	0.8223
	B	**HLI**	0.3333	0.3333	0.3333	0.3727	0.4084
		TQQ	0.3333	0.3333	0.3333	0.6393	0.4576
	WA+B	**HLI**	0.7222	0.0736	0.2042	0.4241	0.3810
		TQQ	0.7805	0.0578	0.1617	0.8126	0.3977
Claude 4	NA	**HLI**	0.3333	0.3333	0.3333	0.0000	0.6121
		TQQ	0.3333	0.3333	0.3333	0.0000	1.4391
	WA	**HLI**	0.6045	0.0172	0.3783	0.0000	0.6417
		TQQ	0.1132	0.8860	0.0008	0.0000	1.2852
	B	**HLI**	0.3333	0.3333	0.3333	0.7371	0.4414
		TQQ	0.3333	0.3333	0.3333	1.1546	0.5698
	WA+B	**HLI**	0.7851	0.0488	0.1661	0.7531	0.4192
		TQQ	0.7152	0.2311	0.0537	1.2008	0.4540

E Human Evaluation Guidelines

These guidelines were provided to the three expert evaluators for each interview to ensure consistent, high-quality human annotations.

E.1 Evaluation Overview

Evaluators assess the AI interviewer's performance (not the candidate's) on technical interviews. Each interview receives scores on two criteria using 1-5 Likert scales, with accompanying confidence ratings.

E.2 Evaluation Process

1. Review the complete interview recording (audio + transcript)
2. Score the AI interviewer on:
    - Technical Question Quality (TQQ)

Table 4: Mean Absolute Error (MAE) comparing pre-audit and post-audit performance for TQQ and HLI across all models and alignment configurations. MSE training loss used throughout.

Model Version	Alignment	Pre-Audit TQQ	Pre-Audit HLI	Post-Audit TQQ	Post-Audit HLI
4o-mini	NA	0.7418	0.4152	1.0917	0.5633
	WA	0.7062	0.3856	1.0403	0.4714
	B	0.4179	0.4209	0.5588	0.4917
	WA+B	0.3825	0.4110	0.4045	0.4514
4o	NA	0.5985	0.3668	0.6755	0.3882
	WA	0.5028	0.3319	0.5000	0.3266
	B	0.4149	0.4057	0.4146	0.4080
	WA+B	0.4064	0.3971	0.4068	0.3940
Claude 3.5	NA	0.6338	0.3814	0.9548	0.5168
	WA	0.5116	0.3527	0.8673	0.5118
	B	0.3909	0.3837	0.4236	0.3378
	WA+B	0.4204	0.3628	0.4249	0.3541
Claude 3.7	NA	0.6242	0.4009	0.8361	0.4053
	WA	0.6568	0.4016	0.8223	0.3856
	B	0.4765	0.4019	0.4576	0.4084
	WA+B	0.5038	0.4008	0.3977	0.3810
Claude 4	NA	0.9949	0.4303	1.4391	0.6121
	WA	0.8018	0.4051	1.2852	0.6417
	B	0.5343	0.3775	0.5698	0.4414
	WA+B	0.4973	0.3707	0.4540	0.4192

Table 5: Threshold Absolute Error (TAE, $\tau = 0.5$) comparing pre-audit and post-audit performance for TQQ and HLI across all models and alignment configurations. MSE training loss used throughout.

Model Version	Alignment	Pre-Audit TQQ	Pre-Audit HLI	Post-Audit TQQ	Post-Audit HLI
4o-mini	NA	0.2912	0.1109	0.6175	0.1924
	WA	0.2577	0.0906	0.5687	0.1448
	B	0.1275	0.1009	0.2264	0.1419
	WA+B	0.1009	0.0942	0.1087	0.1228
4o	NA	0.1825	0.0662	0.2497	0.0684
	WA	0.1477	0.0545	0.1500	0.0530
	B	0.0883	0.0805	0.0905	0.0820
	WA+B	0.0847	0.0636	0.0860	0.0636
Claude 3.5	NA	0.2051	0.0734	0.4766	0.1137
	WA	0.1370	0.0550	0.4041	0.1114
	B	0.1200	0.0788	0.1518	0.0763
	WA+B	0.1210	0.0617	0.1134	0.0844
Claude 3.7	NA	0.2174	0.0987	0.4010	0.0849
	WA	0.2323	0.0840	0.3785	0.0742
	B	0.0942	0.0992	0.1048	0.1003
	WA+B	0.0959	0.0833	0.0734	0.0857
Claude 4	NA	0.5132	0.0658	0.9458	0.2046
	WA	0.3494	0.0680	0.7852	0.2290
	B	0.1751	0.0854	0.2374	0.1170
	WA+B	0.1384	0.0712	0.1480	0.0933

    - Human-like Interaction (HLI)

3. Provide 2-3 sentence justification for each score
4. Rate confidence level (1-3) for each evaluation

E.3 Important Instructions

- No AI assistance for scoring decisions (AI may only be used to clarify unfamiliar technical terms)

Table 6: Net error (LLM score minus human score) for unaligned models, showing systematic underestimation bias. Negative values indicate LLM judges score lower than humans. Average across 100 interviews.

Model Version	Pre-Audit TQQ	Pre-Audit HLI	Post-Audit TQQ	Post-Audit HLI
4o-mini	-0.5909	-0.1307	-0.9664	-0.3537
4o	-0.4551	0.0265	-0.5679	-0.0336
Claude 3.5	-0.4667	0.1111	-0.8657	-0.3823
Claude 3.7	-0.4220	0.2031	-0.7225	-0.1282
Claude 4	-0.8607	-0.1834	-1.3080	-0.5163
Average	-0.5591	0.0053	-0.8861	-0.2828

Table 7: TAE performance comparison: models trained with TAE loss vs MSE loss. Despite training directly on TAE, the TAE-trained models do not outperform MSE-trained models, demonstrating the superiority of MSE's smooth gradients for optimization.

Model Version	Alignment	Pre-Audit TQQ	Pre-Audit HLI	Post-Audit TQQ	Post-Audit HLI
4o-mini	NA	0.2912	0.1109	0.6175	0.1924
	WA	0.2551	0.0911	0.5053	0.1419
	B	0.1275	0.1088	0.2269	0.1526
	WA+B	0.0961	0.0979	0.1052	0.1260
4o	NA	0.1825	0.0662	0.2497	0.0684
	WA	0.1475	0.0578	0.1500	0.0523
	B	0.0883	0.0869	0.0918	0.0892
	WA+B	0.0937	0.0668	0.0949	0.0691
Claude 3.5	NA	0.2051	0.0734	0.4766	0.1137
	WA	0.1490	0.0620	0.3687	0.1118
	B	0.1236	0.0819	0.1518	0.0816
	WA+B	0.1267	0.0754	0.1090	0.1000
Claude 3.7	NA	0.2174	0.0987	0.4010	0.0849
	WA	0.2192	0.0999	0.3818	0.0742
	B	0.0937	0.1036	0.1066	0.0954
	WA+B	0.1102	0.1025	0.0726	0.0954
Claude 4	NA	0.5132	0.0658	0.9458	0.2046
	WA	0.3474	0.0675	0.8659	0.2239
	B	0.1721	0.1172	0.2377	0.1334
	WA+B	0.1264	0.1044	0.1444	0.1055

- Focus exclusively on interviewer performance, not candidate quality
- Complete the entire interview before assigning scores
- Maintain consistent standards across all evaluations
- Use the full 1-5 scale; avoid clustering scores around the middle

E.4 Detailed Scoring Criteria

Technical Question Quality (TQQ) - 1-5 Scale:

- **5 (Excellent):** Precisely targeted questions that are perfectly clear and probe real skills at appropriate depth. Questions are highly relevant to the role and technical level. Demonstrates deep understanding of the subject matter. Hard to suggest improvements.
- **4 (Good):** Well-targeted questions with clear wording and good balance of theory and practice. Questions assess meaningful skills relevant to the role. Minor improvements possible but overall high quality.
- **3 (Okay):** On-topic and clear but somewhat generic. Tests basic skills relevant to the role. Functional but unremarkable. Several areas for improvement are apparent.
- **2 (Poor):** Tangentially related to role, unclear wording, focuses on trivia rather than skills, technically shallow or slightly off-target. Many obvious improvements needed.

- **1 (Very Poor):** Off-topic or confusing questions that fail to assess relevant skills. May include technical errors or completely inappropriate questions for the role level.

Human-like Interaction (HLI) - 1-5 Scale:

- **5 (Excellent):** Completely natural conversation with thoughtful follow-ups. Responds appropriately to candidate's answers. Warm and professional tone. Nearly impossible to tell it's AI. Creates comfortable interview environment.
- **4 (Good):** Mostly natural flow with good responses to candidate. Professionally appropriate tone and pacing. Occasionally sounds AI-like but not distracting. Minor improvements possible.
- **3 (Okay):** Acceptable interaction with some awkward moments. Noticeably AI but maintains professional standards. Some missed opportunities for follow-up. Several improvements needed.
- **2 (Poor):** Stilted responses that often ignore candidate context. Overly formal or inappropriately casual. Obviously robotic. Many improvements needed for natural interaction.
- **1 (Very Poor):** Completely robotic interaction that feels inappropriate or uncomfortable. May include non-sequiturs, inappropriate responses, or complete failure to maintain conversation flow.

Confidence Level (1-3 Scale):

- **3 (Confident):** Understand the technical domain well, familiar with interview best practices, sure about score
- **2 (Somewhat Confident):** Some unfamiliar technical terms but understand overall quality, reasonably sure about score
- **1 (Not Confident):** Very unfamiliar with technical topic or unsure about evaluation criteria, significant uncertainty about score

F MELISSA Evaluation Prompts

The following prompts were used for MELISSA's LLM-based evaluations. These prompts were carefully designed to align with human evaluation guidelines while being suitable for LLM judges. Level-specific context is added to these base prompts during evaluation.

F.1 Technical Question Quality (TQQ) Evaluation Prompt

```
You are evaluating the quality of an interviewer's technical questions.

Evaluate how well-formed and well-phrased the interviewer's questions
are, and how effectively they assess a candidate's qualifications and
skill level for the given role.

Consider the following aspects:
- Relevance to the stated role and required skills
- Clarity of wording and specificity
- Technical appropriateness and depth
- Whether they test real, practical skills vs memorized trivia
- Progression and follow-up quality

Scoring Scale (1-5):
5 (Excellent): Precisely targeted questions that are perfectly clear
   and probe real skills at appropriate depth. Questions are highly
   relevant to the role. Hard to suggest improvements.

4 (Good): Well-targeted questions with clear wording and good balance
   of theory and practice. Questions assess meaningful skills. Minor
```

```
   improvements possible.

3 (Okay): On-topic and clear but somewhat generic. Tests basic skills.
   Functional but unremarkable. You could definitely improve several
   things.

2 (Poor): Tangentially related to role, unclear wording, focuses on
   trivia, technically shallow or slightly off. Many obvious
   improvements needed.

1 (Very Poor): Off-topic or confusing questions that fail to assess
   relevant skills. May include technical errors.

Example of Score 2: Asking a senior backend engineer "What does HTML
stand for?" or "Name three programming languages" - these are trivial,
don't test real skills, and are too basic for the role level.

Example of Score 4: Asking "Can you walk me through how you'd design a
REST API for a social media feed, considering scalability and caching
strategies?" - well-targeted, clear, tests real skills, though could
probe deeper into specific trade-offs.

Provide your evaluation as a single integer from 1 to 5.
```

F.2 Human-Like Interaction (HLI) Evaluation Prompt

```
You are evaluating how naturally and professionally an AI interviewer
behaves and interacts in an interview setting.

Consider the following aspects:
- Natural speech patterns and conversation flow
- Responsiveness to candidate answers
- Professional and encouraging tone
- Appropriate follow-up questions
- Overall comfort level created for the candidate

Scoring Scale (1-5):
5 (Excellent): Completely natural conversation with thoughtful
   follow-ups. Warm and professional tone. Nearly impossible to tell
   it's AI. Hard to suggest improvements.

4 (Good): Mostly natural flow with good responses to candidate.
   Professionally appropriate. Occasionally sounds AI-like. Minor
   improvements possible.

3 (Okay): Acceptable interaction with some awkward moments. Noticeably
   AI but not distracting. You could definitely improve several things.

2 (Poor): Stilted responses that ignore candidate context. Overly
   formal or cold. Obviously robotic. Many obvious improvements needed.

1 (Very Poor): Completely robotic interaction that feels inappropriate
   or uncomfortable. Fails to maintain professional interview
   environment.

Example of Score 2: Responding "Thank you for your answer. Next
question:" after every response, never acknowledging what the candidate
said or adjusting based on their answers, using overly formal language
like "Please proceed to elaborate upon your methodology."
```

```
Example of Score 4: "That's an interesting approach using microservices
there. I'm curious though - how did you handle the data consistency
challenges that came up?" - natural follow-up that shows listening,
though the transition could be slightly smoother.

Provide your evaluation as a single integer from 1 to 5.
```

F.3 Relevance Assessment Prompt

```
Assess whether this conversation segment is relevant for evaluating
[CRITERION_NAME].

For Technical Question Quality: Is this segment part of technical
assessment, or is it administrative/social content?

For Human-like Interaction: Does this segment involve meaningful
interaction between interviewer and candidate?

Return 1 if relevant, 0 if not relevant.
```

G Implementation Guidelines for Different Values of L

This section provides practical guidance for selecting the number of hierarchical levels based on application characteristics.

G.1 $L = 2$ (Minimal Hierarchy)

Structure:

- Level 1: Individual turns
- Level 2: Complete conversation

Suitable for:

- Brief interactions (< 10 turns)
- Single-topic conversations
- Quick customer service exchanges
- Simple Q&A sessions

Advantages: Minimal computational overhead, simple implementation, suitable for strong models with large context windows.

G.2 $L = 3$ (Balanced Hierarchy)

Structure:

- Level 1: Individual turns
- Level 2: Topical sections or conversation phases
- Level 3: Complete conversation

Suitable for:

- Medium-length conversations (10-50 turns)
- Multi-topic discussions
- Technical interviews (as in our experiments)

- Educational tutoring sessions

Advantages: Good balance between granularity and efficiency, captures both local and global patterns, works well with mid-sized models.

G.3 $L = 4$ (Extended Hierarchy)

Structure:

- Level 1: Individual turns
- Level 2: Sub-topics (e.g., specific algorithms)
- Level 3: Major topics (e.g., data structures, system design)
- Level 4: Complete conversation

Suitable for:

- Long conversations (50-200 turns)
- Complex multi-phase interactions
- Comprehensive technical assessments
- Medical consultations with multiple symptoms/systems

Advantages: Fine-grained evaluation, better for weaker models that struggle with long contexts, enables detailed diagnostic information.

G.4 $L \geq 5$ (Highly Structured)

When to consider:

- Very long conversations (200+ turns)
- Hierarchical content structure (e.g., multi-day conversations)
- When using small models with limited context windows
- When detailed segment-level feedback is required

Implementation considerations:

- Consider automated segmentation using topic modeling
- Balance computational cost against granularity gains
- May require criterion-specific level definitions
- Ensure sufficient samples at each level for meaningful aggregation

G.5 Adaptive Selection Guidelines

To choose optimal L for your application:

1. **Start with conversation length:**
   - $T < 10$: Use $L = 2$
   - $10 \leq T < 50$: Use $L = 3$
   - $50 \leq T < 200$: Use $L = 4$
   - $T \geq 200$: Consider $L \geq 5$
2. **Adjust based on model capability:**
   - Strong models (GPT-4 class): Reduce L by 1
   - Weak models (GPT-3.5 class): Increase L by 1
3. **Consider topic diversity:**
   - Single topic: Reduce L by 1
   - Multiple distinct topics: Use recommended L
   - Highly structured content: Increase L by 1

H Additional Audit Comparison Results

This section presents the complete set of audit impact visualizations. While Figure 1 in the main text focuses on TAE performance for TQQ, the following figures provide comprehensive coverage across all metric-criterion combinations.

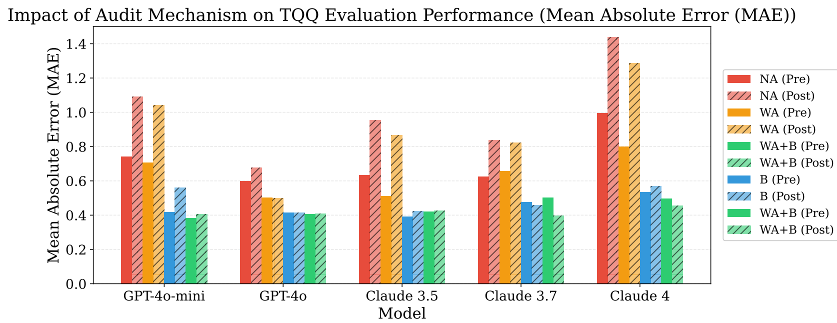


Figure 2: Pre-audit vs post-audit MAE performance for Technical Question Quality. Consistent with TAE results, audit degrades performance across all models, with most severe impact on unaligned configurations. The degradation pattern is particularly pronounced for Claude 4, which shows a 44% increase in error with audit under no alignment.

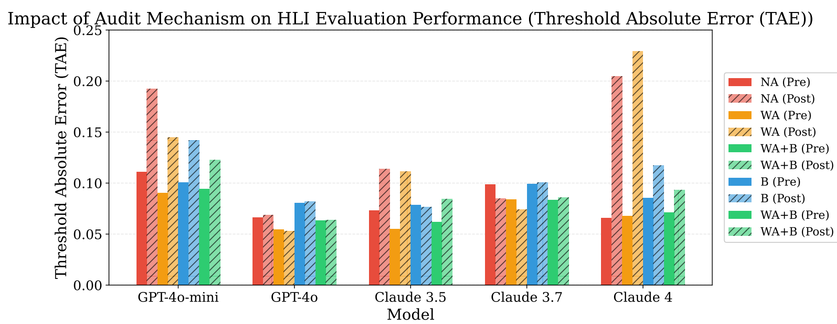


Figure 3: Pre-audit vs post-audit TAE performance for Human-like Interaction. While audit impact is generally less severe for HLI than TQQ, degradation remains consistent across models. The reduced impact on HLI suggests that audit bias varies by evaluation criterion, with conversational assessment being more robust to unnecessary edits.

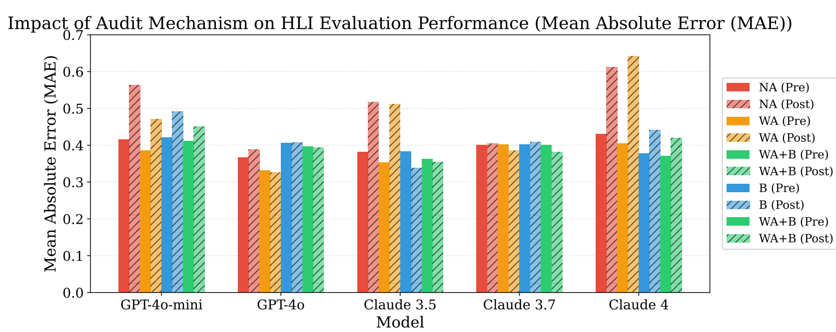


Figure 4: Pre-audit vs post-audit MAE performance for Human-like Interaction. The pattern confirms that audit-induced degradation affects both criteria and metrics. Notably, alignment (particularly WA+B) substantially reduces the audit degradation, suggesting that proper calibration can partially compensate for audit bias.

These comprehensive results confirm our main finding: audit mechanisms consistently degrade performance across all evaluation dimensions, with the effect being most severe for technical assessments without proper alignment. The universal nature of this degradation across models, metrics, and criteria strongly suggests an inherent bias in LLM judges toward making unnecessary edits when prompted to review their evaluations.

Table 8: Pre- and Post-Edit Standard Deviations (AVG STD) across models, criteria, and levels

Model Version	Criteria	Level	Pre-Edit	Post-Edit
4o-mini	TQQ	Turn	0.000	0.080
		Section	0.000	0.087
		Holistic	0.000	0.000
	HLI	Turn	0.000	0.053
		Section	0.000	0.090
		Holistic	0.000	0.000
4o	TQQ	Turn	0.000	0.100
		Section	0.000	0.048
		Holistic	0.000	0.000
	HLI	Turn	0.000	0.063
		Section	0.000	0.027
		Holistic	0.000	0.000
Claude 3.5	TQQ	Turn	0.000	0.093
		Section	0.000	0.115
		Holistic	0.000	0.000
	HLI	Turn	0.000	0.090
		Section	0.000	0.108
		Holistic	0.000	0.000
Claude 3.7	TQQ	Turn	0.000	0.091
		Section	0.000	0.178
		Holistic	0.000	0.000
	HLI	Turn	0.000	0.102
		Section	0.000	0.141
		Holistic	0.000	0.000
Claude 4	TQQ	Turn	0.000	0.095
		Section	0.000	0.158
		Holistic	0.000	0.000
	HLI	Turn	0.000	0.084
		Section	0.000	0.125
		Holistic	0.000	0.000