Decoding Emoji Movie Puzzles with Local LLMs

Adnan Sakib, Kiril Kuzmin

1. Research Question and Hypotheses

Research Question:
How do simple changes in prompting (zero-shot vs. few-shot vs. JSON-constrained) and sampling parameters (temperature, top_p) affect a local LLM's accuracy and latency when decoding emoji movie puzzles?

Hypotheses:

- **H1:** JSON-constrained outputs reduce formatting errors and slightly improve measured accuracy.
- **H2:** A small number of in-context examples (few-shot) improves exact-match rates over zero-shot.
- **H3:** Moderate temperature (0.3–0.4) balances exploration and correctness better than 0.0 or very high temperatures.

2. Experimental Setup

All experiments were conducted locally using the **Mistral 7B** instruction-tuned model via **Ollama**. The dataset (`emoji_puzzles.json`) contained 15 emoji-movie pairs (e.g., 🧑🕷️🏙️ → *Spider-Man*). Each condition combined:

- **Prompting strategy:** zero-shot, JSON-constrained, few-shot
- **Temperature:** {0.0, 0.3, 0.7}
  with **top_p = 0.9** fixed.

The evaluation harness (`evaluate.py`) sent each emoji to the model, recorded the raw output, wall-clock latency, and correctness after text normalization and alias matching. Accuracy @ 1 was computed as the proportion of exact matches after normalization.

3. Results

3.1 Accuracy × Temperature (%)

Prompt Type	0.0	0.3	0.7
Few-Shot	26.7	26.7	26.7
JSON-Constrained	13.3	13.3	13.3
Zero-Shot	0.0	6.7	6.7

3.2 Latency (s)

Prompt Type	Temperature	Mean Latency	P95 Latency
Few-Shot	0.0	1.82	5.17
Few-Shot	0.3	1.85	5.30
Few-Shot	0.7	1.47	2.87
JSON-Constrained	0.0	1.15	1.42
JSON-Constrained	0.3	1.12	1.38
JSON-Constrained	0.7	1.16	1.43
Zero-Shot	0.0	1.47	5.62
Zero-Shot	0.3	1.00	2.32
Zero-Shot	0.7	0.83	1.55

4. Analysis

Prompting effects.
Few-shot prompting achieved the highest accuracy (≈ 26.7 %) across all temperatures, confirming **H2**. Providing in-context examples helped the model recognize emoji-to-movie relationships more reliably. Zero-shot performance was poorest, showing the model’s limited ability to infer meaning without guidance.

JSON-constrained prompting produced well-formed JSON but lower semantic accuracy (≈ 13 %). This partially contradicts **H1**: while formatting errors were eliminated, accuracy dropped because the model prioritized syntax compliance over reasoning. For smaller models like Mistral 7B, enforcing strict JSON structure may restrict expressive freedom, reducing its ability to interpret ambiguous emoji patterns.

Temperature effects.
Accuracy remained largely stable across temperatures, implying that **prompt structure** had a stronger influence than sampling randomness. However, latency decreased slightly at higher temperatures, supporting **H3's** notion that moderate temperatures (0.3–0.4) balance exploration and determinism.

Latency observations.
JSON-constrained outputs were the fastest (≈ 1.1 s mean), as the model generated shorter, structured replies. Few-shot prompts had the highest latency (≈ 1.8 s mean, up to 5 s p95) because longer contexts increased generation time. Zero-shot responses showed higher variability.

Qualitative trends.
Typical success cases included 🧑🕷️🏙️ → *Spider-Man* and 🦁👑🌅 → *The Lion King*, while frequent failures were semantically close but not exact, such as 👸❄️ → *Snow Queen* instead of *Frozen*. This shows that the model captures partial semantics but fails strict string-matching accuracy.

5. Discussion and Future Work

The results demonstrate that **prompt structure plays a more decisive role than sampling temperature** in determining local LLM performance on emoji-to-movie decoding. JSON-constrained prompting reduced formatting variability but also restricted reasoning flexibility, leading to lower semantic accuracy. In contrast, few-shot prompting achieved the most balanced results, showing that minimal contextual examples can guide the model toward more accurate interpretations. Zero-shot prompting performed the weakest, confirming that the model benefits from even limited guidance when handling abstract symbolic inputs.

Future work should investigate **larger or intermediate instruction-tuned models** to see whether increased capacity improves reasoning under structured formats. Expanding the emoji dataset and adopting **semantic-similarity metrics** rather than strict exact-match scoring would provide a fairer evaluation of partial correctness. Additional studies could explore **hybrid prompting strategies** that blend reasoning freedom with controlled structure and examine how higher temperatures or alternative decoding methods influence creativity and reliability in local LLMs.