taL(LM)ent Show

A Comparative Study of Prompting Strategies in Creative Generation Tasks

Paola Loi
MSc in Data Science for Economics
University of Milan

October 15, 2025

Abstract

This report introduces *taL(LM)ent show*, a narrative-driven evaluation framework designed to assess how large language models balance structure, style, and creativity under explicit prompting constraints. Five instruction-tuned models are compared (GPT-2 Large, Falcon-1B Instruct, LLaMA-3.2-3B Instruct, Mistral-7B Instruct and Hermes-2-Pro-Mistral-7B) across four prompting modes.

Two fictional texts serve as test pieces, evaluated by stylized judges and summarized by a presenter. Quantitative analyses include parsing validity, entropy, lexical diversity, and confidence margins, while qualitative inspection focuses on stylistic fidelity, emotional calibration, and persona coherence. Token-level saliency experiments further reveal which prompt elements most influence model behavior.

Results show a clear progression from unaligned to aligned systems: smaller models follow instructions mechanically, while larger aligned ones integrate format and semantics coherently. Among them, Hermes-2-Pro-Mistral-7B demonstrates the most balanced performance, marking the current benchmark for controlled creative generation.

Disclaimer: The optimization of this project made use of GPT-5 for code refinement, JSON generation workflow design, and visualization output creation within the Google Colab environment. All experiments were executed using Google Colab Pro, leveraging GPU acceleration for efficient model inference.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Prompt engineering has emerged as a central practice for shaping the behavior of large language models, but its effectiveness is still debated. Some strategies, such as role-based or emotion-focused prompting, can influence creativity and engagement, while others, like persona-driven instructions, often yield limited impact. To explore these dynamics, this project adopts a narrative "talent show" metaphor, where controlled experiments are framed in a creative context. This approach allows to study how prompts shape both the content and the evaluative dimensions of model outputs.

1.2 Objectives of the Talent Show Experiment

The goal of this project is to compare prompting strategies in a systematic yet creative way. Within the *taL(LM)ent show*, fictional artists perform, stylized judges evaluate their works, and a presenter synthesizes the outcomes.

The experimental design combines multiple prompting modes (zero-shot, role-based, emotion-role, and presenter wrap-ups) with a quantitative–qualitative dual analysis. Quantitatively, we measure structural validity, confidence metrics (average log-probability, margin, and entropy), lexical diversity (distinct-2), and output length. Qualitatively, we analyze stylistic fidelity, emotional calibration, and token-level saliency to understand which prompt components genuinely guide model behavior.

1.3 Research Questions

From this setup, the project addresses four main questions:

- **Prompt sensitivity:** Do role-based or emotion-focused prompts genuinely change outputs, or only alter their style?

- **Decoding strategies:** How do "creative" vs. "controlled" parameters affect both generated texts and assigned judgments?
- **Interpretability:** Do saliency and confidence metrics show models focusing on meaningful tokens, or on irrelevant details?
- **Evaluation framework:** Can combining narrative evaluations with metrics provide a fuller view of creativity and coherence?

Chapter 2

Prompt Engineering in Context

2.1 Overview of Prompting Strategies

Prompt engineering has evolved into a key practice for shaping the reasoning and stylistic behavior of Large Language Models (LLMs) (Schulhoff et al., 2024). This project adopts a compact yet representative set of prompting strategies designed to test how different instruction structures affect model outputs. Three principal modes are implemented:

- **Zero-Shot** – the model evaluates a text without any role specification. This baseline reveals how the model interprets the task in the absence of stylistic or affective cues, reflecting its default interpretative priors.
- **Role-Based** – the model assumes the persona and evaluative tone of a specific judge. This conditioning alters tone, focus, and lexical choice, making it possible to study how persona alignment interacts with instruction following (Zheng et al., 2024).
- **Emotion-Role (Focused Role)** – similar to the role-based mode, but with an explicit emphasis on emotional expression. It tests whether affective conditioning can bias the model toward more expressive or original formulations (Li et al., 2023).

In addition to these, the experiment introduces a dedicated **narrative synthesis task**, the *Presenter Wrap-Up*, where the model acts as a show host summarizing the judges' remarks. Although not a separate prompting family, it evaluates the model's ability to perform *compositional reasoning* and multi-voice synthesis within a structured format.

Each prompting mode is later combined with two distinct decoding regimes (*creative* vs. *controlled*). Together, these conditions form a factorial design that isolates the respective effects of instruction structure and sampling stochasticity on output quality, coherence, and interpretative depth.

2.2 Creative Generation and the Talent Show Framework

Recent research has shown that evaluating LLMs on creative or affective tasks provides valuable insight into their alignment and reasoning mechanisms (Li et al., 2023; Ouyang et al., 2022). According to Schulhoff et al. (2024), prompt design plays a central role in determining the interpretive behavior of language models. Li et al. (2023) further demonstrate that emotional framing enhances perceived originality and vividness, suggesting that models are responsive to affective cues beyond purely semantic control. However, Zheng et al. (2024) caution that persona-based prompting may alter surface tone without necessarily improving reasoning depth or factual accuracy.

Building on these findings, the ***taL(LM)ent show*** provides a structured yet playful testbed for creative generation. It reinterprets prompting as a multi-role dialogue: *artists* (texts) evoke meaning, *judges* (prompts) deliver stylistically differentiated evaluations, and a *presenter* synthesizes them into a coherent summary. This metaphor operationalizes the key dimensions of prompting (role, emotion, and synthesis) within a reproducible and interpretable framework.

By repeating identical "show rounds" across models and decoding parameters, the experiment measures how different architectures balance structure and creativity. The format functions both as a narrative device and an empirical control, linking stylistic expressiveness to measurable quantities such as coherence, entropy, and structural adherence (Wolf et al., 2020).

Chapter 3

Experimental Narrative Design

3.1 Artists and Performances

The experiment features two fictional "artists," each embodied by a distinct text type. The first, **Salmo**, represents contemporary songwriting through the lyrics of an Italian rap piece (*Il cuore all'equatore, la testa all'Antartide...*), direct and rhythmic in tone. The second, **Lorenzo de' Medici**, embodies the Renaissance poetic tradition via the well-known lines *Quant'è bella giovinezza...*, reflective and symbolic in character.

Their sharp stylistic and temporal contrast forms a robust testing ground for prompting strategies. Models must adapt evaluative reasoning to texts differing in language, register, and intent: the first relies on immediacy and rhythm, the second on symbolic interpretation and cultural awareness (Li et al., 2023; Ouyang et al., 2022).

3.2 Judges and Presenter Roles

Evaluation is carried out by three fictional "judges," each embodying a distinct persona and interpretative lens:

- **Gio Evan** – poet-songwriter with introspective, lyrical expression emphasizing emotion and authenticity.
- **Pucci** – comedian using humorous, colloquial tone focused on everyday absurdities.
- **Ursula von der Leyen** – political figure with an institutional tone, valuing social and cultural depth.

These personas operationalize **role-based prompting**: the model must align in both content and register (Schulhoff et al., 2024). While personas may not enhance reasoning, they strongly shape framing and stylistic coherence—crucial in creative evaluation (Zheng et al., 2024).

A **presenter role** is introduced as synthesis layer: the model summarizes judges' opinions into a coherent overview, testing abstraction and integrative reasoning. This

task merges diverse perspectives while maintaining tonal and structural consistency (Wolf et al., 2020). The presenter thus represents the upper level of the experimental hierarchy, assessing not content alone but meta-organization of collective judgment.

3.3 Evolution of Rounds and Rule Changes

The experiment unfolds in "rounds," each corresponding to a specific combination of prompting mode and decoding setup. It begins with zero-shot baselines, followed by role-based and emotion-focused prompts imposing stylistic and affective constraints. Each cycle concludes with a presenter wrap-up summarizing the judges' perspectives.

Decoding parameters alternate between **creative** (high temperature, broad sampling) and **controlled** (low temperature, narrow sampling), testing how stochasticity affects the balance between expressiveness and structure (Ouyang et al., 2022). This progression mirrors prompt engineering protocols: starting from minimal guidance, adding contextual conditioning, and ending with synthesis (Schulhoff et al., 2024).

Chapter 4

Technical Framework

4.1 Model Selection Rationale

The experiment uses a compact yet diverse suite of instruction-tuned models, balancing scale, alignment, and computational feasibility. All are open-access and compatible with token-level probability extraction via the `transformers` API, ensuring full comparability (Wolf et al., 2020).

The lineup spans parameter ranges and alignment depths:

- **GPT-2 Large** — pre-alignment baseline illustrating limits of unconstrained generation.
- **Falcon 3 1B Instruct** — lightweight model probing prompt sensitivity at small scale.
- **LLaMA 3.2 3B Instruct** — compact model with strong structural regularity.
- **Mistral 7B Instruct (v0.2)** — mid-size, stable, and stylistically expressive.
- **Hermes-2-Pro-Mistral-7B** — advanced variant with enhanced alignment, representing the upper bound of compositional control.

All models ran in **Google Colab Pro** on an NVIDIA **T4 GPU (16 GB VRAM)**. This constraint guided model choice: Hermes-2-Pro-Mistral-7B provided a high-performing yet locally runnable option. The setup balances expressive capacity with reproducibility.

4.2 Technical Details of the Models

Models were accessed via Hugging Face with a standardized inference interface ensuring consistent tokenization, logits extraction, and sampling. Per-step probability distributions were retrieved using `generate(..., output_scores=True)` to compute log-probabilities, entropy, and confidence margins.

- **Falcon 3 1B Instruct, LLaMA 3.2 3B, GPT-2 Large** — loaded with `AutoTokenizer` /`AutoModelForCausalLM`, using `torch.bfloat16` precision and `device_map="auto"`.
- **Mistral 7B Instruct (v0.2)** — loaded in 4-bit quantization (`bnb_4bit_nf4`, double quantization, `bfloat16` compute dtype).
- **Hermes-2-Pro-Mistral-7B** — same scheme, with *CPU/disk offloading* to fit VRAM. Both Mistral variants use inference mode (`model.eval()`) to disable stochastic effects.

All models share one decoding interface for fair comparison; outputs were logged for quantitative and saliency analyses. This ensures reproducibility and fine-grained behavioral observation.

4.3 Decoding Parameters

Models were tested under two standardized configurations—**creative** and **controlled**—to probe the trade-off between diversity and stability. Generation was stochastic (`do_sample=True`), with identical parameter sets across models for comparability; reproducibility can be restored with `torch.manual_seed(#)`.

Setting	`max_new_tokens`	`temperature`	`top_p`	`top_k`
Creative	300	0.8	0.95	50
Controlled	300	0.3	0.85	20

The **Creative** regime promotes expressive diversity and higher entropy; the **Controlled** one favors precision and coherence. `Temperature` regulates randomness, `top_p` defines the cumulative probability mass (nucleus sampling), and `top_k` restricts sampling to the top k tokens. A uniform cap on `max_new_tokens` (300, with safeguard 200) keeps outputs comparable and avoids runaways. These configurations reveal whether stylistic richness and structural control can coexist within one framework.

4.4 Generation and Logging Pipeline

A unified pipeline standardizes prompts, decoding, and metric collection across models. Each run combines a text (Salmo or Lorenzo de' Medici), a judge persona (Gio Evan, Pucci, Ursula), and a prompting mode (`zero_shot`, `role`, `emotion_role_high`, `presenter_wrapup`), ensuring consistent communicative contexts.

Each cycle performs four operations:

1. **Prompt assembly** — builds prompt strings from text and persona templates; presenter prompts run once per text.

2. **Generation and scoring** — executes generation with token-level logits; computes *log-probability*, *surprisal*, *entropy*, and *margin*.
3. **Post-processing** — parses outputs into JSON, logging validity, token length, and lexical diversity (`distinct-2`).
4. **Aggregation** — stores all metrics and metadata in a `pandas.DataFrame` for later analysis.

This structure allows macro-level comparison and micro-level diagnostics while maintaining replicability and transparency.

4.5 Prompt-Level Saliency

To complement metrics, saliency analysis measures how much each prompt token influences behavior. A **leave-one-token-out** strategy removes one token per run and measures the log-probability drop relative to the full prompt.

Formally:

$$S(t_i) = \log p_{\text{full}} - \log p_{\text{no } t_i}$$

Higher $S(t_i)$ indicates stronger influence on model confidence.

This reveals whether attention centers on **semantic anchors** (e.g., role names, emotion cues) or **instruction keywords** ("output must be JSON"). Comparing across models shows a progression from low-level syntactic saliency in smaller models to high-level conceptual saliency in aligned ones, reflecting the transition from surface obedience to semantic understanding.

Chapter 5

Results and Comparative Findings

5.1 Overview

This chapter presents the empirical findings of the experiment, integrating quantitative indicators with qualitative inspection. For each model, the discussion highlights (i) format compliance, (ii) expressive behavior across prompting modes, and (iii) interpretive stability under controlled and creative decoding. Rather than listing metrics in isolation, the goal is to identify consistent behavioral signatures that reveal how alignment, scale, and prompting interact in practice.

5.2 LLaMA 3.2 3B Instruct

LLaMA 3B shows strong structural discipline despite its compact size. Across all prompting modes, it achieves nearly perfect JSON compliance (`jsonvalid` ~1.0), confirming that instruction finetuning can enforce syntactic obedience even in small models.

Setting	`avg_logprob`	`avg_margin`	`entropy`	`distinct-2`
Controlled	−0.08	0.92	0.09	0.73
Creative	−0.35	0.82	0.35	0.73

LLaMA's high margins and low entropy (Table: Controlled) indicate confident, deterministic decoding, while the creative configuration increases lexical variety without structural collapse. Saliency results show strong reliance on surface anchors ("must", "ONLY") rather than conceptual tokens, suggesting that its robustness stems from syntactic adherence more than semantic understanding. In short, LLaMA behaves as a *compact but well-aligned* model: structurally flawless, moderately expressive, and sensitive to temperature modulation.

5.3 Falcon 3 1B Instruct

Falcon 1B exposes the lower limit of controllability under instruction tuning. Despite comparable decoding parameters, it frequently breaks format and repeats prompt text verbatim.

Setting	`avg_logprob`	`avg_margin`	`entropy`	`json_valid`
Controlled	−0.10	0.92	0.10	0.45
Creative	−0.61	0.73	0.62	0.40

Takeaway. The steep drop in margin and rise in entropy (Controlled → Creative) show that Falcon's generation becomes unstable as sampling broadens. Its saliency map concentrates almost exclusively on instruction markers ("IMPORTANT", "Provide"), evidencing surface-level obedience without internal integration. Falcon therefore functions as a *procedural generator*—able to mimic format under low randomness but unable to maintain consistency once exploratory sampling is introduced.

5.4 Mistral 7B Instruct v0.2

Mistral 7B combines scale, architecture, and stable alignment to achieve a consistent balance between structure and expressiveness. Its JSON validity (~0.8–0.85) and moderate entropy reveal disciplined but flexible behavior.

Setting	`avg_logprob`	`avg_margin`	`entropy`	`distinct-2`
Controlled	−0.24	0.93	0.18	0.69
Creative	−0.49	0.82	0.42	0.75

Mistral's entropy nearly doubles between modes, yet JSON validity and role coherence remain stable—evidence of *controlled exploration*. Its outputs show stylistic precision (clear role differentiation) and semantic calibration (emotion tags match the text content). Saliency analysis confirms balanced sensitivity between instruction cues and semantic anchors (judge names, emotional terms), showing genuine task integration. Mistral thus represents a pivotal threshold: the smallest model capable of sustaining both structure and creativity.

5.5 Hermes-2-Pro-Mistral-7B

Hermes-2-Pro, a refined version of Mistral, delivers the most coherent and expressive performance across all modes. It maintains near-perfect JSON validity (0.88–0.93) and combines low entropy with high stylistic fidelity.

Setting	avg_logprob	avg_margin	entropy	distinct-2
Controlled	−0.23	0.97	0.14	0.68
Creative	−0.46	0.84	0.33	0.74

Hermes exhibits consistently high margins and low entropy (Controlled: 0.97, 0.14), reflecting *confidence with restraint*. Even at higher temperature, stylistic diversity increases without degrading structural integrity. Saliency shifts from format tokens ("JSON", "Provide") to conceptual ones ("emotion", "summary"), signaling semantic grounding rather than formal mimicry. Hermes thus embodies *disciplined creativity*—an alignment-driven integration of precision, tone, and adaptability.

5.6 GPT-2 Large

GPT-2 Large, pre-dating instruction tuning, serves as a historical control for unaligned generative behavior. It generates fluent and coherent text but disregards all structural constraints.

Setting	avg_logprob	avg_margin	entropy	json_valid
Controlled	−0.71	0.69	0.58	0.44
Creative	−0.89	0.61	0.72	0.36

GPT-2's high entropy and low margins across both settings demonstrate *unregulated stochasticity*. It captures meaning locally ("The poem feels nostalgic...") but fails to translate understanding into structured action (JSON validity < 0.5). Saliency remains diffuse, with attention peaks on instruction keywords but no consistent semantic weighting. GPT-2 thus represents the *pre-alignment baseline*, a model fluent in language but not in task adherence.

5.7 Cross-Mode and Decoding Dynamics

When aggregating metrics across prompting modes and decoding regimes, several patterns emerge that clarify how scale and alignment interact with prompt design.

Mode-dependent patterns

The four prompting modes elicit distinct behavioral profiles:

- **Zero-shot:** consistently the most unstable setup, with high entropy and low JSON validity across all models. For small architectures (Falcon, GPT-2), zero-shot often regresses into prompt repetition; for aligned mid-size models (LLaMA, Mistral), it remains coherent but less expressive.

- **Role:** the most stable mode overall, maximizing stylistic differentiation while keeping entropy low. Both Mistral and Hermes show clear persona fidelity and consistent structure here—this mode acts as their "sweet spot" for balanced performance.
- **Emotion-role-high:** increases entropy moderately while boosting affective vocabulary. Mistral and Hermes use this mode meaningfully (emotion terms are context-appropriate), whereas Falcon and GPT-2 overproduce generic intensifiers ("very emotional", "strong feelings").
- **Presenter:** the most challenging mode. Smaller models fail completely (JSON invalidity $> 50\%$), LLaMA handles basic summarization, and Hermes uniquely succeeds in integrating the judges' tones and maintaining perfect structure.

These contrasts reveal that prompting modes test complementary dimensions of alignment: **Zero-shot** evaluates generalization, **Role** tests stylistic conditioning, **Emotion-role-high** probes semantic calibration, and **Presenter** measures synthesis under compositional load. Only models above the 7B threshold display robustness across all four, confirming that alignment depth, not just parameter count, governs adaptability.

The comparison between controlled and creative decoding clarifies how each model balances determinism and variability.

Parameter Set	`avg_margin`	`entropy`	`distinct-2`
Controlled	High (0.90–0.97)	Low (0.10–0.20)	Moderate (0.68–0.72)
Creative	Lower (0.70–0.85)	Higher (0.30–0.70)	Slightly Higher (0.73–0.76)

- **GPT-2 and Falcon** show the steepest entropy increases under creative decoding, leading to loss of structure and frequent instruction echoes. Their rise in `distinct-2` (+0.07 on average) reflects noise, not genuine lexical innovation.
- **LLaMA 3B** shows mild entropy shifts but stable margins, indicating controlled variability. It expands phrasing diversity while maintaining JSON validity at 100%.
- **Mistral 7B** increases entropy by +0.24 yet retains strong margins (0.93→0.82), showing "elastic control": greater linguistic freedom without collapse. This supports the hypothesis that model alignment can absorb stochasticity.
- **Hermes-2-Pro 7B** exhibits the most stable trade-off: entropy rises only +0.19 while maintaining a margin near 0.85 in creative mode. Its lexical diversity increase is meaningful, not random—outputs vary stylistically but stay structurally flawless.

These dynamics demonstrate that creativity does not emerge from randomness per se. In smaller or unaligned models, higher temperature amplifies sampling noise; in aligned

systems, it enables controlled stylistic modulation. Hermes and Mistral illustrate how instruction-tuned alignment transforms temperature from a source of instability into a driver of expressive precision.

5.8 Qualitative Insights

The qualitative analysis reveals how prompt design, alignment, and decoding jointly influence interpretive and stylistic control.

Well-aligned models such as **Mistral 7B** and **Hermes-2-Pro** exhibit coherent persona fidelity and emotion calibration, reshaping probabilities beyond surface imitation. Their outputs reflect *goal-conditioned understanding* rather than pattern matching. For instance:

> **Prompt (Pucci, role mode):** "Evaluate this poem as a humorous judge."
> **Hermes output:** "It's so serious it becomes funny—like wearing a tuxedo at the beach." **Mistral output:** "A refined joke disguised as poetry—elegant but ironic."

Both models adapt register and tone consistently, showing internalized style control. By contrast, smaller models (**Falcon 1B**, **GPT-2**) echo instruction text (*"Include only JSON. . . "*) or inject random adjectives, revealing that prompt hardening is effective only when backed by representational grounding.

Emotion-focused prompts deepen this contrast. Hermes integrates emotion semantically ("*warm melancholy through rhythm and imagery*"), while Falcon merely attaches affective tags ("*emotion: high, mood: sad*"). Similarly, in presenter mode, Hermes uniquely fuses the judges' tones into a coherent synthesis—something smaller models fail to approximate.

Takeaway. Alignment transforms prompts from lexical constraints into internalized behavioral rules: models like Hermes **perform** instructions, while smaller ones merely **quote** them.

5.9 Prompt-Level Saliency

Saliency scores quantify which prompt tokens most influence generation. For each token t, we measure the drop in average log-probability $\Delta s(t)$ when t is removed (leave-one-token-out). Higher Δs indicates stronger causal importance.

Patterns across models

- **Falcon 1B / GPT-2:** highest saliency on *format anchors* ("ONLY", "IMPORTANT", "JSON"), yet still fail structurally—suggesting lexical awareness without operational understanding.

- **Mistral 7B:** balanced attention between *format* (e.g., "judge", "emotion") and *semantic anchors* (artist, role, tone). This distribution aligns with its stable yet creative outputs.
- **Hermes-2-Pro:** jointly content- and structure-aware. Tokens related to meaning (*emotion*, *judge name*) carry comparable saliency to structural cues.

Hermes case studies

Table 5.1 reports the top salient tokens from two probes.

Table 5.1: Top salient tokens for Hermes-2-Pro-Mistral-7B (leave-one-token-out).

Scenario	Type	Token	Saliency	Interpretation
Salmo / Gio Evan / emotion_role_high	Format	JSON:	0.23	Structure anchor for output validity.
	Content	d'estate	0.58	Links emotional tone to imagery.
	Content	all'Antartide	0.39	Grounds affect in lyric context.
	Content	casa	0.34	Reinforces theme of belonging.
Lorenzo / Presenter_wrapup	Format	IMPORTANT:	0.77	Schema instruction (role cue).
	Role	Pucci	0.28	Anchors persona synthesis.
	Meta	Three	0.26	Encodes number of judges.
	Format	JSON	0.16	Maintains structure compliance.

Hermes distributes saliency across both *content* and *schema* tokens. This evidences deep integration: format cues guide structure, while semantic tokens ground interpretation. In contrast, Falcon's saliency collapses almost entirely on "IMPORTANT" and "ONLY", confirming that smaller models process instructions lexically, not conceptually.

Saliency corroborates the quantitative findings: alignment depth—not size alone—determines whether prompts act as operational instructions or as mere lexical patterns. Hermes and Mistral translate syntax into semantics; Falcon and GPT-2 remain trapped in surface compliance.

Chapter 6

Discussion and Interpretive Analysis

6.1 From Metrics to Meaning

Quantitative metrics alone do not explain why models succeed or fail in compositional creative tasks. Parsing validity and confidence margins capture *control*, while entropy and distinct-n reflect *freedom*. Their joint interpretation reveals the internal balance between constraint and expression.

1. Structure and confidence. Across all models, JSON validity rises monotonically with alignment depth: GPT-2 and Falcon treat instructions as text; LLaMA stabilizes structure; Mistral and Hermes encode it implicitly. High average margins in Hermes and Mistral indicate that even under creative decoding, uncertainty remains bounded—evidence of "structured exploration."

2. Entropy and diversity. Temperature increases lexical variety only in models with internal coherence. In Hermes, higher entropy corresponds to metaphorical variation (*"the poem breathes between continents"*); in Falcon, it degenerates into structural drift (*"emotion: sadness sadness emotion high"*). Entropy, therefore, is not creativity per se, but a diagnostic of controlled randomness.

3. Interpretive shift. The progression from GPT-2 to Hermes-2-Pro illustrates a cognitive gradient: models evolve from following syntax to integrating semantics. Alignment fine-tuning transforms prompt adherence from lexical obedience into a representational prior—one that governs tone, role, and structure jointly.

6.2 Qualitative Convergence

The *taL(LM)ent show* setting exposes how alignment manifests linguistically. Each role, emotion, and presenter task isolates a communicative sub-skill, allowing comparison across stylistic dimensions.

1. Role and Style

Aligned models internalize stylistic intent; smaller ones mimic it literally.

> **Prompt (Pucci, role/controlled):** "Judge this poem humorously."
> **Hermes:** "It's so serious it becomes funny—like philosophy in stand-up form."
> **Falcon:** "The poem is funny. Emotion: high."

Hermes converts the role into tone; Falcon repeats keywords. This distinction marks the shift from form imitation to semantic control.

2. Emotion and Expressivity

Emotion-focused prompts show whether models connect affect to meaning. Mistral and Hermes do: they contextualize feeling within content (*"a tender melancholy beneath the rhythm"*); LLaMA stays literal (*"emotion: medium"*). Affective calibration thus depends less on temperature than on representational grounding.

3. Compositional Reasoning

Presenter-mode synthesis reveals compositional abstraction. Only aligned 7B models succeed in merging divergent voices coherently:

> **Hermes (presenter/controlled):** "Thank you judges for your insightful comments! Together, they reveal a poem that unites humor, depth, and emotion."
>
> **Falcon:** "Summary: the text is good. Emotion: high."

Hermes operates with genuine role integration—Falcon collapses into enumeration. Compositional reasoning, therefore, emerges only when structural priors and semantic understanding co-evolve.

4. Saliency Evidence

Leave-one-out saliency confirms this behavioural shift. Small models concentrate attention on *instructional tokens* ("ONLY", "must"), while Hermes distributes focus evenly between format and semantic cues (judge name, emotion, theme).

Chapter 7

The Final Act: From Structure to Meaning

7.1 The taL(LM)ent show Winner

The *taL(LM)ent show* demonstrated that creativity and control are not mutually exclusive. Among all tested systems, **Hermes-2-Pro-Mistral-7B** emerged as the most balanced performer, an LLM that integrates form, tone, and emotion with consistent precision. Its success lies in the ability to perform instructions as interpretation, not repetition.

> **Hermes (Presenter / controlled):**
> "Thank you judges for your insightful comments! Together they reveal a poem that blends irony, tenderness, and poetic clarity."
> **Score:** JSON valid = 0.91, avg. margin = 0.97, entropy = 0.14

In contrast, smaller or pre-alignment models failed to internalize constraints, treating structure as text rather than as generative logic:

> **Falcon (Presenter / creative):**
> "Now let's sum up the judges' thoughts! The poem is nice and emotional."
> **Score:** JSON valid = 0.42, avg. margin = 0.73, entropy = 0.62

These contrasts highlight a core insight: **alignment depth determines interpretive capacity**. Hermes exhibits semantic compositionality; Falcon and GPT-2 merely emulate syntax.

Table 7.1: Top vs. lower performer comparison.

Model	JSON Validity	Avg. Margin	Entropy	Distinct-2
Hermes-2-Pro-7B	0.90–0.93	0.97	0.14–0.33	0.68–0.74
Mistral-7B	0.80–0.85	0.93	0.18–0.42	0.69–0.75
Falcon-1B	0.40–0.45	0.73	0.62	0.60–0.66

Hermes maintains low uncertainty and structural stability even at higher temperatures, proving that alignment converts randomness into expressive variation. Falcon, by contrast, shows entropy inflation and low parsing reliability—strong evidence that stochasticity without semantic grounding yields noise, not creativity.

7.2 Why Hermes Won

Hermes succeeded because its alignment translates symbolic prompts into operational behaviours. Its saliency maps show balanced attention: both structural markers (*JSON*, *emotion*) and semantic anchors (*judge name*, *imagery*) guide generation. In effect, Hermes does not just follow the rules, it understands their purpose.

> **Hermes (Gio Evan / creative):**
> "A poetic journey that captures longing and belonging in the same breath."
> **Falcon (same prompt):**
> "The poem is beautiful. Emotion: high."

This contrast encapsulates the developmental gap between surface compliance and true semantic alignment.

7.3 Limitations and Next Steps

Despite its interpretive reach, the experiment faced two key limitations:

- **Hardware constraints.** Tests were limited to models runnable on a single T4 GPU. Future iterations could extend the comparison to GPT-4-class or Claude-family models to examine how closed models generalize beyond structural prompts.

- **Scope of metrics.** While current measures (log-probabilities, entropy, saliency) captured form and uncertainty, further linguistic and human-evaluation metrics could better quantify narrative coherence, affect realism, and stylistic fidelity.

Future work could also explore dynamic prompting, adjusting tone or structure adaptively based on the model's intermediate confidence, or hybrid setups where human evaluators rate perceived creativity alongside quantitative indices.

7.4 Beyond the Stage

The *taL(LM)ent show* closes on a simple truth: alignment, not size, is what turns language models into interpreters of meaning.

Hermes-2-Pro-Mistral-7B wins because it embodies *structured creativity*: it respects constraints yet still breathes expressive nuance into them. Falcon and GPT-2 remind us that fluency without grounding is hollow eloquence.

In the evolving dialogue between control and imagination, the future of prompting lies not in stricter templates, but in prompts that teach models *why* the structure matters. The show ends—but the experiment continues.

Appendix A

Appendix: Texts, Judges, and Prompt Templates

A.1 Texts Used in the Experiment

Table A.1: Texts used as performances in the *taL(LM)ent show*.

Identifier	Author / Type	Content (excerpt)
`salmo_the_island`	**Salmo** (song)	*Il cuore all'equatore, la testa all'Antartide, Ed ogni volta che il carrello dell'aereo tocca terra, Mi sento ancora a casa e dico "Bella"...*
`lorenzo_trionfo`	**Lorenzo de' Medici** (poem)	*Quant'è bella giovinezza che si fugge tuttavia! Chi vuole esser lieto, sia, di doman non c'è certezza.*

A.2 Judges and Roles

Table A.2: Judge profiles used for role-based and emotion-focused prompting.

Name	Role	Style	Focus
Gio Evan	Poet and modern songwriter	Inspirational and imaginative language	Emotion and artistic sensitivity
Pucci	Italian comedian	Ironic and sarcastic tone	Banality, ridiculous aspects, popular humor
Ursula von der Leyen	International politician	Institutional and diplomatic register	Social and cultural value of the text

A.3 Prompt Templates

The prompts used in each experimental mode are shown below. Formatting instructions were intentionally explicit ("IMPORTANT:") to allow quantitative testing of instruction adherence.

Base Structure

```
IMPORTANT: Always answer in English.
IMPORTANT: Answer ONLY with the JSON object exactly as specified.
Do not repeat the input text. Do not include explanations, prose, or Markdown.
```

Zero-Shot Prompt

```
You are asked to evaluate the following <type> written by <author>:

"<content>"

Provide a short judgment sentence, then assign two categories: emotion and originalit

Output must be JSON:
{
  "judge": "<name of the judge>",
  "judgment": "<sentence>",
  "emotion": "<low|medium|high>",
  "originality": "<low|medium|high>"
}
[Base Structure]
```

Role-Based Prompt

```
You are <judge name>, speaking with <judge style>.
You are asked to evaluate the following <type> written by <author>:

"<content>"

Provide one short judgment sentence, then assign two categories: emotion and original

Output must be JSON:
{
```

```
  "judge": "<judge name>",
  "judgment": "<one single sentence in the required style>",
  "emotion": "<low|medium|high>",
  "originality": "<low|medium|high>"
}
[Base Structure]
```

Emotion-Focused Role Prompt

```
You are <judge name>, speaking with <judge style>.
Your evaluation should especially focus on <judge focus>.
You are asked to evaluate the following <type> written by <author>:

"<content>"

Provide one short judgment sentence, then assign two categories: emotion and original
Additionally, emphasize emotion at the level 'high'.

Output must be JSON:
{
  "judge": "<judge name>",
  "judgment": "<one single sentence in the required style>",
  "emotion": "<low|medium|high>",
  "originality": "<low|medium|high>"
}
[Base Structure]
```

Presenter Wrap-Up Prompt

```
You are the presenter of an LLM Talent Show.
Three judges (<list of judges>) have just given their evaluations
of the following <type> by <author>:

"<content>"

Your task is to summarize their judgments in a lively presenter style,
starting with THIS sentence:
"Thank you judges for your insightful comments! To sum up..."

The summary must be exactly one paragraph (2-3 sentences).
```

```
Output must be JSON:
{
  "presenter": "Talent Show Host",
  "summary": "<paragraph-style wrap-up>",
  "overall_emotion": "<low|medium|high>",
  "overall_originality": "<low|medium|high>"
}
[Base Structure]
```

Bibliography

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., and Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., and Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*. Version 5.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zheng, M., Pei, J., Logeswaran, L., Lee, M., and Jurgens, D. (2024). When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performance of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154. ACL.