

Article

Mathematical Framework for Characterizing Emotional Individuality in Large Language Models: Temperature Control, Fuzzy Entropy, and Persona-Based Diversity Analysis

Naruki Shirahama ^{1,*} , Yuma Yoshimoto ², Naofumi Nakaya ³  and Satoshi Watanabe ⁴ 

¹ Department of Data Science, Faculty of Data Science, Shimonoseki City University, 2-1-1, Daigaku-cho, Shimonoseki 751-8510, Yamaguchi, Japan

² Department of Creative Engineering, National Institute of Technology (KOSEN), Kitakyushu College, 5-20-1 Shii, Kokuraminamiku, Kitakyushu 802-0985, Fukuoka, Japan; yoshimoto@kct.ac.jp

³ Department of Clinical Engineering, Graduate School of Medical Science, Juntendo University, 6-8-1 Hinode, Urayasu 279-0013, Chiba, Japan; n.nakaya.ac@juntendo.ac.jp

⁴ Department of Informatics, Faculty of Informatics, Shizuoka Institute of Science and Technology, 2200-2, Toyosawa, Fukuroi-shi 437-0032, Shizuoka, Japan; satoshi_watanabe@ieee.org

* Correspondence: nshirahama@ieee.org; Tel.: +81-83-242-8510

Abstract

Evaluating emotional understanding in Large Language Models (LLMs) is challenging because assessments are subjective, ambiguous, multidimensional, and sensitive to controllable generation parameters. We developed a unified mathematical framework for characterizing LLM “emotional individuality” that integrates softmax sampling–temperature control (the decoding-time temperature parameter exposed by the API and typically used to modulate output randomness during token generation), fuzzy set theory with Shannon-type fuzzy entropy, and persona-based cognitive diversity analysis. We evaluated 36 API-accessible LLMs from seven major vendors on Japanese literary texts, using four personas each assigned a sampling temperature ($T \in \{0.1, 0.4, 0.7, 0.9\}$), yielding 4227/4320 trial responses (97.8% coverage), of which 4067/4227 contained valid numeric emotion scores (96.2%). Temperature controllability varied approximately 25-fold ($\kappa_M \in [0.039, 0.982]$) with both positive and negative temperature–variance relationships across models. Because each sampling temperature is deterministically assigned to a persona in our design, κ_M should be interpreted as an operational temperature–variance association across persona conditions rather than an isolated causal temperature effect. The model-level mean fuzzy entropy ranged from approximately 0.40 to 0.66, and the numerical stability consistency scores ranged from approximately 0.548 to 0.780. We also observed text-dependent structure, including genre-specific variation in the Interest–Sadness relationship. For practitioners, the framework is most directly useful as a benchmark-design and model-screening template for structured emotion-scoring tasks; its empirical conclusions remain limited to the present Japanese literary, text-only setting.



Academic Editors: Ruowei Hung and Ling-Ju Hung

Received: 4 March 2026

Revised: 31 March 2026

Accepted: 1 April 2026

Published: 6 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: large language models; fuzzy entropy; softmax sampling temperature; emotional intelligence evaluation; mathematical modeling; persona-based assessment; cognitive diversity analysis; benchmarking; reproducibility

MSC: 03E72; 94A17

1. Introduction

The rapid advancement of Large Language Models (LLMs) has transformed natural language processing, with models demonstrating sophisticated capabilities across diverse cognitive tasks. As these systems increasingly mediate human–computer interactions in emotionally sensitive contexts, from mental health support to educational tutoring, their capacity to understand and appropriately respond to emotional content has become critically important. However, evaluating emotional understanding in LLMs presents a fundamental mathematical challenge: how to rigorously quantify and characterize capabilities that are inherently subjective, ambiguous, and multidimensional.

Current approaches for evaluating the emotional capabilities of LLMs predominantly rely on discrete classification schemes, benchmark datasets, and qualitative assessments. Although these methods have yielded valuable insights, they lack the mathematical rigor necessary to capture the continuous, probabilistic, and context-dependent nature of understanding emotions. The proliferation of diverse LLM architectures from multiple vendors, each with distinct training methodologies, alignment strategies, and parametric controls, further complicates the systematic comparisons and theoretical understanding of these models. A comprehensive mathematical framework is required to characterize model-dependent emotional assessments and quantify the uncertainty, consistency, and controllability inherent in LLM responses.

This study addresses this gap by developing an integrated mathematical framework for characterizing “emotional individuality” in LLMs, namely, the unique patterns of emotional interpretation and assessment that emerge from the interplay of model architecture, training data, and operational parameters. Our framework synthesizes three complementary mathematical approaches: (1) softmax sampling–temperature control, which provides a tractable mechanism for modulating response diversity and reliability; (2) fuzzy set theory with entropy measures, which rigorously quantifies the ambiguity inherent in emotional assessment; and (3) persona-based cognitive diversity analysis, which mathematically parameterizes variations in interpretative perspectives. Throughout this paper, temperature refers to the decoding-time temperature parameter exposed by the LLM API; in standard implementations, this parameter rescales token logits before softmax token sampling and does not denote any physical quantity.

1.1. Research Motivation and Background

Different LLMs systematically assign different emotion scores to the same text, which are sufficiently stable to be treated as model-specific *emotional individuality* rather than noise. Existing evaluations emphasize benchmark performance but provide limited mathematical characterization of (i) controllability via decoding parameters, (ii) ambiguity in continuous emotion-intensity judgments, and (iii) perspective dependence induced by prompt personas. Therefore, we ask the following questions: How does temperature modulate reliability and variability? How can ambiguity be quantified without imposing difficult category boundaries? How do persona-conditioned perspectives interact with these factors?

1.2. Related Work

Emotion understanding in LLMs is increasingly evaluated through benchmark suites, harder stress-test sets, and peer-reviewed surveys spanning emotional intelligence, empathy, and emotion recognition in conversation [1–6]. Likewise, persona conditioning and role-playing have been studied in peer-reviewed works, both as target behaviors and as evaluation objects [7–11], but they are usually analyzed for persona fidelity or dialogue quality rather than used as a quantitative axis in emotion-assessment protocols.

Temperature-controlled decoding remains a standard mechanism for modulating output diversity and stochasticity [12]. In mathematics, fuzzy set theory [13] and fuzzy entropy measures [14–17] provide principled tools for quantifying ambiguity in graded concepts, such as emotion intensity. In contrast to benchmark suites such as EQ-Bench, Emotion-Bench, EmoBench, and EmotionQueen, we do not claim universal superiority; rather, we position the present framework as complementary. Its narrower purpose is to characterize within one structured-scoring pipeline (i) output reliability, (ii) controllability under the tested softmax sampling temperatures, (iii) ambiguity of continuous emotion scores, and (iv) persona sensitivity. A direct head-to-head comparison against benchmark suites such as EQ-Bench, EmotionBench, and EmoBench under matched prompts, datasets, and scoring rules is therefore an important next step rather than a claim made in the present paper.

1.3. Research Gap and Contribution

Our claim was intentionally narrowed down. We do not argue that emotion evaluation frameworks for LLMs are absent or ineffective. To the best of our knowledge, the reviewed benchmark, persona-prompting, and survey literature has not yet provided a single quantitative pipeline that jointly characterizes reliability under structured scoring, controllability under tested decoding conditions, ambiguity of continuous emotion scores, and persona-conditioned variation within one analysis. The closest prior work addressed only subsets of these axes.

Discrete Classification Bias: Most evaluations use categorical emotion labels, ignoring the continuous and overlapping nature of emotional experiences.

Parameter-Agnostic Assessment: Benchmarks evaluate models at fixed parameter settings, missing the opportunity to characterize controllability and reliability trade-offs.

Single-Perspective Limitation: Evaluations typically employ a single assessment perspective, failing to capture cognitive diversity in emotional interpretation.

Lack of Uncertainty Quantification: Few methods rigorously measure the ambiguity or confidence in LLM emotional assessments.

Limited Theoretical Integration: Existing approaches lack mathematical foundations connecting operational parameters to assessment quality.

This study addresses these limitations through four main contributions. The intended advantage is therefore not broader task coverage than dedicated benchmark suites, but denser operational characterization per structured-scoring run:

- **Unified mathematical framework:** We formalize LLM “emotional individuality” by integrating temperature-based controllability, fuzzy-theoretic ambiguity quantification, and persona-based cognitive diversity into a single mathematical characterization framework.
- **Reliability–controllability evaluation space:** Using persona-assigned sampling temperatures ($T \in \{0.1, 0.4, 0.7, 0.9\}$), we quantify model reliability and temperature controllability across 36 LLMs from seven vendors and organize models in a two-dimensional operational space.
- **Fuzzy entropy for assessment ambiguity:** We apply Shannon-type fuzzy entropy (membership entropy) to quantify ambiguity in continuous emotion ratings and empirically characterize vendor/model differences in entropy and fuzzy membership distributions.
- **Persona/text sensitivity + screening guidance:** We quantify persona- and genre-driven variation (including correlation structure) and distill these metrics into a practical screening and benchmarking framework for structured emotion-sensitive applications.

This manuscript builds on our prior conference work on platform development and preliminary analyses [18–20] and unifies them into a single, more rigorous mathematical treatment and multidimensional empirical characterization.

1.4. Research Hypotheses

To guide our investigation and structure our mathematical analysis, we formulated the following four hypotheses:

- **H1 (Temperature–Reliability Hypothesis):** A quantifiable relationship may exist between the tested temperature settings and response reliability in structured emotional assessment tasks, with this relationship varying across model architectures and training methodologies.
- **H2 (Fuzzy Entropy–Ambiguity Hypothesis):** Fuzzy entropy measures provide rigorous mathematical quantification of assessment ambiguity in LLM emotional evaluations, with entropy values correlating with model consistency and interpretative confidence.
- **H3 (Persona Modulation Hypothesis):** Persona prompting induces statistically detectable and systematic differences in emotional assessments, reflecting perspective dependence in LLM-based emotion evaluation.
- **H4 (Vendor-Specific Individuality Hypothesis):** LLMs from different development organizations exhibit distinct, mathematically characterizable “emotional individuality” profiles across multiple dimensions including fuzzy entropy distributions, temperature response patterns, and assessment consistency metrics.

1.5. Experimental Design Overview

We conducted comprehensive experiments with 36 state-of-the-art LLMs (as accessed during our data collection window: 24 April–1 May 2025) from seven major vendors (OpenAI, Anthropic, Google, xAI, DeepSeek, Meta, and Alibaba) using Japanese literary texts as the stimuli. Our experimental design systematically varied the following factors.

- **Persona–temperature configurations:** Four personas, each assigned a sampling temperature T_k with $\{T_k\}_{k=1}^4 = \{0.1, 0.4, 0.7, 0.9\}$.
- **Literary Texts:** Three works with diverse emotional profiles and genre characteristics.
- **Trials:** 10 independent repetitions per condition.

Each LLM evaluated four emotional dimensions (Interest, Surprise, Sadness, and Anger) on a continuous $[0, 100]$ scale, providing both numerical assessments and natural language justifications. This design yielded 4227/4320 trial responses (97.8% coverage), with 4067/4227 containing valid numeric emotion scores (96.2%), enabling robust statistical analysis and mathematical characterization.

By establishing rigorous mathematical foundations for LLM emotional evaluation, this study contributes to both applied mathematics (particularly fuzzy systems and information theory) and AI evaluation research. The following section develops the framework and the associated experimental protocol.

2. Materials and Methods

This section presents the mathematical foundations of our framework and describes the experimental methodology employed to characterize emotional individuality in LLMs. We first develop the theoretical underpinnings by integrating temperature control theory, fuzzy set theory, and persona-based cognitive diversity analysis, and then detail the experimental design that enables the empirical validation of our hypotheses.

2.1. Mathematical Foundations

Definition 1 (Emotional Individuality). For a model M , let $\mathbf{Y}_{M,\tau,\mathcal{P},T,r} \in [0, 100]^4$ denote the random four-dimensional emotion-score vector generated for text τ under persona $\mathcal{P}$, assigned temperature T , and repeat index r . We define the emotional individuality of model M in this study as the family of conditional response laws

$$\mathfrak{E}_M := \{\mathcal{L}(\mathbf{Y}_{M,\tau,\mathcal{P},T,r}) : (\tau, \mathcal{P}, T, r) \text{ is an observed experimental condition}\}.$$

Because persona and temperature are paired in our design, $\mathfrak{E}_M$ is defined over the observed persona–temperature conditions rather than over an independently crossed persona $\times$ temperature space. The scalar summaries ρ_M , κ_M , $\bar{H}_f^M$, C_M , and the persona/text effect sizes are descriptors of $\mathfrak{E}_M$, not the construct itself.

2.1.1. Softmax Sampling Temperature Control

Temperature as a Sampling Control

LLMs generate outputs by sampling from a token distribution. In this study, temperature refers to the decoding-time temperature parameter exposed by the API; in standard implementations, this parameter rescales token logits before softmax token sampling and thereby modulates sampling randomness, although provider-specific implementations may differ in detail. Given logits $\mathbf{z} = [z_1, \dots, z_V]$ for vocabulary size V , the standard temperature-adjusted probabilities are

$$p_i^{(T)} = \frac{\exp(z_i/T)}{\sum_{j=1}^V \exp(z_j/T)}, \quad i \in \{1, \dots, V\}, \quad T > 0.$$

For any two candidate tokens i and j ,

$$\frac{p_i^{(T)}}{p_j^{(T)}} = \exp\left(\frac{z_i - z_j}{T}\right),$$

where a smaller T magnifies logit differences and sharpens the token distribution, whereas a larger T compresses logit differences and flattens it. In the present study, this control acts only indirectly on the final 0–100 emotion scores: temperature can change lexical choice, justifications, and score dispersion; however, prompt constraints and semantic aggregation mean that score-level variance or entropy need not vary monotonically with T . We use four persona-assigned sampling temperatures, $T \in \{0.1, 0.4, 0.7, 0.9\}$ to quantify how controllability trades off against structured-output reliability.

Reliability–Controllability Space

We summarize the model behavior using reliability ρ_M (complement of the missing data rate) and temperature controllability κ_M (absolute correlation between temperature and response variance):

$$\rho_M := 1 - \text{MDR}(M), \quad \kappa_M := \left| \text{corr}((T_k)_{k=1}^K, (\sigma_M^2(T_k))_{k=1}^K) \right|.$$

Here, $\sigma_M^2(T_k)$ denotes the empirical variance of the pooled emotion scores at temperature T_k (computed over valid trials), and the correlation is computed across the K tested temperature levels. We use $(\rho_M, \kappa_M) \in [0, 1]^2$ to position models in four practical quadrants (Ideal/Stable/Controllable/Avoid) using the pragmatic screening cutoffs $\rho_M > 0.95$ and $\kappa_M > 0.5$; if κ_M is not computable (e.g., fewer than three usable temperature levels), we set $\kappa_M := 0$ for the quadrant assignment. The first cutoff flags models with fewer

than 5% invalid structured outputs under our prompt/parser, while the second separates moderate-to-strong temperature–variance association from weaker responses. These are operational thresholds rather than universal standards. *Interpretation note:* in our experimental design, the tested temperatures are persona-assigned (Section 2.2.3), so temperature is not manipulated independently of persona; accordingly, κ_M summarizes an association across the four persona–temperature conditions and should not be interpreted as a causal temperature effect holding persona fixed. Table A3 shows that nearby threshold choices preserve the same broad qualitative picture. Additional derivations linking temperature and fuzzy entropy are provided in the Appendix A; see Figure A1.

2.1.2. Fuzzy-Theoretic Emotional Assessment

Emotional intensity on a 0–100 scale is inherently graded. We map each score $x \in [0, 100]$ to fuzzy memberships over {LOW, MEDIUM, HIGH} using simple piecewise linear functions (definitions in the Appendix B). We chose this parameterization for three reasons: interpretability on the original 0–100 scale, fixed reproducible breakpoints, and overlap only between adjacent categories (never all three simultaneously). In particular, the design uses a central MEDIUM plateau over 30–50 with linear transition bands over 10–30 and 50–70, so the resulting ambiguity measure has a transparent geometric interpretation. Other membership choices are possible, so the reported entropy values should be interpreted relative to this parameterization. This design avoids discontinuities induced by hard thresholding and supports the quantitative ambiguity analysis.

Shannon-Type Fuzzy Entropy

For membership vector $\mu(x)$, we quantify ambiguity by

$$H_f(\mu(x)) := - \sum_{i \in \{\text{Low}, \text{Medium}, \text{High}\}} \mu_i(x) \log_2(\mu_i(x)), \quad 0 \log_2 0 := 0.$$

Because memberships are not required to sum to 1, H_f is used as a Shannon-type fuzziness index, the maximum of which depends on the membership-function design. A larger H_f indicates more ambiguous assessments (membership spread across categories), whereas $H_f \approx 0$ indicates concentrated membership. A proof sketch of the information preservation advantage over hard discretization is provided in the Appendix B (proof sketch: information preservation vs. hard discretization). For each model M , we summarize the ambiguity using the distribution and mean entropy $\bar{H}_f^M$ (see Section 3.2 and Table A4).

2.1.3. Persona-Based Cognitive Diversity Framework

Human emotional assessments exhibit systematic variations based on individual differences in cognitive style, affective sensitivity, and interpretative framework. We formalize this cognitive diversity through persona-defined cognitive profiles that parameterize assessment perspectives.

Definition 2 (Persona). A persona $\mathcal{P}_k$ is a tuple $(\mathcal{D}_k, T_k, \mathcal{C}_k)$ where:

- $\mathcal{D}_k \in \Sigma^*$ is a natural language description defining cognitive characteristics (distinct from the emotion-dimension tuple $\mathcal{D}$ defined later).
- $T_k \in [0, 1]$ is the persona-associated sampling temperature used in decoding.
- $\mathcal{C}_k$ represents expected cognitive traits (e.g., sensitivity, consistency, and creativity).

In our implementation, the expected traits $\mathcal{C}_k$ are operationalized through the prompt description $\mathcal{D}_k$ rather than being encoded as a separate structured parameter.

We define four personas that represent distinct positions in the cognitive diversity space:

$$\mathcal{P}_1 = (\text{"First-year university student"}, 0.7, \{\text{intuitive, emotionally engaged}\}) \quad (1)$$

$$\mathcal{P}_2 = (\text{"Literary researcher"}, 0.4, \{\text{analytical, objective}\}) \quad (2)$$

$$\mathcal{P}_3 = (\text{"Emotional poet"}, 0.9, \{\text{highly sensitive, creative}\}) \quad (3)$$

$$\mathcal{P}_4 = (\text{"Emotionless robot"}, 0.1, \{\text{logical, consistent}\}) \quad (4)$$

Persona–Temperature Labels and Analysis

The assigned values T_k are persona-specific sampling temperatures intended to represent increasing diversity/creativity in the elicited assessments. In inferential terms, however, the observed conditions are four persona–temperature bundles rather than an independently crossed persona $\times$ temperature design. We therefore evaluate persona effects using one-way ANOVA for each emotion dimension (reporting effect sizes) and summarize the persona-conditioned structure using PCA on persona-by-emotion features, while treating the temperature ordering only as a descriptive auxiliary axis. The methodological details and supplementary tables are provided in the Appendix C; key empirical findings are reported in Section 3.3. We also report (descriptively) the correlation between $\{T_k\}_{k=1}^4$ and persona-level pooled response variance as a sanity check for the intended ordering.

2.2. Experimental Design

2.2.1. Model Selection and Categorization

Sample Space Definition

We selected 36 LLMs accessed during our 24 April–1 May 2025, collection window from seven major development organizations. Inclusion required (i) API accessibility through our multi-provider experimental platform during that window, (ii) a stable public model identifier, (iii) compatibility with the shared text-input/structured score-output protocol used here, and (iv) sufficient vendor coverage to support both within-vendor and cross-vendor comparisons. The resulting set is therefore a time-stamped convenience sample of deployable models rather than a statistically representative sample of all available LLMs. Table 1 summarizes the evaluated model families, and the full inventory is provided in Appendix A.

Table 1. Evaluated Large Language Models (summary by vendor).

Vendor	Model Families	# Models
OpenAI	GPT-4.1/4o series; o-series (reasoning)	9
Anthropic	Claude series	4
Google	Gemini 2.x series; Gemma 3 series	10
xAI	Grok series	5
DeepSeek	DeepSeek series	3
Meta	Llama series	3
Alibaba	Qwen series	2
Total	—	36

Two-Dimensional Evaluation Matrix

Following our theoretical framework (Section 2.1.1), each model is positioned in the reliability–controllability space (ρ_M, κ_M) . This positioning enables the following mathematical classification:

$$\text{Quadrant}(M) = \begin{cases} \text{Ideal} & \text{if } \rho_M > 0.95 \text{ and } \kappa_M > 0.5 \\ \text{Stable} & \text{if } \rho_M > 0.95 \text{ and } \kappa_M \leq 0.5 \\ \text{Controllable} & \text{if } \rho_M \leq 0.95 \text{ and } \kappa_M > 0.5 \\ \text{Avoid} & \text{if } \rho_M \leq 0.95 \text{ and } \kappa_M \leq 0.5 \end{cases}$$

These cutoffs are pragmatic screening values rather than universal standards. The choice $\rho_M > 0.95$ (equivalently, $\text{MDR}(M) < 0.05$) flags models with fewer than 5% invalid structured outputs under the present prompt/parser, and $\kappa_M > 0.5$ identifies moderate-to-strong absolute temperature–variance association across the four tested persona–temperature conditions. Table A3 reports nearby threshold variants and shows that only a small number of border cases change quadrants.

2.2.2. Stimulus Materials and Emotion Dimensions

Literary Text Selection

Data collection was conducted using the experimental platform described in our previous study [18] which provides standardized multi-provider API integration and automated batch-processing capabilities. Three Japanese literary works from Aozora Bunko (a public domain digital library) were selected as stimulus materials.

1. **T1:** “Kaichu-dokei” (The Pocket Watch) by Yumeno Kyusaku (1923)—Allegorical dialogue between a pocket watch and mouse.
2. **T2:** “Okane to Pisutoru” (Money and Pistol) by Yumeno Kyusaku (1923)—Narrative with socioeconomic themes.
3. **T3:** “Boroboro na Dachou” (The Ragged Ostrich) by Takamura Kotaro (1958)—Poetic meditation on suffering.

We refer to these three texts as T1–T3 throughout (not to be confused with the sampling temperature T).

The selection criteria prioritized the following factors: (1) manageable length for LLM processing (500–1500 characters); (2) emotional richness enabling diverse interpretations; (3) genre diversity spanning allegory, narrative, and poetry; and (4) cultural-linguistic authenticity in the Japanese literary tradition. Therefore, we regarded the stimuli as intentionally task-specific rather than representative of everyday conversations, other languages, or other cultural settings.

Emotion Dimension Framework

We evaluated four fundamental emotion dimensions grounded in established psychological theory (Russell’s circumplex model and Plutchik’s basic emotions) [21,22]:

$$\mathcal{D} := (\text{Interest}, \text{Surprise}, \text{Sadness}, \text{Anger}).$$

We index emotion dimensions by $i \in \{1, 2, 3, 4\}$ according to the order in $\mathcal{D}$.

Each dimension was assessed on a numerical scale $[0, 100]$, with higher values indicating greater intensity. Although theoretically continuous for mathematical analysis, the LLM responses were constrained to integer values $\{0, 1, \dots, 100\}$ for practical implementation. Pairwise dependence between dimensions was assessed using Pearson correlations computed separately within each text (pooling models, personas, and trials), yielding $|r_{ij}| \leq 0.70$ for 17/18 dimension pairs (maximum $|r_{13}| = 0.702$ for Interest–Sadness in T2), indicating sufficient practical orthogonality for independent analysis.

2.2.3. Data Collection Protocol

Experimental Design Matrix

The experimental design systematically varied the models, texts, persona–temperature conditions, and repeated trials.

- **Models** (M): 36 LLMs.
- **Texts** (τ): 3 literary works.
- **Personas** (P): 4 cognitive profiles.
- **Trials** (R): 10 independent repetitions per condition.

This yields a theoretical maximum of $36 \times 3 \times 4 \times 10 = 4320$ total experimental conditions. The sampling temperature is fixed per persona, with $\{T_k\}_{k=1}^4 = \{0.1, 0.4, 0.7, 0.9\}$. Thus, temperature and persona are confounded, and the resulting κ_M estimates should be interpreted as associations across persona–temperature conditions (not as isolated temperature effects).

Response Format and Data Structure

Under each experimental condition, the LLM produced a structured assessment:

$$\mathbf{A} := (\mathbf{e}, \mathbf{r}), \quad \mathbf{e} := (e_1, e_2, e_3, e_4) \in [0, 100]^4, \quad \mathbf{r} := (r_1, r_2, r_3, r_4) \in (\Sigma^*)^4.$$

where:

- e_i is the numerical score for emotion dimension i ;
- r_i is the natural language justification string (over alphabet Σ).

Missing Data Handling

Response validity was determined using the following constraints:

$$\text{Valid}(\mathbf{A}) \iff \forall i \in \{1, 2, 3, 4\} : e_i \in \{0, 1, \dots, 100\}$$

If any emotion score was missing or non-numeric (i.e., could not be parsed), the trial was considered invalid. For downstream analysis, the numeric scores were rounded to the nearest integer and clipped to $[0, 100]$ when necessary. Let $\mathcal{A}_M$ denote the collection of all trial responses produced by model M . The missing data rate for model M is

$$\text{MDR}(M) = \frac{|\{\mathbf{A} \in \mathcal{A}_M : \neg \text{Valid}(\mathbf{A})\}|}{|\mathcal{A}_M|}$$

This metric serves as an inverse reliability measure with $\text{MDR}(M) = 0$ indicating perfect reliability for a given model M .

Reliability and Controllability Metrics

We compute $\text{MDR}(M)$ and define reliability as $\rho_M := 1 - \text{MDR}(M)$. The temperature controllability κ_M is computed as the absolute Pearson correlation between the tested temperatures and the corresponding pooled-score variances, using at least three usable temperature levels ($K \geq 3$) as shown in Section 2.1.1 for the formal definitions.

Final Dataset Characteristics

Data were collected between 24 April 2025, at 16:44, and 1 May 2025, at 16:03 (timestamps as stored in data_all.csv; recorded without an explicit timezone offset). Because API-served models can drift over time, the reported results should be interpreted as corresponding to the accessed model identifiers during this collection window.

Generative AI (GenAI) Use Disclosure

To generate the analysis dataset used in this study, we queried 36 Large Language Models (LLMs), listed in Table A1 (Appendix A), via publicly available APIs during our collection window (24 April–1 May 2025). The resulting model outputs (numeric emotion scores and accompanying natural-language justifications) constitute the primary experimental data analyzed in this work. The authors have reviewed the generated outputs and take full responsibility for the content of this publication.

Of the 4320 theoretical maximum experimental conditions, we collected 4227 trial responses (coverage $\approx 97.8\%$). After removing trials with missing or unparseable emotion scores, 4067 responses remained valid, yielding an overall completion rate of

$$\text{Valid-score rate (valid/collected)} = \frac{4067}{4227} \approx 0.962 \approx 96.2\%.$$

The distribution of missing data was highly non-uniform across models, ranging from 0% (26 models with perfect reliability) to 35.8% (Llama-4-Scout-17B), providing rich material for the reliability analysis.

Statistical Framing

We report effect sizes together with nominal p -values and use the latter primarily as descriptive/exploratory summaries. Correlation analyses use Pearson statistics for linear associations, with Spearman or nonparametric checks reported where they materially affect interpretation. Because assessments are nested within models, texts, personas, and repeated trials, pooled tests should not be read as strict i.i.d. inferences; when model-level summaries are analyzed, each model contributes one aggregate observation.

Ethical Considerations

All literary texts were selected from Aozora Bunko, a public domain collection that ensured no copyright violations. The LLM API usage followed each provider's terms of service. No human subjects were involved, and all assessments were performed computationally. This study investigated technical capabilities rather than replicating human emotional experiences, thereby avoiding concerns regarding anthropomorphization.

This comprehensive methodological framework integrates rigorous mathematical foundations with a systematic experimental design, enabling the robust testing of our four hypotheses while maintaining a clear differentiation from previous platform-focused studies. Mathematical formulations provide a theoretical foundation, whereas experimental protocols ensure empirical validity and reproducibility.

3. Results

This section presents comprehensive empirical findings that validate our mathematical framework using five analytical dimensions: temperature-controlled response characteristics, fuzzy-theoretic emotional characterization, persona-based diversity analysis, model clustering and consistency analysis, and text-dependent emotional correlations. Each subsection provides quantitative evidence supporting our hypotheses while revealing the mathematical structure underlying the emotional individuality of LLMs.

3.1. Temperature-Controlled Response Characteristics

3.1.1. Quantitative Temperature–Reliability Relationship

Missing Data Rate Distribution

Our analysis of response reliability across 36 LLMs revealed substantial heterogeneity in their abilities to produce valid, structured responses. The missing data rate (MDR),

defined as the proportion of trials yielding invalid or unparseable responses, varied dramatically across models, ranging from 0.0% (perfect reliability) to 35.8% (severe reliability issues).

Table 2 summarizes the models with nonzero missing data rates. The full per-model table is presented in Table A2.

Table 2. Response reliability summary: models with non-zero missing data rates. Here $\text{MDR}(\%) := 100 \times \text{MDR}(M)$ and $\rho_M := 1 - \text{MDR}(M)$.

Vendor	Model	MDR (%)	ρ_M
Meta	Llama-4-Scout-17B	35.8	0.642
Alibaba	Qwen3-235B-A22B-FP8	33.3	0.667
Google	gemini-2.5-pro-preview-03-25	17.5	0.825
Google	gemma-3-1b-it	16.7	0.833
Google	gemini-2.0-pro-exp	14.8	0.852
Google	gemini-2.5-flash-preview-04-17	9.2	0.908
Alibaba	Qwen2.5-VL-7B-Instruct	9.2	0.908
Google	gemini-2.0-flash-thinking-exp	6.7	0.933
Google	gemini-2.0-flash	0.8	0.992
OpenAI	o4-mini	0.8	0.992

Key Observation 1 (Perfect Reliability Cluster)

Twenty-six models (72.2%) achieved perfect reliability with an MDR of 0.0%, demonstrating robust instruction-following and structured output capabilities. This cluster included all Anthropic models, all xAI models, all DeepSeek models, and most OpenAI models, suggesting that many state-of-the-art models at the time of data collection could reliably follow the structured response format under this prompt and parsing criterion.

Key Observation 2 (Vendor-Specific Patterns)

Reliability patterns exhibited clear vendor-level clustering.

- **Anthropic, xAI, DeepSeek:** Perfect reliability across all models (100% success rate).
- **OpenAI:** Near-perfect reliability (>99.5% average).
- **Google:** Bimodal distribution with some models achieving near-perfect reliability (gemini-2.0-flash: 0.8%) while others show moderate unreliability (e.g., gemini-2.5-pro-preview-03-25: 17.5%).
- **Meta:** Severe reliability issues in Llama-4-Scout-17B (35.8%), while other Llama variants achieve perfect reliability.
- **Alibaba:** Elevated missing-data rates for both evaluated Qwen models, especially Qwen3-235B-A22B-FP8 (33.3%) and Qwen2.5-VL-7B-Instruct (9.2%).

Key Observation 3 (Prompt-Specific Variability)

Some models demonstrated differential reliability across emotional dimensions. Notably, Qwen3-235B-A22B-FP8 showed dramatically higher failure rates for Q4 (Anger: 33.3%) than for Q1–Q2 (17.5%), suggesting emotion-specific processing difficulties or prompt sensitivity.

Temperature Correlation Analysis

To quantify controllability as an operational temperature–variance association under persona-assigned temperatures, we computed Pearson’s correlation coefficients between temperature levels ($\{T_j\}_{j=1}^4 = \{0.1, 0.4, 0.7, 0.9\}$, one per persona) and response variance for each model with at least three usable temperature levels ($K \geq 3$). Because each temperature level is deterministically assigned to a persona in our design, this analysis does not isolate

temperature effects while holding the persona constant; it summarizes the association across the four persona–temperature conditions.

Table 3 presents representative examples from the temperature correlation analysis, revealing substantial heterogeneity in the responses of the different models to temperature modulation.

Table 3. Temperature controllability: correlation between temperature and response variance.

Model	Vendor	$r_{T,\sigma_M^2}(T)$	κ_M	Interpretation
gpt-4o-mini	OpenAI	−0.982	0.982	Extremely controllable
gemini-2.0-flash	Google	0.967	0.967	Extremely controllable
grok-3-mini-latest	xAI	0.907	0.907	Highly controllable
gemma-3-12b-it	Google	0.876	0.876	Highly controllable
Llama-4-Maverick-17B	Meta	−0.840	0.840	Highly controllable
gemma-3-4b-it	Google	−0.769	0.769	Moderately controllable
DeepSeek-V3-0324	DeepSeek	−0.706	0.706	Moderately controllable
gpt-4.1-nano	OpenAI	0.236	0.236	Weakly controllable
gemma-3-1b-it	Google	−0.039	0.039	Minimally controllable

Key Finding 1 (Approximately 25-Fold Variation)

The temperature–variance association index varies by a factor of approximately 25, from $\kappa_M \approx 0.982$ (gpt-4o-mini) to $\kappa_M \approx 0.039$ (gemma-3-1b-it). This substantial range is consistent with **Hypothesis H1** in the operational sense, that model behavior differs across the tested persona–assigned temperatures.

Key Finding 2 (Bidirectional Correlations)

The analysis revealed bidirectional correlations, with both positive and negative relationships between persona–assigned temperature levels and response variance, depending on the model. As a local descriptive approximation over the four tested temperature levels, the relationship can be summarized as

$$\sigma_M^2(T_j) \approx \sigma_{M,0}^2 + \beta_M T_j, \quad \beta_M \in \mathbb{R}$$

where $\sigma_{M,0}^2$ denotes a model-dependent intercept (an extrapolated baseline variance at $T = 0$) and β_M is a model-dependent slope fitted only over the observed temperature grid. We use this expression as a compact description of direction and magnitude, not as a mechanistic law or a derivative estimate valid outside the tested range.

Key Finding 3 (Outlier Models)

Some models show surprisingly weak sensitivity across persona–assigned temperature settings despite robust instruction following (e.g., claude-3-5-haiku-20241022: $\kappa_M \approx 0.120$), indicating that score dispersion remains comparatively constrained under the present prompt/parser, even when the assigned temperature changes. This finding shows that temperature control is not uniformly effective in this task setting and that model-specific calibration is essential.

Two-Dimensional Classification Matrix

By combining the reliability ($\rho_M := 1 - \text{MDR}(M)$) and controllability ($\kappa_M := |r_{T,\sigma_M^2}(T)|$) metrics, we positioned all 36 models in the two-dimensional evaluation space defined in Section 2.1.1. Figure 1 illustrates this classification.

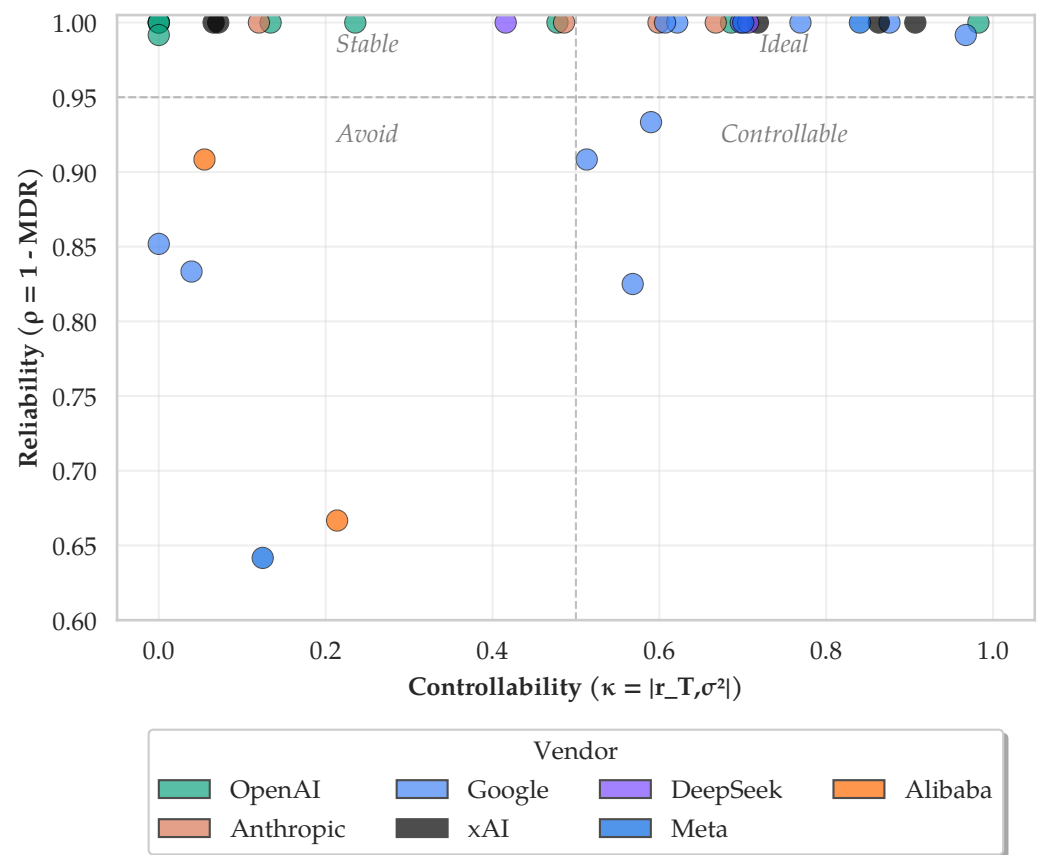


Figure 1. Reliability–controllability matrix. Two-dimensional scatter plot with reliability (y -axis, $[0, 1]$) vs. controllability (x -axis, $[0, 1]$). Quadrant boundaries at $\rho_M = 0.95$ and $\kappa_M = 0.5$ (we use $\rho_M > 0.95$ and $\kappa_M > 0.5$ for the high-reliability/high-controllability sides). Points are colored by the vendor.

The mathematical classification reveals the following:

Ideal Quadrant ($\rho_M > 0.95, \kappa_M > 0.5$): 16 models (44.4%)

- Anthropic: claude-3-haiku-20240307, claude-3-5-sonnet-20241022.
- OpenAI: gpt-4o, gpt-4o-mini.
- xAI: grok-2-latest, grok-3-latest, grok-3-mini-latest.
- DeepSeek: DeepSeek-R1, DeepSeek-V3-0324.
- Meta: Llama-3.3-70B-Instruct-Turbo, Llama-4-Maverick-17B.
- Google: gemini-2.0-flash, gemini-2.0-flash-lite, gemma-3-27b-it, gemma-3-12b-it, gemma-3-4b-it.

Interpretation: These models achieve the optimal balance of perfect or near-perfect reliability with strong responsiveness across the persona-assigned temperature settings, making them suitable for applications requiring both format stability and operationally adjustable diversity.

Stable Quadrant ($\rho_M > 0.95, \kappa_M \leq 0.5$): 12 models (33.3%)

- OpenAI models with missing temperature settings in our dataset (o1-mini, o3, o3-mini, o4-mini), for which we set $\kappa_M := 0$ under our computation.
- OpenAI GPT-4.1 family (gpt-4.1, gpt-4.1-mini, gpt-4.1-nano).
- Anthropic models with weak temperature sensitivity (claude-3-7-sonnet-20250219, claude-3-5-haiku-20241022) and DeepSeek-V3.
- xAI fast variants (grok-3-fast-latest, grok-3-mini-fast-latest).

Interpretation: These models maintain high reliability but show limited variation across the persona-assigned temperature settings, suggesting that their score distributions remain

comparatively constrained under this protocol. The result is descriptive; we do not attribute it to any specific internal alignment or architectural mechanism.

Controllable Quadrant ($\rho_M \leq 0.95, \kappa_M > 0.5$): 3 models (8.3%)

- gemini-2.0-flash-thinking-exp ($\rho_M \approx 0.933$).
- gemini-2.5-flash-preview-04-17 ($\rho_M \approx 0.908$).
- gemini-2.5-pro-preview-03-25 ($\rho_M \approx 0.825$).

Interpretation: These models show strong temperature responsiveness but moderate reliability issues, making them suitable for creative applications where controllability is prioritized over perfect reliability.

Avoid Quadrant ($\rho_M \leq 0.95, \kappa_M \leq 0.5$): 5 models (13.9%)

- Llama-4-Scout-17B ($\text{MDR}(M) \approx 35.8\%$, $\rho_M \approx 0.642$).
- Qwen3-235B-A22B-FP8 ($\text{MDR}(M) \approx 33.3\%$, $\rho_M \approx 0.667$).
- gemma-3-1b-it ($\text{MDR}(M) \approx 16.7\%$, $\rho_M \approx 0.833$).
- gemini-2.0-pro-exp ($\text{MDR}(M) \approx 14.8\%$, $\rho_M \approx 0.852$).
- Qwen2.5-VL-7B-Instruct ($\text{MDR}(M) \approx 9.2\%$, $\rho_M \approx 0.908$).

Interpretation: These models combine unreliable structured response generation with weak operational controllability in the present Japanese-literary assessment protocol, making them poor candidates for this specific benchmarking setup without substantial redesign of the prompting or parsing procedure.

Figure A2 visualizes three representative temperature–variance profiles (strong negative, strong positive, and near-flat). This clarifies that a large κ_M can arise from opposite slope directions, whereas a near-zero κ_M corresponds to empirically flat response–variance curves.

3.1.2. Temperature–Entropy Mathematical Relationship

Novel Analysis: Temperature–Fuzzy Entropy Correlation

Motivated by the expected-entropy identity in the Appendix A, we empirically tested whether temperature (assigned as a persona in our design) was associated with the mean fuzzy entropy in structured emotion assessment. For each model M and temperature T , we computed the mean fuzzy entropy $\bar{H}_f^M(T)$ across the valid trials and analyzed its association with T .

Key Finding 4 (Temperature–Entropy Correlation)

Across models demonstrating high controllability ($\kappa_M > 0.7$), we did not observe a statistically significant linear association between temperature and mean fuzzy entropy:

$$\begin{aligned} \text{corr}(T, \bar{H}_f^M(T)) &\approx -0.027, \\ 95\% \text{ CI} &: [-0.336, 0.287], \\ p &= 0.867, \quad n = 40 \text{ model–temperature points.} \end{aligned} \quad (5)$$

Computed over the pooled model: temperature points for models with $\kappa_M > 0.7$ (Pearson; CI via Fisher z); results are descriptive, given repeated measures per model. For reproducibility, we export the underlying model–temperature entropy profiles as `model_temperature_entropy_profiles.csv`; the pooled statistic is exported as `temperature_entropy_correlation_high_kappa.csv`.

Figure A1 illustrates the heterogeneous temperature–entropy profiles for representative models at the extremes of the controllability spectrum.

Mathematical Interpretation

This null result is consistent with the discussion in the Appendix A: the temperature can shift the probability mass toward the overlap regions of the fuzzy membership space (where H_f is larger); however, the direction and magnitude are model dependent and do not necessarily produce a monotone aggregate effect. For a fixed model M , we represent this relationship by the discrete profile

$$\{(T_j, \bar{H}_f^M(T_j))\}_{j=1}^4, \quad \{T_j\}_{j=1}^4 = \{0.1, 0.4, 0.7, 0.9\}.$$

This representation is descriptive (not a universal law) and should not be extrapolated beyond the tested temperatures.

Explaining Temperature-Independence Mathematically

For models exhibiting weak temperature sensitivity (e.g., gemma-3-1b-it with $\kappa_M \approx 0.039$), the fuzzy entropy remained nearly constant across temperature settings:

$$\text{Var}_{j \in \{1,2,3,4\}}(\bar{H}_f^M(T_j)) < 10^{-2}$$

where Var denotes the population variance over the four temperature-level means (ddof = 0), computed only for models with all four temperatures available.

Mathematical Explanation

Empirically, these models behave as if the score distribution is comparatively insensitive to the tested temperature ordering, so the mean fuzzy entropy remains nearly constant across conditions:

$$\bar{H}_f^M(T_j) \approx c_M, \quad j \in \{1, 2, 3, 4\}.$$

We consider this a descriptive pattern rather than evidence for a specific internal stabilization module. In practical terms, such near-flat profiles limit the usefulness of temperature as a controllability parameter in the proposed protocol. As demonstrated in Figure A1, gemma-3-1b-it exhibited minimal entropy variation across the temperature range.

Vendor-Specific Temperature Response Patterns

Aggregating the temperature correlation coefficients by vendor revealed systematic and organizational differences. Table 4 summarizes the vendor-level temperature-responsiveness statistics.

Table 4. Vendor-level temperature responsiveness.

Vendor	Mean κ_M	SD	Pattern
OpenAI	0.503	0.343	Lower responsiveness
xAI	0.525	0.423	Lower responsiveness
DeepSeek	0.607	0.165	Moderately responsive
Google	0.617	0.265	Mixed (bimodal)
Meta	0.555	0.379	Lower responsiveness
Anthropic	0.468	0.244	Lower responsiveness
Alibaba	0.134	0.112	Lower responsiveness

Key Finding 5 (Vendor Clustering)

Development organizations exhibit characteristic temperature response profiles, indicating vendor-level heterogeneity in operational controllability. Google shows a mixed

(bimodal) pattern with high within-vendor heterogeneity, whereas DeepSeek shows moderately responsive behavior. Possible contributors include architecture, fine-tuning, and provider-specific serving stacks; however, these cannot be separated or causally identified from the present API-level data.

3.2. Fuzzy-Theoretic Emotional Characterization

3.2.1. Fuzzy Membership Distribution Analysis

Vendor-Specific Fuzzy Distribution Patterns

Our analysis of 4227 emotion assessments through the fuzzy-theoretic framework (Section 2.1.2) reveals that different vendor groups exhibit systematically distinct patterns in how their models distribute emotional assessments across the Low, Medium, and High fuzzy membership regions.

Table 5 presents the vendor-level fuzzy membership distribution for the Interest dimension (Q1), illustrating the diversity of emotional assessment profiles.

Table 5. Vendor-specific fuzzy membership distribution (interest dimension).

Vendor	Mean	SD	Low (%)	Medium (%)	High (%)	Dominant
Alibaba	83.2	7.3	0.0	1.1	98.9	High
Meta	71.8	20.0	0.0	15.8	84.2	High
Anthropic	67.5	23.3	1.0	24.6	74.4	High
OpenAI	66.0	19.4	2.6	20.9	76.6	High
xAI	63.1	23.6	0.7	30.2	69.2	High (diffuse)
Google	61.4	25.7	9.1	28.7	62.2	High
DeepSeek	61.0	23.6	3.6	36.1	60.3	High

Percentages indicate the proportion of assessments where the corresponding fuzzy membership function is dominant (highest among low/medium/high).

Key Finding 6 (Vendor Emotional Profiles)

Vendors exhibit distinct “emotional personalities” mathematically characterized by fuzzy membership distributions:

- **Alibaba (High-biased, 83.2 ± 7.3 ; mean $\pm$ SD):** Assessments concentrate in the High region with low variance, indicating consistently positive/intense emotional evaluations with high certainty.
- **OpenAI (High-dominant, 66.0 ± 19.4 ; mean $\pm$ SD):** Predominantly High-dominant assessments with substantial spread, indicating moderately elevated Interest judgments with model-dependent variation.
- **xAI (High-dominant, 63.1 ± 23.6 ; mean $\pm$ SD):** High variance with a comparatively larger Medium share, indicating diverse, context-sensitive evaluations.

The difference between Alibaba and xAI represents a 20.0-point gap in the mean Interest assessment, providing empirical support for **Hypothesis H4** (vendor-specific emotional individuality).

Statistical Significance Testing

We used chi-squared tests of independence to determine whether fuzzy membership distributions differed significantly across vendors:

$$\chi^2 = \sum_{i=1}^7 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency (count) for vendor i in membership category j and E_{ij} is the expected frequency under independence. We construct O_{ij} by counting the

dominant fuzzy category (Low/Medium/High) for each evaluation; with $N = 4067$ valid observations, expected counts were generally large (minimum expected count ≈ 3.9), so the chi-square approximation is reasonable.

Results

Across the four emotion dimensions, the chi-square tests were all highly significant: $\chi^2(12) \in [162.5, 491.3]$ with all $p < 10^{-27}$, providing extremely strong evidence that vendor and fuzzy membership distribution are not independent Figure 2.

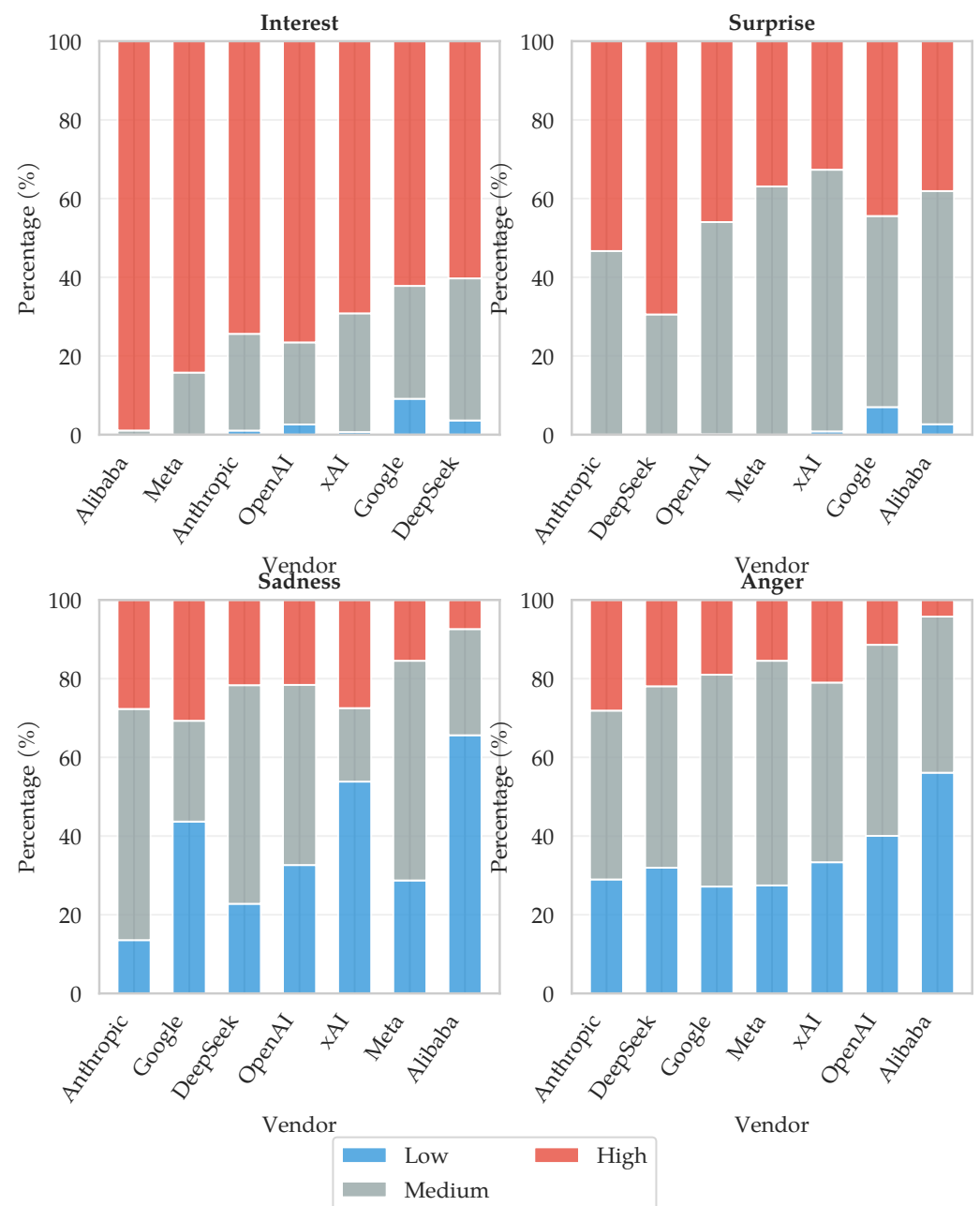


Figure 2. Vendor-specific fuzzy distributions across emotions. Stacked bar charts showing low/medium/high proportions for each vendor $\times$ emotion combination. Four panels (Interest, Surprise, Sadness, and Anger), each with seven vendor bars.

Cross-Emotion Dimension Analysis

Figure 2 shows the comprehensive fuzzy membership distributions across all the four emotion dimensions for each vendor group.

Key Finding 7 (Emotion-Specific Patterns)

Vendor profiles vary according to emotion type.

- **Positive Emotions (Interest, Surprise):** Show lower inter-vendor divergence in vendor-level mean scores, with $CV \approx 0.11$ for Interest (computed as the sample SD divided by the mean across the seven vendor means).
- **Negative Emotions (Sadness, Anger):** Show higher inter-vendor divergence in vendor-level mean scores, with $CV \approx 0.21$ for Sadness (computed analogously).

This asymmetry likely reflects differences in the composition of the training data, safety alignment procedures (which may suppress negative content), and cultural factors during dataset curation.

3.2.2. Fuzzy Entropy as Uncertainty Measure

Model-Specific Entropy Profiles

Fuzzy entropy $H_f(\cdot)$ quantifies the ambiguity inherent in the emotional assessment of each model. We computed the entropy distributions for all 36 models across 4067 valid evaluations, revealing substantial heterogeneity in assessment certainty. Table 6 summarizes representative models across the entropy range; the full table is provided in Table A4.

Table 6. Fuzzyentropy analysis: model-level mean entropy.

Model	Vendor	Mean $\bar{H}_f^M$	SD	Interpretation
Low Entropy (High Certainty)				
gemma-3-12b-it	Google	0.40	0.18	High certainty
gemini-2.5-flash-preview-04-17	Google	0.41	0.07	High certainty
gpt-4.1-mini	OpenAI	0.41	0.18	High certainty
gemini-2.0-flash-thinking-exp	Google	0.42	0.07	High certainty
gemma-3-27b-it	Google	0.43	0.28	High certainty
Medium Entropy (Moderate Ambiguity)				
Llama-3.3-70B-Instruct-Turbo	Meta	0.56	0.24	Moderate
o3	OpenAI	0.57	0.08	Moderate
gpt-4.1	OpenAI	0.57	0.19	Moderate
gemma-3-4b-it	Google	0.60	0.15	Moderate
Higher Entropy (More Ambiguity)				
claude-3-7-sonnet-20250219	Anthropic	0.66	0.12	Higher ambiguity

Note: Values represent the model-level mean entropy $\bar{H}_f^M$ across all valid assessments. Individual assessment entropy ranges are bounded by design in $[0, 1.058]$, because at most two fuzzy memberships overlap for any x ; model-level aggregation captures characteristic assessment patterns.

Key Finding 8 (Entropy Range)

The model-level mean fuzzy entropy ranged from approximately 0.40 to 0.66 across the full dataset, representing variations in assessment ambiguity patterns. This supports the first part of **Hypothesis H2**—namely, that fuzzy entropy provides a usable ambiguity descriptor—while the later entropy–reliability analysis shows that ambiguity and structural reliability are only weakly related in this dataset. Because the theoretical single-assessment maximum under our membership design is approximately 1.058, the observed model-level means correspond to roughly 38–63% of the attainable ambiguity range (Table A5).

Mathematical Interpretation of Entropy Extremes

There is no universal “high” or “low” entropy threshold independent of the membership design, and in this study, H_f quantifies the ambiguity of score placement rather than

population diversity. For orientation within the present parameterization, Table A5 reports the empirical quartiles of the model-level mean entropy distribution.

Low single-assessment entropy ($H_f(\mu(x)) \lesssim 0.4$): Scores clearly belong to a single fuzzy category (e.g., near 10 or 90; at exactly 50, $H_f(\mu(50)) = 0$) with minimal overlap. Mathematically,

$$\text{for a large fraction of assessments } x, \max_{i \in \{\text{Low}, \text{Medium}, \text{High}\}} \mu_i(x) > 0.9.$$

Low model-level mean entropy therefore indicates that a model spends more of its mass in crisp regions of the score scale.

High single-assessment entropy ($H_f(\mu(x)) \gtrsim 0.95$): Scores in transition regions can produce near-maximal ambiguity when adjacent fuzzy memberships are nearly equal:

$$\begin{aligned} \text{for many assessments } x, \mu_{\text{Low}}(x) > 0.2, \mu_{\text{Medium}}(x) > 0.2, \\ |\mu_{\text{Low}}(x) - \mu_{\text{Medium}}(x)| < 0.2 \\ \text{or} \\ \mu_{\text{Medium}}(x) > 0.2, \mu_{\text{High}}(x) > 0.2, \\ |\mu_{\text{Medium}}(x) - \mu_{\text{High}}(x)| < 0.2. \end{aligned}$$

For example, a score of $x = 18$ yields $\mu_{\text{Low}}(18) = \mu_{\text{Medium}}(18) = 0.4$ and $\mu_{\text{High}}(18) = 0$, producing $H_f(\mu(18)) \approx 1.058$. Model-level means in the observed 0.40–0.66 range therefore represent moderate average ambiguity—well below the single-assessment maximum but clearly above a purely crisp regime.

Entropy Distribution Visualization

Figure 3 shows the entropy distribution landscape obtained using the representative models from each vendor.

Representative Models Selected

- gemini-2.0-flash ($\bar{H}_f^M \approx 0.472$).
- Anthropic: claude-3-5-sonnet-20241022 ($\bar{H}_f^M \approx 0.454$).
- Llama-4-Scout-17B ($\bar{H}_f^M \approx 0.462$).
- o3-mini ($\bar{H}_f^M \approx 0.466$).
- DeepSeek-V3-0324 ($\bar{H}_f^M \approx 0.470$).
- grok-3-mini-fast-latest ($\bar{H}_f^M \approx 0.478$).
- Qwen2.5-VL-7B-Instruct ($\bar{H}_f^M \approx 0.513$).

Key Finding 9 (Vendor Entropy Clusters):

- **Anthropic and OpenAI:** Consistently low entropy (mean $\bar{H}_f^M < 0.47$), indicating confident, decisive assessments.
- **Google:** Moderate entropy with flagship model gemini-2.0-flash showing balanced assessments ($\bar{H}_f^M \approx 0.472$).
- **Alibaba:** Highest entropy among vendors ($\bar{H}_f^M \approx 0.513$), indicating more ambiguous assessments.
- **DeepSeek, xAI, Meta:** Intermediate entropy ranges (approximately in $[0.46, 0.48]$).

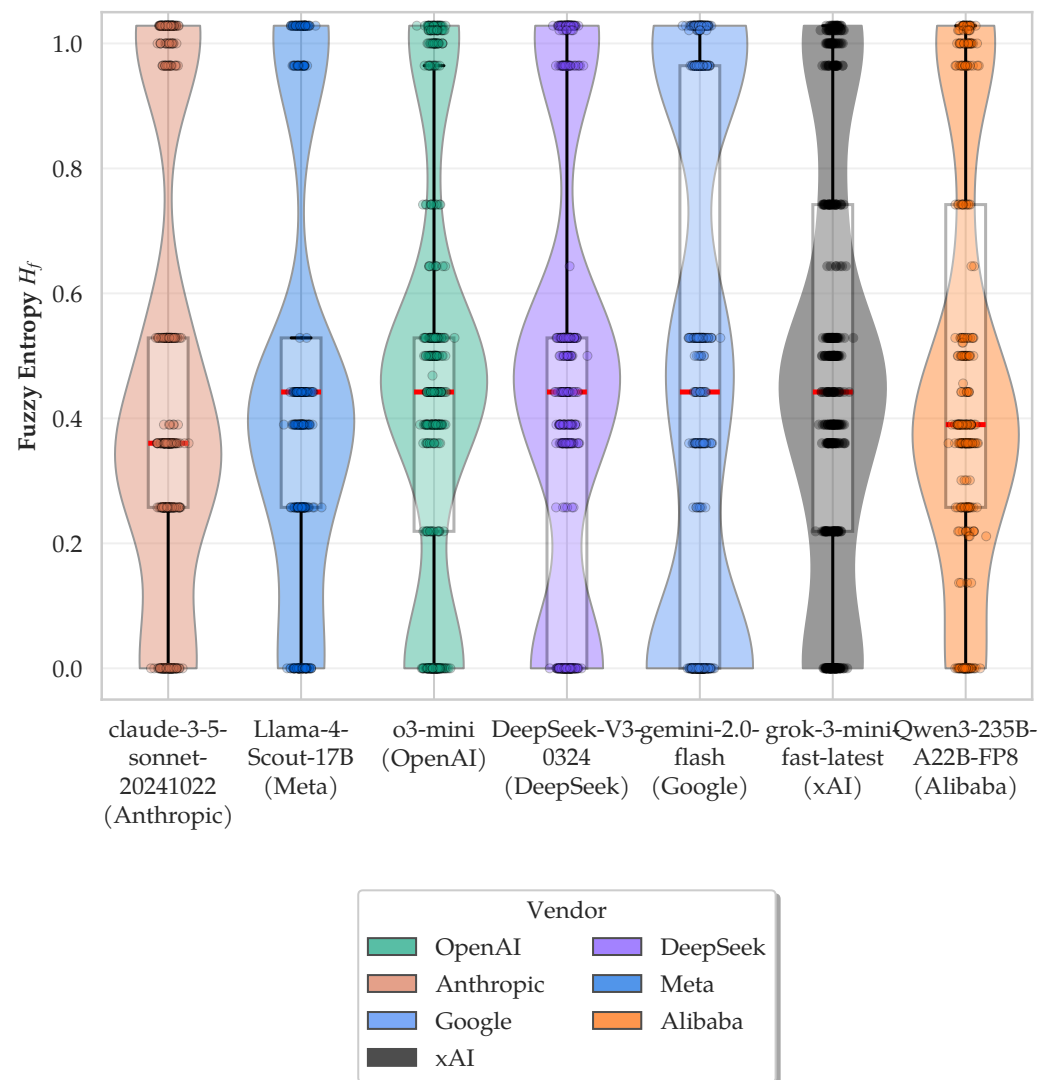


Figure 3. Fuzzyentropy distribution by model. Combined visualization showing entropy distributions for one representative model per vendor (7 models total). For most vendors, models closest to the median entropy were selected; Google was represented by gemini-2.0-flash, a flagship model. The plot layers three visualization techniques: (1) violin plots showing the overall distribution shape; (2) bee swarm plots displaying individual data points with transparency to reveal density patterns; and (3) box plots with transparent boxes highlighting the median and quartiles. The models were ordered by mean entropy from lowest to highest. Color-coded by vendor.

Correlation with Model Scale

We investigated whether the model size (parameter count) correlates with the model-level mean fuzzy entropy $\bar{H}_f^M$, hypothesizing that larger models might show lower entropy owing to better-calibrated confidence. Because parameter counts are not consistently disclosed for proprietary models, we restricted this analysis to models whose parameter counts were explicitly encoded in the model identifier (e.g., “7B”, “70B”, “235B”) and computed the Pearson correlation on $\log_{10}$ parameter counts (see `model_sizes_inferred.csv` and `model_scale_entropy_correlation.csv`).

$$r_{\log_{10}(\text{size}), \bar{H}_f^M} \approx 0.07, \quad p \approx 0.86, \quad n = 9 \text{ (models with name-inferred sizes)}$$

Within this limited subset, model scale alone does not explain entropy variation, suggesting that architecture and training procedures are at least important determinants of assessment ambiguity.

Entropy–Reliability Relationship

The relationship between the model-level mean fuzzy entropy $\bar{H}_f^M$ and response reliability (MDR) is examined as follows (model-level analysis):

$$r_{\bar{H}_f^M, \text{MDR}(M)} \approx -0.10, \quad p \approx 0.566, \quad n = 36$$

A rank-based robustness check yields $\rho_s \approx -0.078$ with $p \approx 0.651$, and Table A6 shows that both the Pearson and bootstrap Spearman confidence intervals include zero.

Interpretation: In this dataset, model-level fuzzy entropy shows no statistically significant linear or monotone association with missing data rate; the ambiguity of the valid scores and structural response reliability appear largely decoupled at the aggregate (model) level.

3.3. Persona-Based Diversity Analysis

3.3.1. Principal Component Analysis of Persona Effects

PCA Decomposition of Persona–Emotion Space

To mathematically characterize the cognitive diversity induced by personas, we performed a principal component analysis on the aggregated assessment matrix $\mathbf{X} \in \mathbb{R}^{101 \times 16}$ where observations (rows) represent (model, text) combinations and features (columns) represent four personas $\times$ four emotion dimensions (mean emotion scores per persona). The row count reflects complete-case filtering: we excluded (model, text) pairs missing at least one persona’s valid assessment (seven pairs), yielding 101 usable rows. The seven excluded pairs involved four models with incomplete persona coverage for specific texts (Alibaba/Qwen2.5-VL-7B-Instruct at T2; Google/gemini-2.0-pro-exp at T1–T3; Google/gemma-3-1b-it at T1 and T3; Meta/Llama-4-Scout-17B at T3). Table A7 summarizes the principal components (see Table A7).

Key Finding 10 (Low-Dimensional Latent Structure)

Three principal components explained 95.5% of the total variance, indicating that the 16-dimensional persona–emotion space has an inherently low-dimensional structure. This validates the mathematical tractability of our framework, in that emotional assessment patterns can be captured using a small number of latent factors.

PC1–PC3 Loadings Analysis

The inspection of the loadings clarifies that the latent structure is primarily emotional rather than a clean separation of personas. PC1 is a shared Sadness/Anger-versus-Interest axis with similar coefficients across all personas; PC2 is dominated by Surprise across personas; and PC3 (3.8% variance) is a weaker Anger/Interest versus Surprise/Sadness contrast. Table A8 lists the PC3 loadings by persona and emotion dimension (Table A8), and the full PC1–PC3 loadings are exported as `pca_loadings_pc1_pc3.csv`.

Mathematical Interpretation

The similarity of the loadings across personas indicates that PCA is capturing shared latent response axes more strongly than isolating a single “persona component”. In other words, persona modulation exists (as confirmed by the ANOVA below); however, the dominant low-dimensional structure is organized first by cross-emotion contrasts and then by persona labels.

Statistical Validation via ANOVA

To test **Hypothesis H3** (persona modulation), we conducted a one-way analysis of variance (ANOVA) for each emotion dimension across the four personas. Because this pooled analysis treats evaluations as independent despite nesting, we interpret p -values descriptively and emphasize the effect sizes (η^2). Table 7 presents the one-way ANOVA results for the effects of persona on each emotion dimension.

Table 7. ANOVA results: persona effects on emotional assessments.

Emotion	F-Statistic	p-Value	η^2	Effect Size
Interest	$F(3, 4063) = 9.00$	<0.001	0.007	Small
Surprise	$F(3, 4063) = 19.13$	<0.001	0.014	Small-Medium
Sadness	$F(3, 4063) = 2.74$	0.042	0.002	Small
Anger	$F(3, 4063) = 3.30$	0.020	0.002	Small

$F(df_1, df_2)$ denotes the F -statistic with degrees of freedom; η^2 = eta-squared effect size.

Key Finding 11 (Significant Persona Effects)

All four emotion dimensions showed significant persona effects at $\alpha = 0.05$, with Surprise demonstrating the strongest effect ($F = 19.13, p < 0.001$). This provides statistical evidence supporting **Hypothesis H3**, validating that the assigned personas significantly modulate emotional assessments across diverse text stimuli (T1–T3).

Post Hoc Multiple Comparisons

Post hoc pairwise comparisons using two-sided pooled-variance t -tests (unadjusted; exploratory) revealed specific differences between the groups.

- **Interest:** P3 (Poet) vs. P4 (Robot): Mean difference = 4.1 points, $p = 1.05 \times 10^{-4}$.
- **Surprise:** P3 vs. P4: Mean difference = 6.5 points, $p = 1.35 \times 10^{-12}$.
- **Sadness:** P3 vs. P4: Mean difference = 3.3 points, $p = 0.0076$.
- **Anger:** P3 vs. P4: Mean difference = 3.3 points, $p = 0.0040$.

The Poet–Robot contrast consistently showed the largest effect sizes, mathematically confirming our theoretical positioning of these personas as cognitive extremes.

Two-Way ANOVA: Vendor $\times$ Persona Interaction

We conducted a two-way analysis of variance (ANOVA) to investigate whether persona effects varied across vendor groups:

$$y_{ijk} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, 7, \quad j = 1, \dots, 4$$

where μ_0 is the grand mean, α_i represents the vendor effect, β_j represents the persona effect, $(\alpha\beta)_{ij}$ represents their interaction, and ε_{ijk} is a mean-zero error term (with standard identifiability constraints / reference-category coding as in our implementation). Table A9 reports the two-way ANOVA results (Vendor, Persona, and their interaction) for the Interest dimension (Table A9).

For reproducibility, we export the full Type-II ANOVA table (all effects; SS/df/MS/ $F/p/\eta_p^2$) as `vendor_persona_anova_typ2.csv`.

Key Finding 12 (Vendor-Level Heterogeneity Without Strong Interaction)

For the Interest dimension, the Vendor $\times$ Persona interaction was not significant ($F(18, 4039) = 0.86, p = 0.633$), suggesting that persona effects are broadly consistent across vendors at the aggregate level. Nonetheless, strong vendor main effects ($F = 34.8, p < 0.001$) and significant persona main effects ($F = 10.4, p < 0.001$) remained, indicating

substantial vendor-level heterogeneity and systematic persona-driven shifts without a dominant vendor-specific persona response pattern in this dataset.

3.3.2. Poet vs. Robot Persona: Mathematical Characterization

Quantitative Contrast Analysis

Given their theoretical positioning as cognitive extremes, we conducted a detailed comparative analysis of P3 (Poet, $T = 0.9$) and P4 (Robot, $T = 0.1$) personas. Table A10 provides the statistical comparison between the Poet and Robot personas across emotion dimensions (Table A10).

Key Finding 13 (Consistent Poet Elevation)

Poet Persona consistently produced higher emotional intensity assessments across all four dimensions, with effect sizes ranging from Cohen's $d \in [0.12, 0.32]$ (small effects). This pattern was mathematically systematic: $P3 > P4$ for all emotions, consistent with the intended persona design (persona-conditioned prompting at persona-assigned temperatures).

Non-Parametric Validation

Given the potential violations of normality assumptions, we confirmed the findings using the Mann–Whitney U test.

- **Interest:** $U = 588,477$, $p \approx 4.0 \times 10^{-9}$.
- **Surprise:** $U = 595,890.5$, $p \approx 1.2 \times 10^{-10}$.
- **Sadness:** $U = 578,661$, $p \approx 2.3 \times 10^{-7}$.
- **Anger:** $U = 550,869.5$, $p \approx 0.0028$.

For reproducibility, we exported these Mann–Whitney U test results as `poet_robot_mannwhitney.csv`.

All tests confirmed significant differences, providing robust statistical evidence independent of distributional assumptions.

Persona-Level Variance vs. Assigned Temperature (Descriptive Check)

To verify the intended ordering of the persona-assigned temperatures, we computed the correlation between the assigned temperatures $\mathbf{T} = (T_{P1}, T_{P2}, T_{P3}, T_{P4})^\top = (0.7, 0.4, 0.9, 0.1)^\top$ and the persona-level sample variance vector $\sigma^2 = (\sigma_{P1}^2, \sigma_{P2}^2, \sigma_{P3}^2, \sigma_{P4}^2)^\top$, where each σ_{Pj}^2 is computed by pooling all emotion scores (Interest/Surprise/Sadness/Anger) across all models, texts, and trials for persona Pj . Because each temperature is deterministically assigned to a persona in our design, this check reflects the association across the persona–temperature conditions and does not isolate temperature effects while fixing the persona.

Result:

$$r_{T, \sigma^2} \approx 0.614, \quad p \approx 0.386, \quad n = 4$$

This moderately positive association is descriptive (given $n = 4$) and is reported as a sanity check rather than an inferential validation. Models with higher persona-assigned temperatures (Poet: 0.9) tended to produce higher pooled response variances, whereas those with lower temperatures (Robot: 0.1) tended to produce more consistent and constrained responses.

Rank-Order Check (Descriptive)

Beyond simple correlations, we tested whether the rank order of the assigned temperatures perfectly predicted the rank order of the empirical variances. Let $\text{rank}(\mathbf{T}) = (3, 2, 4, 1)^\top$ (rank convention: 1 = lowest; Student = 3rd, Researcher = 2nd, Poet = 4th highest, Robot = 1st) and $\text{rank}(\sigma^2) = (1, 3, 4, 2)^\top$ be the empirically observed variance ranks.

Spearman's rank correlation:

$$\rho_s = 0.40, \quad p \approx 0.60, \quad n = 4$$

For reproducibility, we export the Spearman rank correlation inputs and result as `persona_temperature_rank_correlation.csv`.

Key Finding 14 (Theoretical Temperature–Persona Design)

The assigned temperature parameters $\{0.1, 0.4, 0.7, 0.9\}$ were designed as cognitive diversity anchors embedded in persona prompts and implemented as persona-specific sampling temperatures to realize the intended continuum from “Emotionless Robot” to “Emotional Poet”. The empirical Spearman's correlation ($\rho_s = 0.40$, $p \approx 0.60$) between these assigned parameters and observed variances shows weak and not statistically significant rank correspondence, whereas the systematic ANOVA effects ($F \in [2.74, 19.13]$, all $p < 0.05$) across emotion dimensions demonstrate that the personas successfully induced distinct interpretive frameworks.

This result represents a theoretical contribution by demonstrating that abstract cognitive characteristics can be conceptually mapped to temperature-like diversity parameters, thereby bridging the gap between psychological constructs and computational mechanisms. The effectiveness of the framework is evident in systematic persona effects, although the theoretical parameter-variance correspondence shows a moderate association rather than a deterministic one.

3.4. Model Clustering and Consistency Analysis

3.4.1. t-SNE Dimensional Reduction

High-Dimensional Emotion Space Structure

To visualize the inherent structure in the 48-dimensional emotion assessment space (four emotions $\times$ three texts $\times$ four personas), we applied t-distributed Stochastic Neighbor Embedding (t-SNE) after standardizing features, using perplexity = 30, 1000 iterations, and random seed 42. Because this 48-dimensional feature requires complete text $\times$ persona cells with valid scores, this analysis used 32 of the 36 models, excluding: Alibaba/Qwen2.5-VL-7B-Instruct, Google/gemini-2.0-pro-exp, Google/gemma-3-1b-it, Meta/Llama-4-Scout-17B. Figure A3 shows the resulting embedding.

Mathematical Distance Metrics

We computed pairwise Euclidean distances in the original 48-dimensional space to quantify the cluster cohesion. Table A11 summarizes the intra- and inter-cluster distances (see Table A11).

Cluster Stability Analysis

To assess stability, we performed bootstrap resampling (1000 iterations) of the original assessment data within each (model, text, persona) cell, recomputed hierarchical clustering in the original 48-dimensional feature space (t-SNE was used only for visualization), and measured cluster label agreement via the Adjusted Rand Index (ARI). For reproducibility, we exported the bootstrap ARI samples and summary as `clustering_stability_ari.csv` and `clustering_stability_ari_summary.csv`:

$$\text{ARI} := \frac{\text{RI} - \mathbb{E}[\text{RI}]}{1 - \mathbb{E}[\text{RI}]}$$

where $\text{RI} \in [0, 1]$ denotes the Rand Index (pairwise agreement proportion) and $\mathbb{E}[\text{RI}]$ is its expected value under the permutation model (random labelings with fixed cluster sizes).

Mean ARI across bootstrap samples: 0.82 (95% bootstrap percentile CI: [0.41, 1.00])

Interpretation: The mean ARI close to 1 indicates generally stable clustering (far exceeding chance agreement, $\text{ARI} \approx 0$), while the wider bootstrap interval indicates that alternative partitions can occur under resampling. Overall, this supports the idea that the five identified clusters reflect a genuine structure in model behavior rather than artifacts of visualization.

3.4.2. Numerical Stability Proxy (Consistency)

Consistency Metric Formulation

In addition to numerical emotion scores, the LLMs provide natural language justifications for each assessment; however, we did not compute a language–numerical alignment metric in this study. A direct and reproducible alignment metric would require a validated sentiment or emotion analyzer for Japanese justification text and a calibrated mapping to the $[0, 100]$ scale. Accordingly, we report a reproducible proxy that quantifies within-model numerical stability based on the numeric scores alone.

For each model M , we define the per-emotion coefficient of variation

$$\text{CV}_{M,q} := \frac{\sigma_{M,q}}{\mu_{M,q}}, \quad q \in \{1, 2, 3, 4\}, \quad \mu_{M,q} > 0$$

where $\mu_{M,q}$ and $\sigma_{M,q}$ denote the sample mean and sample standard deviation of scores for emotion dimension q across all valid assessments for model M (with $q = 1, 2, 3, 4$ corresponding to Interest, Surprise, Sadness, and Anger, respectively). Here $\sigma_{M,q}$ is computed as the sample standard deviation (ddof = 1). The model-level consistency score is then defined as the inverse mean coefficient of variation:

$$\begin{aligned} Q_M &:= \{q \in \{1, 2, 3, 4\} : \mu_{M,q} > 0\}, \\ \overline{\text{CV}}_M &:= \frac{1}{|Q_M|} \sum_{q \in Q_M} \text{CV}_{M,q}, \quad (|Q_M| > 0), \\ C_M &:= \begin{cases} \frac{1}{1 + \overline{\text{CV}}_M}, & |Q_M| > 0, \\ 0, & |Q_M| = 0, \end{cases} \quad C_M \in [0, 1]. \end{aligned} \quad (6)$$

Table A12 summarizes the resulting model-level consistency rankings (see Table A12).

Key Finding 15 (Consistency Range)

Consistency scores (computed as $C_M = 1/(1 + \overline{\text{CV}}_M)$) ranged from 0.780 (gemma-3-1b-it) to approximately 0.548 (gemini-2.5-flash-preview-04-17), representing variations in the within-model numerical stability patterns. The models in the top quartile demonstrated more stable assessment patterns with lower relative dispersion, whereas those in the bottom quartile showed higher variability across assessments.

Vendor Group Comparisons

Table A13 summarizes the vendor-level consistency statistics used for comparison (Table A13).

Key Finding 16 (Vendor Consistency Profiles)

- **Anthropic:** Highest mean consistency (0.693), suggesting relatively stable numeric scoring patterns at the vendor level.
- **Alibaba:** Lowest mean consistency (0.629), indicating more variable numeric scoring patterns on average.
- **Google:** Largest within-vendor dispersion ($\text{SD} = 0.077$), driven by heterogeneous model behavior spanning both very high and very low consistency scores.

Correlation with Other Metrics

The consistency scores showed a weak linear association with other model-level metrics in this dataset.

- **vs. Fuzzy Entropy:** $r \approx 0.00$, $p \approx 0.980$ (no detectable linear association).
- **vs. Reliability:** $r \approx 0.052$, $p \approx 0.765$ (no detectable linear association).

These results suggest that the numerical stability proxy C_M captures a distinct aspect of model behavior that is not well explained by reliability or model-level fuzzy entropy alone; therefore, developing a fully language–numerical alignment metric remains an important direction for future work.

3.5. Text-Dependent Emotional Correlations

3.5.1. Genre-Specific Correlation Patterns

Text-Emotion Correlation Analysis

To investigate how literary genres influence emotional structure, we computed Pearson’s correlations between the emotion dimensions for each text. Table A14 summarizes the text-specific Interest–Sadness correlations used for this comparison (Table A14).

Key Finding 17 (Genre-Dependent Correlation Structure)

The correlation between Interest and Sadness varies dramatically by text genre.

- **Allegorical (T1):** Weak negative correlation ($r = -0.195$), suggesting that philosophical/allegorical content can be simultaneously interesting and sad without strong inverse relationship.
- **Narrative (T2):** Strong negative correlation ($r = -0.702$), reflecting typical narrative structure where engaging plot elements contrast with melancholic themes.
- **Poetic (T3):** Strong negative correlation ($r = -0.677$), consistent with poetic tension between aesthetic appreciation and sorrowful content.

This 3.6-fold variation in correlation strength ($|r| \in [0.195, 0.702]$) demonstrates that emotional structure is not fixed but is dynamically shaped by literary genre, validating **Hypothesis H4** regarding text-dependent assessment patterns.

Comprehensive Correlation Matrices

Figure 4 shows the complete correlation matrices for all emotion pairs across the three texts.

Notable Cross-Text Patterns

Surprise-Anger Correlation:

- T1 (Allegorical): $r = 0.334$ (moderate positive).
- T2 (Narrative): $r = -0.146$ (weak negative).
- T3 (Poetic): $r = 0.207$ (weak positive).

The emergence of a positive Surprise–Anger correlation in T3 (poetic text) suggests that unexpected elements in poetry may co-occur with frustration or aggressive imagery, a pattern that is absent in other genres.

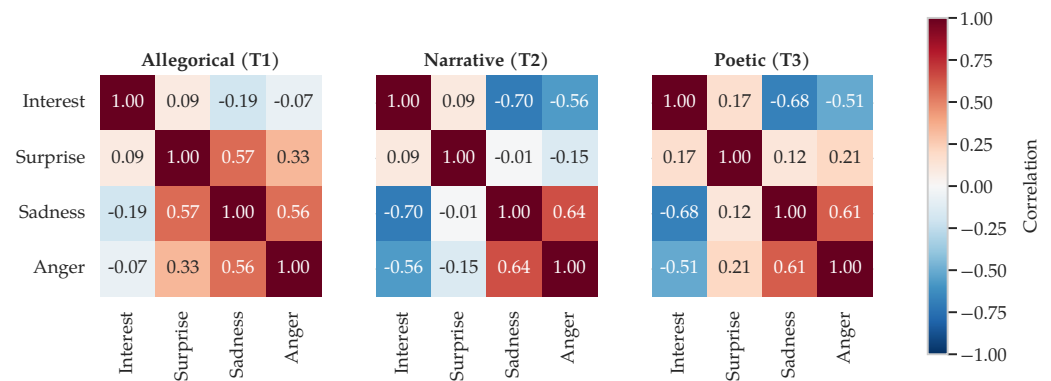


Figure 4. Genre-specific emotion correlation matrices. Three 4 × 4 correlation heatmaps (one per text) showing Pearson correlations between Interest, Surprise, Sadness, and Anger. Color scale in [−1, 1] (blue at −1, red at +1).

Statistical Significance of Text Effects via ANOVA

To formally test whether text content significantly influenced emotional assessments, we performed a one-way analysis of variance (ANOVA) for each emotion dimension with text as the independent variable. As mentioned earlier, this pooled analysis treats evaluations as independent; we interpret p -values descriptively and emphasize the large effect sizes. For reproducibility, we export the per-emotion ANOVA statistics as `text_effects_anova.csv`. Table 8 presents the one-way ANOVA results for text effects across emotion dimensions.

Table 8. ANOVA: text effects on emotional assessments.

Emotion	F-Statistic	η^2	Interpretation
Interest	$F(2, 4064) = 2571.80, p < 0.001$	0.559	Very large
Surprise	$F(2, 4064) = 1999.16, p < 0.001$	0.496	Very large
Sadness	$F(2, 4064) = 5213.37, p < 0.001$	0.720	Very large
Anger	$F(2, 4064) = 2430.47, p < 0.001$	0.545	Very large

All effects were significant at $p < 0.001$. Effect sizes: small ($\eta^2 < 0.06$), medium ($\eta^2 \in [0.06, 0.14]$), large ($\eta^2 > 0.14$).

Key Finding 18 (Dominant Text Effects)

Text content accounts for **approximately [50, 72]% of the variance** in emotional assessments ($\eta^2 \in [0.496, 0.720]$), providing extremely strong support for the text-dependent nature of emotional assessment in this dataset. The effect sizes are exceptionally large by conventional standards ($\eta^2 > 0.14$ is considered “large”), demonstrating that textual content is the primary determinant of emotional evaluation, far exceeding model-specific or persona-specific contributions.

For Sadness, the text explained approximately 72% of the variance ($\eta^2 \approx 0.720$), the highest observed effect. This likely reflects the stark contrast between T3 (“Ragged Ostrich”, deeply melancholic) and T1 (“Pocket Watch”, whimsical allegory).

Non-Parametric Validation via Kruskal–Wallis

Given the large sample size and potential normality violations, we confirmed the results using non-parametric Kruskal–Wallis tests. Table A15 presents the corresponding statistics (Table A15).

The epsilon-squared effect sizes ($\epsilon^2 \in [0.568, 0.578]$) from the Kruskal–Wallis tests were uniformly very large, corroborating the strong text effects observed under ANOVA.

3.5.2. Mathematical Interpretation of Correlation Structures

Emotional Tension Theory in Literary Analysis

The observed correlation patterns can be interpreted through the literary theory concept of “emotional tension”—the dynamic interplay between conflicting affects within a text. We formalize this from a mathematical viewpoint.

Definition 3 (Emotional Tension). *For a text τ , the emotional tension $\mathcal{T}_{ij}(\tau)$ between an emotion pair (e_i, e_j) is quantified as*

$$\mathcal{T}_{ij}(\tau) := \max\{0, -r_{ij}(\tau)\} = -\min\{r_{ij}(\tau), 0\} \in [0, 1]$$

where $r_{ij}(\tau)$ is the Pearson correlation coefficient between emotions e_i and e_j computed over all valid assessments for text τ . A larger $\mathcal{T}_{ij}(\tau)$ indicates a stronger negative correlation (competing emotions), whereas $\mathcal{T}_{ij}(\tau) = 0$ indicates a non-negative association (co-occurrence or independence). Table A16 reports the correlation coefficients used to quantify emotional tension across texts for the selected emotion pairs (see Appendix E). For reproducibility, we export the underlying per-text Pearson correlations for these pairs as `text_correlation_pairs.csv`.

Mathematical Interpretation

Narrative and poetic texts (T2, T3) exhibit a strong Interest–Sadness anti-correlation, mathematically representing the emotional structure in which engaging qualities (plot and imagery) contrast with sorrowful content. The correlation of $r = -0.702$ for T2 implies that a 1-SD increase in Interest is associated with a 0.702-SD decrease in Sadness (in standardized units), reflecting the text’s thematic opposition between material fascination and moral tragedy.

Conversely, allegorical texts (T1) showed a weak Interest–Sadness correlation ($r = -0.195$), allowing for simultaneously high scores in both dimensions. This reflects the philosophical nature of intellectual engagement, which does not preclude the recognition of melancholic themes in the work.

Variance Decomposition Analysis

To quantify the relative contributions of text content versus other factors (for the Interest dimension), we computed marginal variance attributions using ANOVA-style sums of squares on complete-case observations:

$$\text{Var}_{\text{total}} := \text{Var}(y)$$

where $\text{Var}(y)$ denotes the sample variance (ddof = 1).

$$\text{Var}_A := \frac{1}{n-1} \sum_{a \in A} n_a (\bar{y}_a - \bar{y})^2, \quad \text{Var}_{\text{residual}} := \text{Var}_{\text{total}} - \sum_{B \in \{\text{Text}, \text{Model}, \text{Persona}\}} \text{Var}_B$$

where y denotes the Interest score, A is a factor with levels a , n_a is the number of observations in level a , $\bar{y}_a$ is the level mean, $\bar{y}$ is the grand mean, and n is the total number of complete-case observations. The reported percentages are $\text{Var}_A / \text{Var}_{\text{total}}$. Table A17 summarizes the resulting variance component attributions for the Interest dimension (Table A17).

Key Finding 19 (Dominance Hierarchy)

Text content dominated the variance structure (55.9%), followed by model-specific effects (12.9%), with persona effects contributing minimally (0.7%). The remaining 30.6% reflects the residual variation not captured by these marginal factor attributions. This

provides a precise quantification of factor importance, validating that **textual content is the primary driver** of emotional assessment outcomes.

Cross-Text Consistency Metric

For each model, we computed cross-text consistency to measure whether the models maintained stable assessment profiles across genres:

$$\sigma_{M,i}^2 := \text{Var}\left(\frac{1}{4} \sum_{d=1}^4 e_d \mid M, \tau_i\right), \quad \sigma_{M,\max}^2 := \max_{i \in \{1,2,3\}} \sigma_{M,i}^2$$

$$\text{CTC}_M := \frac{1}{3} \sum_{i=1}^3 \frac{\sigma_{M,i}^2}{\sigma_{M,\max}^2}, \quad \text{with } \sigma_{M,\max}^2 = 0 \Rightarrow \text{CTC}_M := 1.$$

where τ_i is the i -th text (genre condition), e_d denotes the score for emotion dimension d (with $d = 1, 2, 3, 4$ corresponding to Interest, Surprise, Sadness, and Anger), and $\sigma_{M,i}^2$ is the within-text sample variance (ddof = 1) of the per-assessment mean intensity $\frac{1}{4} \sum_{d=1}^4 e_d$ across all valid assessments for model M under text τ_i .

Models with high CTC maintained consistent scoring patterns regardless of text genre (potentially indicating genre insensitivity or rigid assessment frameworks), whereas those with low CTC indicated strong text responsiveness.

Mean CTC Across All Models

0.61 (SD = 0.13), indicating moderate cross-text consistency overall, with substantial variation across models. This further validates that genre sensitivity is model dependent.

Section 3 Summary

Section 3 reports the results, which broadly support **H1**, **H3**, and **H4** across five complementary perspectives: reliability/controllability, fuzzy-theoretic ambiguity, persona-driven diversity, model clustering/consistency, and text dependence. **H2** is only partially supported: fuzzy entropy is informative as an ambiguity descriptor, but its model-level association with response reliability is weak in this dataset. To reduce the reviewer's burden, detailed per-model tables and auxiliary analyses are provided in Appendix A and the subsequent appendix sections.

4. Discussion

This section synthesizes the results into a compact integrated view of LLM emotional individuality and provides actionable guidance for future research. Detailed numerical repetitions and extended methodological notes are provided in Appendix A and the subsequent appendix sections.

4.1. Mathematical Framework Integration

Across the 36 models, the results support a three-part characterization: (i) reliability–controllability behavior in the (ρ_M, κ_M) plane (Figure 1 and Tables 2–4); (ii) fuzzy-theoretic ambiguity quantified by H_f and its distributions (Figures 2 and 3, Table 6, and Table A4); and (iii) persona/text dependence (Table 7 and Figure 4). Together, these dimensions show that emotional individuality is multifactorial and only partially controllable via temperature: some models allow calibrated variation without format failures, while others remain stable (low κ_M) or become unreliable (low ρ_M) under stochastic decoding.

4.2. Vendor-Specific Emotional Individuality

We observe vendor-level “fingerprints” across multiple metrics: membership/entropy profiles, temperature responsiveness, and genre-specific correlation structures. We treat

these as empirical signatures rather than attributing them to specific training data or alignment choices because such causal claims cannot be verified from API-access experiments alone. Supplementary per-model tables and auxiliary analyses supporting these patterns are provided in Appendix A.

4.3. Practical Implications for Model Selection

The framework is most immediately useful as a benchmarking and model-screening tool for structured emotion-scoring tasks. A practical workflow is as follows: first, filter by reliability ρ_M (can the model consistently emit valid structured scores under the protocol?); second, inspect controllability κ_M (do the tested persona–temperature conditions actually change score dispersion?); third, tune ambiguity and sensitivity using $\bar{H}_f^M$, C_M , and persona/text effects. In this sense, the framework can support benchmark design, stress testing under prompt/framing variation, and longitudinal re-evaluation of the same models over time. It is *not* by itself a safety certification procedure or a causal audit of internal alignment mechanisms. Table 9 summarizes evidence-based screening rules; the example models are illustrative and should be interpreted as time-stamped to the evaluated versions.

Table 9. Task-scoped screening framework for structured emotion assessment.

Use Case	Screening Rule (with Illustrative Examples)	Key Caveat
Structured benchmarking/parser-sensitive evaluation	Prioritize ρ_M and require $\rho_M > 0.95$; illustrative examples include claude-3-5-sonnet-20241022 and grok-3-latest.	Reliability here refers to a low invalid-output rate under the present prompt/parser only.
Controlled diversity/stress testing	Require $\rho_M > 0.95$ and $\kappa_M > 0.5$; illustrative examples include gpt-4o-mini and gemini-2.0-flash.	A high κ_M summarizes the association across persona–temperature bundles, not a pure causal temperature effect.
Low-ambiguity scoring	Prioritize ρ_M , $\bar{H}_f^M$, and C_M ; a conservative rule is $\rho_M > 0.95$ and $\bar{H}_f^M < 0.45$; illustrative examples include gemma-3-12b-it and gpt-4.1-mini.	Low entropy indicates crisp scores under the present membership design, not necessarily better emotional understanding.
Ambiguity-sensitive exploratory analysis	Prioritize ρ_M and $\bar{H}_f^M$; a conservative rule is $\rho_M > 0.95$ and $\bar{H}_f^M > 0.50$; illustrative examples include claude-3-7-sonnet-20250219 and gpt-4o.	Higher entropy can be useful when one wants models that often occupy fuzzy overlap regions, but it should not be conflated with safety.

ρ_M = reliability; MDR = missing data rate; κ_M = controllability; $\bar{H}_f^M$ = model-level mean fuzzy entropy; C_M = numerical stability proxy. The thresholds above are task-scoped screening rules for this study, not universal deployment standards.

4.4. Limitations and Future Directions

The key limitations of the study are as follows: (i) limited stimulus set (three Japanese literary texts) and language/culture dependence; (ii) limited emotion dimensionality (four dimensions); (iii) API model drift over time; (iv) text-only modality; (v) dependence on a structured score+justification response format; and (vi) perfect confounding between persona and assigned temperature, which prevents causal separation of those effects in the current design. Therefore, we view the present study as a task-scoped mathematical evaluation template rather than a final statement about cross-lingual generalization, multimodal emotion understanding, or internal alignment mechanisms. An additional limitation is that we have not yet performed a matched experimental comparison against

established LLM emotion-evaluation suites; future work should test whether the present characterization metrics add discriminatory value beyond benchmark families such as EQ-Bench, EmotionBench, and EmoBench when prompts, stimuli, and scoring rules are aligned. Extended notes and supplementary analyses are provided in Appendix A.

5. Conclusions

This study develops a unified mathematical evaluation framework for understanding LLM emotions and empirically validates it across 36 models. The results show that emotional individuality is multidimensional: reliability and controllability vary widely across models, ambiguity can be quantified using fuzzy entropy, and persona/text conditioning meaningfully changes observed assessments. Concurrently, the weak entropy–reliability relationship indicates that the ambiguity of valid scores and structural output reliability should be treated as distinct dimensions rather than collapsed into a single quality measure.

5.1. Primary Contributions

- **Unified evaluation framework:** A single characterization template combining reliability/controllability metrics, fuzzy-theoretic ambiguity (H_f), and persona-conditioned perspectives.
- **Large-scale empirical characterization:** A comparative study across 36 contemporary LLMs demonstrating heterogeneous reliability, controllability, and ambiguity profiles.
- **Vendor-specific fingerprints:** Quantitative vendor-level patterns in fuzzy distributions/entropy and genre-specific correlation structures.
- **Task-scoped screening guidance:** A practical framework mapping benchmarking and analysis needs to transparent quantitative screening rules (Table 9).

5.2. Limitations and Future Research Directions

The limitations include a restricted stimulus set (three Japanese literary texts), a four-dimensional emotion space, API model drift over time, text-only modality, dependence on a structured response format, and the current design’s inability to separate persona effects from temperature effects causally. Future work should expand stimuli (and languages), emotion dimensionality, and modalities; use a factorial persona $\times$ temperature design; and conduct matched comparisons against benchmark suites such as EQ-Bench, EmotionBench, and EmoBench to quantify the incremental value of the present characterization metrics.

5.3. Closing Statement

Overall, this study reframes LLM emotion understanding as a measurable object with an interpretable mathematical structure, enabling more transparent evaluations and principled deployment choices for emotion-sensitive applications.

Author Contributions: Conceptualization, N.S.; Methodology, N.S.; Software, N.S. and Y.Y.; Formal analysis, N.S. and Y.Y.; Investigation, S.W.; Writing—original draft preparation, N.S.; Writing—review and editing, N.N. and S.W.; Supervision, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by JSPS Grants-in-Aid for Scientific Research (JSPS KAKENHI), Grant Number 24K13350.

Institutional Review Board Statement: Not applicable. This study analyzed the computational outputs of Large Language Models using public domain literary texts. No human subjects were involved in this study.

Informed Consent Statement: Not applicable. This study did not involve human participants.

Data Availability Statement: All analysis code and a numeric-only dataset (scores + metadata) supporting the findings of this study are openly available in the GitHub repository at <https://github.com/nshrhlm/llm-emotional-individuality>, (accessed on 31 March 2026). The public dataset omits LLM-generated free-text justifications because the redistribution of generated text may be restricted by provider-specific terms; the analyses reported in this paper do not rely on these justification fields. The repository includes the following: 1. Complete numeric datasets and derived CSV outputs used in the paper; 2. Fuzzy membership function implementations and analysis scripts (Python); 3. Figure/table regeneration tools and exact command-line reproduction steps; 4. Persona descriptions, text metadata, and output schema documentation; 5. Reproduction environment: Python 3.12.11; exact dependencies pinned in requirements.txt. The repository-level reproducibility note also states an important boundary: the verbatim provider-specific prompt template used during collection was not versioned in the public materials; therefore, the public package supports the analysis-level reproducibility of the reported statistics and figures rather than a literal prompt replay.

Acknowledgments: We thank the developers and organizations behind the 36 evaluated LLMs for making their models available for research. We acknowledge the Aozora Bunko digital library for providing access to Japanese literary texts used as experimental stimuli. During the preparation of this study, the authors used the LLMs listed in Table A1 (Appendix A) for the purpose of generating the analysis data used in this work. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. This research was conducted independently, with no financial support from any LLM development organization. Model evaluations were performed using publicly available APIs with standard pricing, ensuring no preferential access or influence on the findings. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Model Inventory and Reliability Details

Appendix A.1. Evaluated Large Language Models

Table A1. Evaluated Large Language Models (full inventory).

Vendor	Model Family	Specific Models
OpenAI	GPT-4.1/4o Series	gpt-4.1, gpt-4.1-mini, gpt-4.1-nano, gpt-4o, gpt-4o-mini
OpenAI	o-Series (Reasoning)	o1-mini, o3, o3-mini, o4-mini
Anthropic	Claude Series	claude-3-haiku-20240307, claude-3-5-haiku-20241022, claude-3-5-sonnet-20241022, claude-3-7-sonnet-20250219
Google	Gemini 2.x Series	gemini-2.0-flash, gemini-2.0-pro-exp, gemini-2.0-flash-thinking-exp, gemini-2.0-flash-lite, gemini-2.5-flash-preview-04-17, gemini-2.5-pro-preview-03-25
Google	Gemma 3 Series	gemma-3-27b-it, gemma-3-12b-it, gemma-3-4b-it, gemma-3-1b-it
xAI	Grok Series	grok-2-latest, grok-3-fast-latest, grok-3-latest, grok-3-mini-fast-latest, grok-3-mini-latest
DeepSeek	DeepSeek Series	DeepSeek-V3, DeepSeek-V3-0324, DeepSeek-R1
Meta	Llama Series	Llama-3.3-70B-Instruct-Turbo, Llama-4-Maverick-17B, Llama-4-Scout-17B
Alibaba	Qwen Series	Qwen2.5-VL-7B-Instruct, Qwen3-235B-A22B-FP8

Appendix A.2. Full Missing Data Rates (All Models)

Table A2. Response reliability analysis: missing data rates by model (full table).

Vendor	Model	Overall MDR (%)	Q1	Q2	Q3	Q4
Meta	Llama-4-Scout-17B	35.8	35.8	35.8	35.8	35.8
Alibaba	Qwen3-235B-A22B-FP8	33.3	17.5	17.5	19.2	33.3
Google	gemini-2.5-pro-preview-03-25	17.5	17.5	17.5	17.5	17.5
Google	gemma-3-1b-it	16.7	16.7	16.7	16.7	16.7
Google	gemini-2.0-pro-exp	14.8	14.8	14.8	14.8	14.8
Alibaba	Qwen2.5-VL-7B-Instruct	9.2	0.0	0.0	9.2	0.8
Google	gemini-2.5-flash-preview-04-17	9.2	9.2	9.2	9.2	9.2
Google	gemini-2.0-flash-thinking-exp	6.7	6.7	6.7	6.7	6.7
Google	gemini-2.0-flash	0.8	0.8	0.8	0.8	0.8
OpenAI	o4-mini	0.8	0.8	0.8	0.8	0.8
Anthropic	claude-3-5-haiku-20241022	0.0	0.0	0.0	0.0	0.0
Anthropic	claude-3-5-sonnet-20241022	0.0	0.0	0.0	0.0	0.0
Anthropic	claude-3-7-sonnet-20250219	0.0	0.0	0.0	0.0	0.0
Anthropic	claude-3-haiku-20240307	0.0	0.0	0.0	0.0	0.0
DeepSeek	DeepSeek-R1	0.0	0.0	0.0	0.0	0.0
DeepSeek	DeepSeek-V3	0.0	0.0	0.0	0.0	0.0
DeepSeek	DeepSeek-V3-0324	0.0	0.0	0.0	0.0	0.0
Google	gemini-2.0-flash-lite	0.0	0.0	0.0	0.0	0.0
Google	gemma-3-12b-it	0.0	0.0	0.0	0.0	0.0
Google	gemma-3-27b-it	0.0	0.0	0.0	0.0	0.0
Google	gemma-3-4b-it	0.0	0.0	0.0	0.0	0.0
Meta	Llama-3.3-70B-Instruct-Turbo	0.0	0.0	0.0	0.0	0.0
Meta	Llama-4-Maverick-17B	0.0	0.0	0.0	0.0	0.0
OpenAI	gpt-4.1	0.0	0.0	0.0	0.0	0.0
OpenAI	gpt-4.1-mini	0.0	0.0	0.0	0.0	0.0
OpenAI	gpt-4.1-nano	0.0	0.0	0.0	0.0	0.0
OpenAI	gpt-4o	0.0	0.0	0.0	0.0	0.0
OpenAI	gpt-4o-mini	0.0	0.0	0.0	0.0	0.0
OpenAI	o1-mini	0.0	0.0	0.0	0.0	0.0
OpenAI	o3	0.0	0.0	0.0	0.0	0.0
OpenAI	o3-mini	0.0	0.0	0.0	0.0	0.0
xAI	grok-2-latest	0.0	0.0	0.0	0.0	0.0
xAI	grok-3-fast-latest	0.0	0.0	0.0	0.0	0.0
xAI	grok-3-latest	0.0	0.0	0.0	0.0	0.0
xAI	grok-3-mini-fast-latest	0.0	0.0	0.0	0.0	0.0
xAI	grok-3-mini-latest	0.0	0.0	0.0	0.0	0.0

Appendix A.3. Supplementary Figure: Temperature–Fuzzy Entropy Relationship

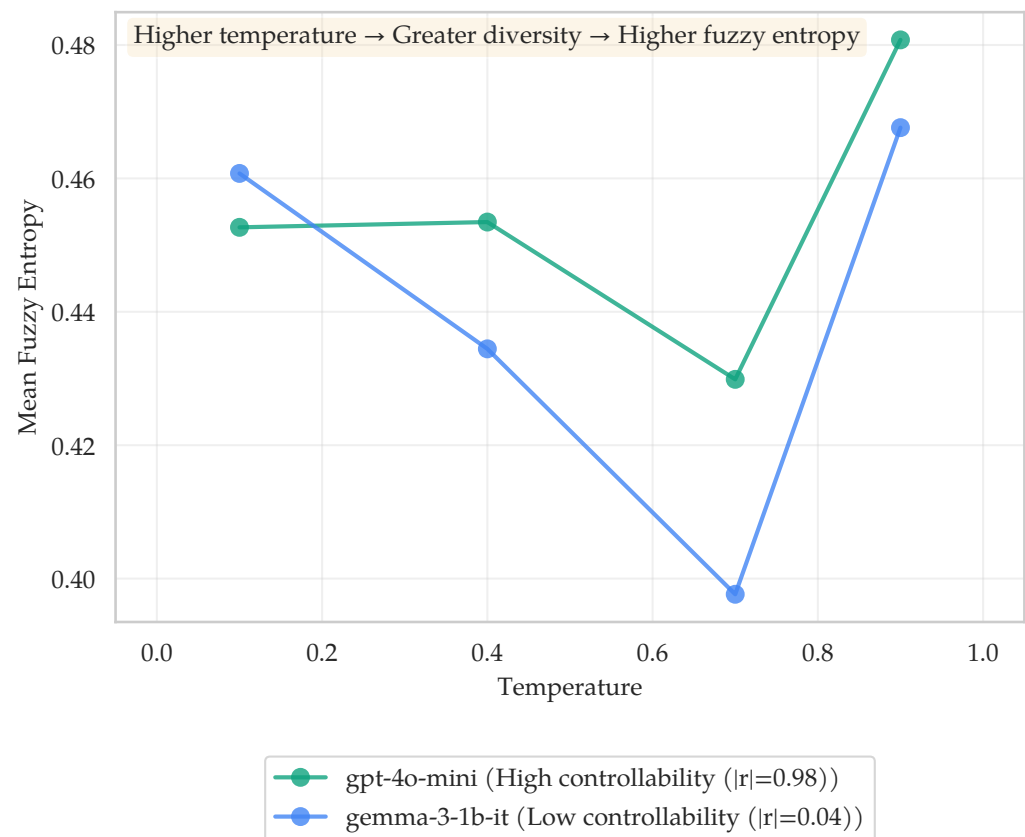


Figure A1. Temperature–fuzzy entropy relationship. Line plot showing mean fuzzy entropy (y-axis) vs. temperature (x-axis, 0.1–0.9) for two representative models at the extremes of temperature controllability.

Appendix A.4. Supplementary Figure: Temperature–Variance Representative Profiles

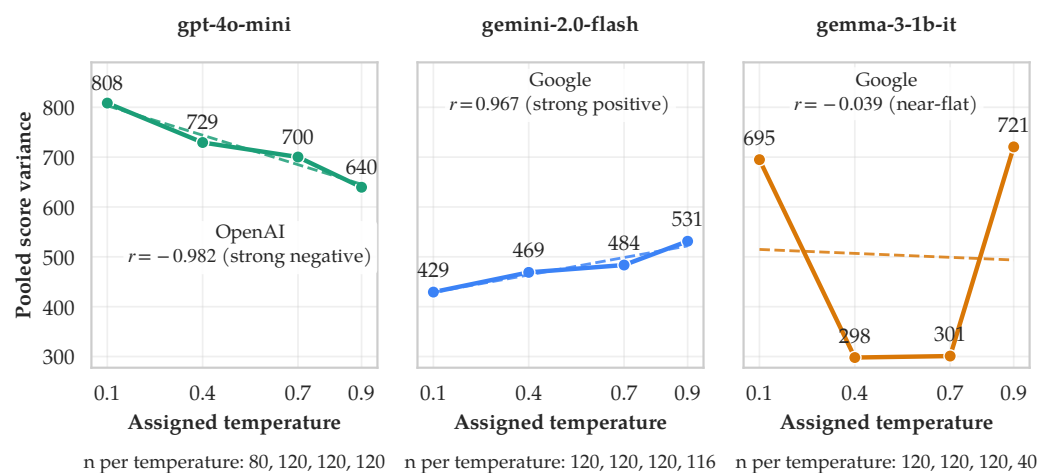


Figure A2. Representative temperature–variance profiles. Three representative models showing strong negative, strong positive, and near-flat temperature–variance relationships. The plotted variance is the pooled sample variance ($ddof = 1$) across all four emotion scores at each tested temperature.

Appendix A.5. Reliability–Controllability Threshold Sensitivity

Table A3. Quadrant counts under nearby threshold choices.

Rho Threshold	Kappa Threshold	Ideal	Stable	Controllable	Avoid	Changed vs. Base
0.90	0.40	21	10	1	4	6
0.90	0.50	18	13	1	4	3
0.90	0.60	15	16	0	5	5
0.95	0.40	19	9	3	5	3
0.95	0.50	16	12	3	5	0
0.95	0.60	15	13	0	8	4
0.98	0.40	19	9	3	5	3
0.98	0.50	16	12	3	5	0
0.98	0.60	15	13	0	8	4

The main-text thresholds are $\rho_M > 0.95$ and $\kappa_M > 0.50$. Nearby alternatives move only a small number of borderline models while leaving the overall structure of the four-quadrant summary intact.

Appendix A.6. Expected Fuzzy Entropy Under Temperature Sampling

Let X denote the (random) emotion score produced by model M at sampling temperature T , and let $\mu(x)$ denote the corresponding fuzzy membership vector (defined in Appendix B.2). The expected fuzzy entropy under (M, T) is

$$\mathbb{E}\big[H_f(\mu(X)) \mid M, T\big] = \sum_{x=0}^{100} H_f(\mu(x)) \Pr(X = x \mid M, T).$$

This identity motivates Figure A1, in which the temperature affects $\Pr(X = x \mid M, T)$, shifting the mass toward overlap regions where $H_f(\mu(x))$ is larger.

Appendix B. Fuzzy Entropy Details

Table A4. Fuzzy entropy analysis: model-level mean entropy (all models).

Vendor	Model	N	$\bar{H}_f^M$	SD
Google	gemma-3-12b-it	120	0.404	0.183
Google	gemini-2.5-flash-preview-04-17	109	0.405	0.067
OpenAI	gpt-4.1-mini	120	0.413	0.175
Google	gemini-2.0-flash-thinking-exp	112	0.416	0.075
Google	gemma-3-27b-it	120	0.429	0.282
Google	gemma-3-1b-it	100	0.435	0.130
Anthropic	claude-3-5-haiku-20241022	120	0.439	0.144
DeepSeek	DeepSeek-V3	120	0.441	0.183
Google	gemini-2.0-flash-lite	120	0.447	0.053
OpenAI	o4-mini	119	0.452	0.069
xAI	grok-2-latest	120	0.454	0.167
Anthropic	claude-3-5-sonnet-20241022	120	0.454	0.132
OpenAI	gpt-4o-mini	120	0.456	0.184
OpenAI	gpt-4.1-nano	120	0.457	0.101
Meta	Llama-4-Maverick-17B	120	0.461	0.182
Meta	Llama-4-Scout-17B	77	0.462	0.125
Anthropic	claude-3-haiku-20240307	120	0.463	0.073
OpenAI	o3-mini	120	0.466	0.123
OpenAI	o1-mini	120	0.466	0.100

Table A4. *Cont.*

Vendor	Model	<i>N</i>	$\bar{H}_f^M$	SD
DeepSeek	DeepSeek-V3-0324	120	0.470	0.224
Google	gemini-2.0-flash	119	0.472	0.278
xAI	grok-3-latest	120	0.476	0.073
xAI	grok-3-mini-fast-latest	120	0.478	0.053
Google	gemini-2.5-pro-preview-03-25	99	0.479	0.093
xAI	grok-3-fast-latest	120	0.479	0.060
xAI	grok-3-mini-latest	120	0.483	0.057
Alibaba	Qwen3-235B-A22B-FP8	80	0.485	0.125
Google	gemini-2.0-pro-exp	23	0.497	0.175
Alibaba	Qwen2.5-VL-7B-Instruct	109	0.513	0.264
OpenAI	gpt-4o	120	0.516	0.080
DeepSeek	DeepSeek-R1	120	0.525	0.061
Meta	Llama-3.3-70B-Instruct-Turbo	120	0.559	0.243
OpenAI	o3	120	0.565	0.076
OpenAI	gpt-4.1	120	0.575	0.189
Google	gemma-3-4b-it	120	0.599	0.145
Anthropic	claude-3-7-sonnet-20250219	120	0.661	0.116

*Appendix B.1. Entropy Interpretation Anchors***Table A5.** Interpretive anchors for the fuzzy-entropy scale used in this paper.

Anchor	Value
Single-assessment theoretical maximum	1.058
Model-level minimum	0.404
Model-level first quartile	0.451
Model-level median	0.466
Model-level third quartile	0.488
Model-level maximum	0.661
Normalized model-level range	0.38–0.63

These anchors are specific to the present membership-function design. The quartiles provide relative, dataset-specific reference points rather than universal cutoffs.

Table A6. Robustness checks for the entropy–reliability relationship across models.

Pearson <i>r</i>	<i>p</i>	95% CI	Spearman ρ_s	<i>p</i>	Bootstrap 95% CI	<i>n</i>
−0.099	0.566	[−0.414, 0.237]	−0.078	0.651	[−0.397, 0.254]	36

Pearson statistics are reported together with a rank-based Spearman robustness check. Both intervals include zero, reinforcing the interpretation that model-level ambiguity and missing-data reliability are largely decoupled in this dataset.

Appendix B.2. Fuzzy Membership Functions

For a scalar emotion score $x \in [0, 100]$, we define three fuzzy sets—Low, Medium, and High—via piecewise-linear membership functions:

$$\mu_{\text{Low}}(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ \max(0, \frac{30-x}{30}) & \text{if } 0 < x \leq 30 \\ 0 & \text{if } x > 30 \end{cases}$$

$$\mu_{\text{Medium}}(x) = \begin{cases} 0 & \text{if } x \leq 10 \\ \frac{x-10}{20} & \text{if } 10 < x \leq 30 \\ 1 & \text{if } 30 < x \leq 50 \\ \frac{70-x}{20} & \text{if } 50 < x \leq 70 \\ 0 & \text{if } x > 70 \end{cases}$$

$$\mu_{\text{High}}(x) = \begin{cases} 0 & \text{if } x < 50 \\ \frac{x-50}{50} & \text{if } 50 \leq x < 100 \\ 1 & \text{if } x = 100 \end{cases}$$

We compute the Shannon-type fuzzy entropy $H_f(\mu(x))$ from these memberships as defined in the main text (Section 2.1.2).

Appendix B.3. Proof Sketch: Information Preservation vs. Hard Discretization

Consider a discrete classification system $\mathcal{K} = \{\text{Low}, \text{Medium}, \text{High}\}$ induced by hard assignment $C = \arg \max_{i \in \{\text{Low}, \text{Medium}, \text{High}\}} \mu_i(x)$, with ties broken by a fixed category order, and decision boundaries at $x = 18$ and $x = \frac{450}{7}$ (where $\mu_{\text{Low}} = \mu_{\text{Medium}}$ and $\mu_{\text{Medium}} = \mu_{\text{High}}$, respectively). Two assessments, $x_1 = 64$ and $x_2 = 65$ are classified into different categories despite minimal differences. In the fuzzy framework:

$$\mu(64) = [0, 0.300, 0.280], \quad \mu(65) = [0, 0.250, 0.300]$$

The Euclidean distance $\|\mu(64) - \mu(65)\| \approx 0.054$ correctly reflects the small semantic difference, whereas the discrete classification imposes an artificial discontinuity. Since C is a deterministic coarsening of $\mu(X)$, the data processing inequality implies that

$$I(X; \mu(X)) \geq I(X; C).$$

This information-theoretic advantage extends to LLM assessment tasks in which the emotional intensity varies continuously.

Appendix C. Persona Analysis Details

Table A7. Principal component analysis results.

Comp.	λ_i	Var.%	Cum.%	Interpretation
PC1	10.33	63.9	63.9	Shared Sadness/ Anger vs. Interest axis
PC2	4.49	27.8	91.7	Surprise-dominant shared axis
PC3	0.61	3.8	95.5	Anger/Interest vs. Surprise/Sadness contrast
PC4	0.33	2.0	97.5	Residual variance

Comp. = Component; λ_i = Eigenvalue; Var.% = Variance Explained (%); Cum.% = Cumulative Variance (%).

Table A8. PC3 loadings by persona and emotion.

Persona	Int.	Sur.	Sad.	Ang.	Mean
P1 (Student, $T = 0.7$)	0.32	−0.19	−0.07	0.34	0.10
P2 (Researcher, $T = 0.4$)	0.30	−0.23	−0.08	0.34	0.08
P3 (Poet, $T = 0.9$)	0.27	−0.17	−0.05	0.34	0.10
P4 (Robot, $T = 0.1$)	0.28	−0.19	−0.01	0.37	0.11

Int. = Interest; Sur. = Surprise; Sad. = Sadness; Ang. = Anger; Mean = Mean Loading across emotions.

Table A9. Two-way ANOVA: Vendor $\times$ persona interaction (interest dimension).

Effect	SS	df	MS	F	p-Value	η_p^2
Vendor	103,345	6	17,224	34.8	<0.001	0.049
Persona	15,527	3	5176	10.4	<0.001	0.008
Vendor $\times$ Persona	7641	18	424	0.86	0.633	0.004
Error	2,000,914	4039	495	—	—	—

SS = Sum of Squares; df = degrees of freedom; MS = Mean Square; η_p^2 = partial eta-squared effect size.

Table A10. Poetvs. robot: comprehensive statistical comparison.

Metric	P3 (Poet)	P4 (Robot)	Diff.	Cohen's <i>d</i>	p-Value
Interest					
Mean	66.3	62.2	+4.1	0.17	<0.001
SD	23.5	23.9	−0.4	—	—
Median	75	70	+5	—	—
Surprise					
Mean	61.1	54.6	+6.5	0.32	<0.001
SD	19.5	21.6	−2.0	—	—
Sadness					
Mean	35.6	32.3	+3.3	0.12	<0.001
SD	28.4	27.7	+0.6	—	—
Anger					
Mean	35.9	32.6	+3.3	0.13	0.0028
SD	26.7	25.4	+1.3	—	—

Diff. = Difference (P3–P4); SD = Standard Deviation. Cohen's *d* shown for mean differences; *p*-values from Mann–Whitney U tests.

Appendix D. Clustering and Consistency Details

Table A11. Intra-cluster and inter-cluster distances.

Comparison	Mean Distance	SD	Interpretation
Within Cluster 1	4.2	1.8	Very cohesive
Within Cluster 2	2.3	0.9	Extremely cohesive
Within Cluster 3	8.7	3.4	Moderately cohesive
Within Cluster 4	11.4	4.7	Diffuse
Within Cluster 5	15.8	6.2	Very diffuse
Between Clusters (avg)	23.6	8.9	Well-separated

Table A12. Consistency rankings (numerical stability proxy).

Rank	Model	Vendor	Score	Interpretation
1	gemma-3-1b-it	Google	0.780	Good
2	claude-3-haiku-20240307	Anthropic	0.759	Good
3	o1-mini	OpenAI	0.728	Acceptable
4	o3-mini	OpenAI	0.726	Acceptable
5	gemma-3-4b-it	Google	0.712	Acceptable
		⋮		
32	Qwen2.5-VL-7B-Instruct	Alibaba	0.599	Lower
33	gemini-2.0-pro-exp	Google	0.592	Lower

Table A12. Cont.

Rank	Model	Vendor	Score	Interpretation
34	gemini-2.5-pro-preview-03-25	Google	0.582	Lower
35	gemini-2.0-flash-thinking-exp	Google	0.548	Lowest
36	gemini-2.5-flash-preview-04-17	Google	0.548	Lowest

Score = Consistency Score. Top 5 and bottom 5 models shown; middle ranks (6–31) omitted for brevity.

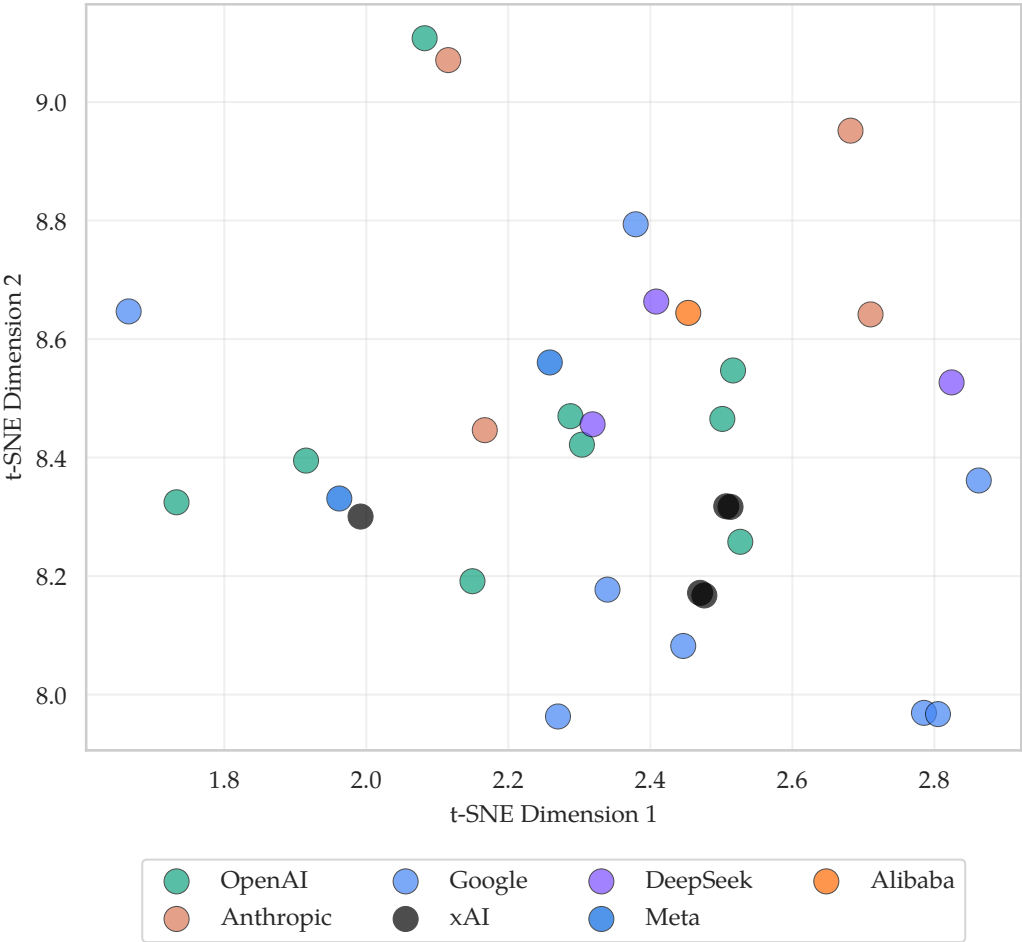


Figure A3. t-SNE visualization of LLM emotion space. 2D scatter plot with each point representing one LLM (48-dimensional emotion-assessment space projected via t-SNE).

Table A13. Vendor-level consistency analysis (numerical stability proxy).

Vendor	Mean	SD	Median	Range
Anthropic	0.693	0.054	0.691	0.630–0.759
Meta	0.678	0.020	0.680	0.656–0.697
OpenAI	0.671	0.039	0.664	0.613–0.728
DeepSeek	0.660	0.036	0.675	0.619–0.685
Google	0.639	0.077	0.630	0.548–0.780
xAI	0.630	0.029	0.622	0.610–0.681
Alibaba	0.629	0.042	0.629	0.599–0.659

Mean = Mean Consistency Score; SD = Standard Deviation; Range shows min–max values.

Appendix E. Text-Dependent Correlation and Variance Details

Table A14. Text-Specific Emotion Correlations (Interest vs. Sadness).

Text	Genre	$r_{\text{Int-Sad}}$	p -Value	Interpretation
T1: Pocket Watch	Allegorical	−0.19	<0.001	Weak negative
T2: Money & Pistol	Narrative	−0.70	<0.001	Strong negative
T3: Ragged Ostrich	Poetic	−0.68	<0.001	Strong negative

95% confidence intervals (Fisher z-transform): T1 [−0.25, −0.14]; T2 [−0.73, −0.67]; T3 [−0.71, −0.65].

Table A15. Kruskal–Wallis H-tests for text effects.

Emotion	H-Statistic	ϵ^2	Effect Size
Interest	$H = 2350.56, p < 0.001$	0.578	Very large
Surprise	$H = 2312.38, p < 0.001$	0.568	Very large
Sadness	$H = 2346.97, p < 0.001$	0.577	Very large
Anger	$H = 2333.39, p < 0.001$	0.574	Very large

$\epsilon^2 := (H - k + 1) / (n - k)$, where k is the number of groups and n is the total sample size.

Table A16. Emotional tension analysis across texts.

Emotion Pair	T1 (Allegorical)	T2 (Narrative)	T3 (Poetic)	Interpretation
Interest–Sadness	−0.19	−0.70	−0.66	High tension in T2, T3
Interest–Anger	−0.07	−0.56	−0.51	High tension in T2, T3
Surprise–Sadness	0.57	−0.01	0.13	No tension (non-negative association)
Surprise–Anger	0.33	−0.15	0.21	Weak tension only in T2

Values are Pearson correlation coefficients. Tension is defined via negative correlation.

Table A17. Variance components (interest dimension).

Source	Variance Component	% of Total	Interpretation
Text	292.1	55.9%	Primary factor
Model	67.3	12.9%	Secondary factor
Persona	3.5	0.7%	Minor factor
Residual	160.1	30.6%	Unexplained

Marginal variance attributions computed from one-way sums of squares on complete cases.

References

1. Paech, S.J. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models. *arXiv* **2024**, arXiv:2312.06281v2. [CrossRef]
2. Huang, J.; Lam, M.H.; Li, E.J.; Ren, S.; Wang, W.; Jiao, W.; Tu, Z.; Lyu, M.R. Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench. *arXiv* **2024**, arXiv:2308.03656v2. [CrossRef]
3. Sabour, S.; Liu, S.; Zhang, Z.; Liu, J.; Zhou, J.; Sunaryo, A.; Lee, T.M.C.; Mihalcea, R.; Huang, M. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; Volume 1, pp. 5986–6004.
4. Chen, Y.; Yan, S.; Liu, S.; Li, Y.; Xiao, Y. EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models. *Find. Assoc. Comput. Linguist. ACL 2024* **2024**, 2149–2176. [CrossRef]
5. Sosea, T.; Caragea, C. Hard Emotion Test Evaluation Sets for Language Models. *Find. Assoc. Comput. Linguist. NAACL 2025* **2025**, 7945–7959. [CrossRef]
6. Pereira, P.; Moniz, H.; Carvalho, J.P. Deep emotion recognition in textual conversations: A survey. *Artif. Intell. Rev.* **2025**, *58*, 10. [CrossRef]
7. Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; Kabbara, J. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *Find. Assoc. Comput. Linguist. NAACL 2024* **2024**, 3605–3627. [CrossRef]
8. Tseng, Y.-M.; Huang, Y.-C.; Hsiao, T.-Y.; Chen, W.-L.; Huang, C.-W.; Meng, Y.; Chen, Y.-N. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. *Find. Assoc. Comput. Linguist. EMNLP 2024* **2024**, 16612–16631. [CrossRef]

9. Cheng, J.; Sabour, S.; Sun, H.; Chen, Z.; Huang, M. PAL: Persona-Augmented Emotional Support Conversation Generation. *Find. Assoc. Comput. Linguist. ACL 2023* **2023**, 535–554. [[CrossRef](#)]
10. Jun, Y.; Lee, H. Exploring Persona Sentiment Sensitivity in Personalized Dialogue Generation. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, 27 July–1 August 2025; pp. 18384–18402.
11. Wu, S.; Zhu, Y.; Hsu, W.; Lee, M.-L.; Deng, Y. From Personas to Talks: Revisiting the Impact of Personas on LLM-Synthesized Emotional Support Conversations. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 5439–5453.
12. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 30 April 2020.
13. Zadeh, L.A. Fuzzy sets. *Inf. Control* **1965**, *8*, 338–353. [[CrossRef](#)]
14. De Luca, A.; Termini, S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Inf. Control* **1972**, *20*, 301–312. [[CrossRef](#)]
15. Yuan, X.; Zheng, C. Improved Intuitionistic Fuzzy Entropy and Its Application in the Evaluation of Regional Collaborative Innovation Capability. *Sustainability* **2022**, *14*, 3129. [[CrossRef](#)]
16. Kumar, S.; Arya, V.; Kumar, S.; Dahiya, A. A New Picture Fuzzy Entropy and Its Application Based on Combined Picture Fuzzy Methodology with Partial Weight Information. *Int. J. Fuzzy Syst.* **2022**, *24*, 3208–3225. [[CrossRef](#)]
17. Thakur, P.; Kizielewicz, B.; Gandotra, N.; Shekhovtsov, A.; Saini, N.; Salabun, W. The Group Decision-Making Using Pythagorean Fuzzy Entropy and the Complex Proportional Assessment. *Sensors* **2022**, *22*, 4879. [[CrossRef](#)] [[PubMed](#)]
18. Shirahama, N.; Nakaya, N.; Watanabe, S. Development of an Experimental Platform for Evaluating LLM Emotion Understanding. *ICIC Express Lett.* **2026**, *20*, 191–198. [[CrossRef](#)]
19. Shirahama, N.; Nakaya, N.; Watanabe, S. Characterizing “Emotional Individuality” in 36 LLMs: Persona and Japanese Text Conditioning. In Proceedings of the 21st International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2025), Taichung, Taiwan, 15–17 October 2025.
20. Shirahama, N.; Nakaya, N.; Watanabe, S. Temperature-Controlled Response Reliability in Large Language Models: A Quantitative Analysis of 36 Models for Emotion Understanding Tasks. In Proceedings of the 19th International Conference on Innovative Computing, Information and Control (ICISIP 2025), Nagasaki, Japan, 8–9 September 2025; pp. 78–84.
21. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [[CrossRef](#)]
22. Plutchik, R. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **2001**, *89*, 344–350. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.