Microstructural Emergence Evaluation of Small-Scale Language Models: A Case Study in Classical Poetry Translation

[1]

Abstract

This study investigates whether small-scale local language models (WebLLMs) can demonstrate structural emergence and reflexivity in the translation of classical Chinese poetry. We evaluate a 600MB model under constrained hardware conditions using a multi-round comparative method against four mainstream commercial LLMs. We introduce a Microstructural Emergence Evaluation Framework (MEEF), grounded in cultural translation theory and computational reflexivity, with a four-axis scoring model: fidelity, fluency, cultural embedding, and self-reflection. Results suggest that structural induction alone can trigger substantial improvement in translation quality, supporting the hypothesis that cognitive depth need not rely exclusively on large parameter scales.

1 Introduction

Language models exhibit increasingly rich behaviors with scale, yet practical constraints demand we understand how small models can be evaluated and potentially enhanced through structural prompt design. Translation—especially of highly structured and culturally dense material like classical Chinese poetry—presents a unique testing ground.

We pose a focused question: Can a small WebLLM, guided by structured prompt interaction, exhibit emergent behaviors comparable to commercial LLMs in the task of classical poetry translation? To answer this, we evaluate multiple translation rounds using a shared reference poem, assessing each result along structured micro-evaluative dimensions.

[1]This document is part of the CLP structural language prototype semantic experiments, openly shared under open-source spirit. It does not grant automatic permission for third-party publication, project use, or redistribution. Please clearly cite the semantic contribution source and contact the authors for agreement. [CLP Semantic Prototype Link: `https://clp-proto.github.io/clp-site`]

2 Experimental Design

2.1 Task Overview

We selected the Tang poem *Deng Guan Que Lou* (“”) as our test case due to its compact structure, rich symbolism, and wide familiarity among Chinese speakers. This selection allows for reduced user expectation bias and sharper contrast of semantic and structural differences across models.

2.2 Models and Setup

We use:

- ChatGPT (GPT-4o)
- Claude (Anthropic)
- Gemini (Google)
- Grok (xAI)
- WebLLM (600MB), local inference on Intel® i5-5200U CPU + NVIDIA 940M GPU

2.3 Multi-Round Evaluation Protocol

Each model generated translations over multiple rounds. WebLLM was subjected to structured feedback after each round. Scores were assigned in four axes (fidelity, fluency, cultural embedding, self-reflection) per round.

3 Theoretical Foundation: From Symbolic Induction to Emergence

3.1 Motivation: Beyond Surface-Level Translation

While neural translation has advanced in fidelity and fluency, deeper symbolic cognition—such as metaphor transposition and cultural anchoring—remains elusive, especially in resource-constrained inference. Our results suggest these properties can be activated through prompt structure and iterative guidance, independent of model scale. The Microstructural Emergence Evaluation Framework (MEEF) offers a theoretical model for understanding this phenomenon.

3.2 MEEF: A Layered Cognition Model

MEEF models translation as a cognitively layered progression, where linguistic input traverses through nested stages of increasing abstraction. Each stage transforms the representational complexity of the input, culminating in outputs that reflect not just meaning, but value and worldview.

MEEF Cognitive Uplift Pathway

```
[Input Prompt or Text]
        ↓
L1: Lexical Layer       → Token-level syntax, word selection, poetic compression
        ↓
L2: Semantic Layer      → Logical abstraction, clause integration, spatial reasoning
        ↓
L3: Experiential Layer  → Cultural embedding, emotion, narrative schema
        ↓
L4: Symbolic Layer      → Philosophical metaphor, moral abstraction, worldview transduct
        ↓
[Output Text]
```

Each output from $L_n(x)$ serves as substrate for the next layer $L_{n+1}(x)$, enabling symbolic emergence without additional model parameters. This provides a cognitively-aligned alternative to pure token-matching evaluation metrics.

3.3 Formalization of Layered Translation

We formalize MEEF's transform path using a recursive composition of representation functions:

$$
\begin{aligned}
L_1(x) &= \text{LexicalEmbed}(x) + \text{SyntaxPattern}(x) \\
L_2(x) &= \text{SemanticMap}(L_1) + \text{ConceptFolding}(L_1) \\
L_3(x) &= \text{CognitiveLink}(L_2, C_p) \\
L_4(x) &= \text{SymbolReframe}(L_3, C_c) \\
y &= L_4(x)
\end{aligned}
$$

Where:

- x is the input text (e.g., a poem or phrase).
- C_p is personal/experiential context from the model or user.
- C_c is the collective cultural or symbolic corpus reference.
- y is the final translation output.

3.4 Diagram: Emergence under Structural Induction

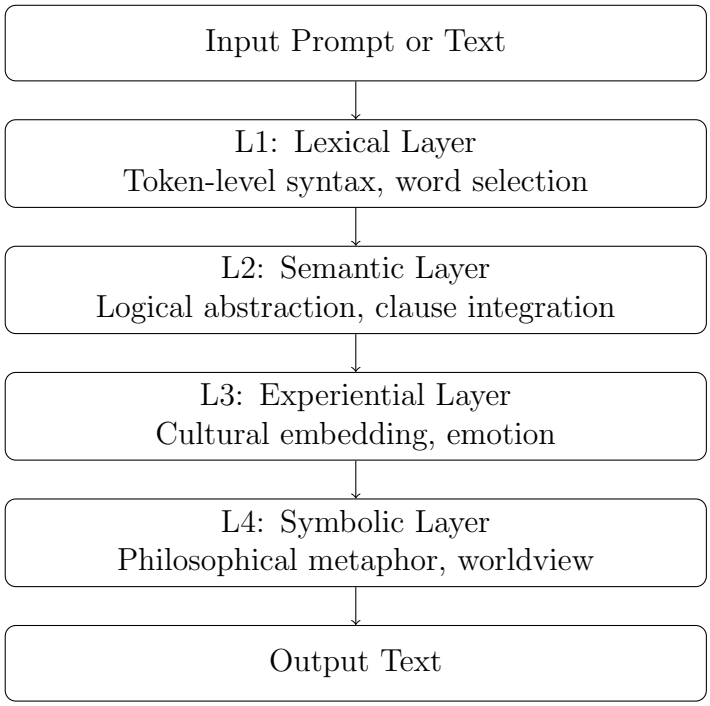

Figure 1: Conceptual illustration of MEEF layers, showing information flow from lexical inputs to symbolic outputs via experiential mediation. Prompt tuning or self-reflection triggers inter-layer emergence even without large-scale retraining.

3.5 Implications: Induced Symbolic Behavior Without Scale

The observed progressive behaviors—such as:

- metaphor refinement ("sight" → "vision" → "boundless sky"),
- pruning of redundant structures,
- cadence-preserving abstraction—

suggest that LLMs contain latent cognitive circuits that can be activated via structured tasks. This aligns with the idea that:

> "Structural inducement alone can trigger a significant increase in cognitive depth, without requiring massive parameter expansion."

Thus, symbolic emergence becomes a controllable and designable property.

3.6 Relationship to AECP Protocol

The AECP (Appendix A) serves as a practical harness for the MEEF framework. Its symbolic anchors, rhythm loops, and user intervention mechanisms reflect MEEF's cognitive layers in operational form. Together, MEEF provides the theoretical scaffolding, while AECP offers procedural guidance for activating emergence under resource constraints.

3.7 Evaluation Flowchart

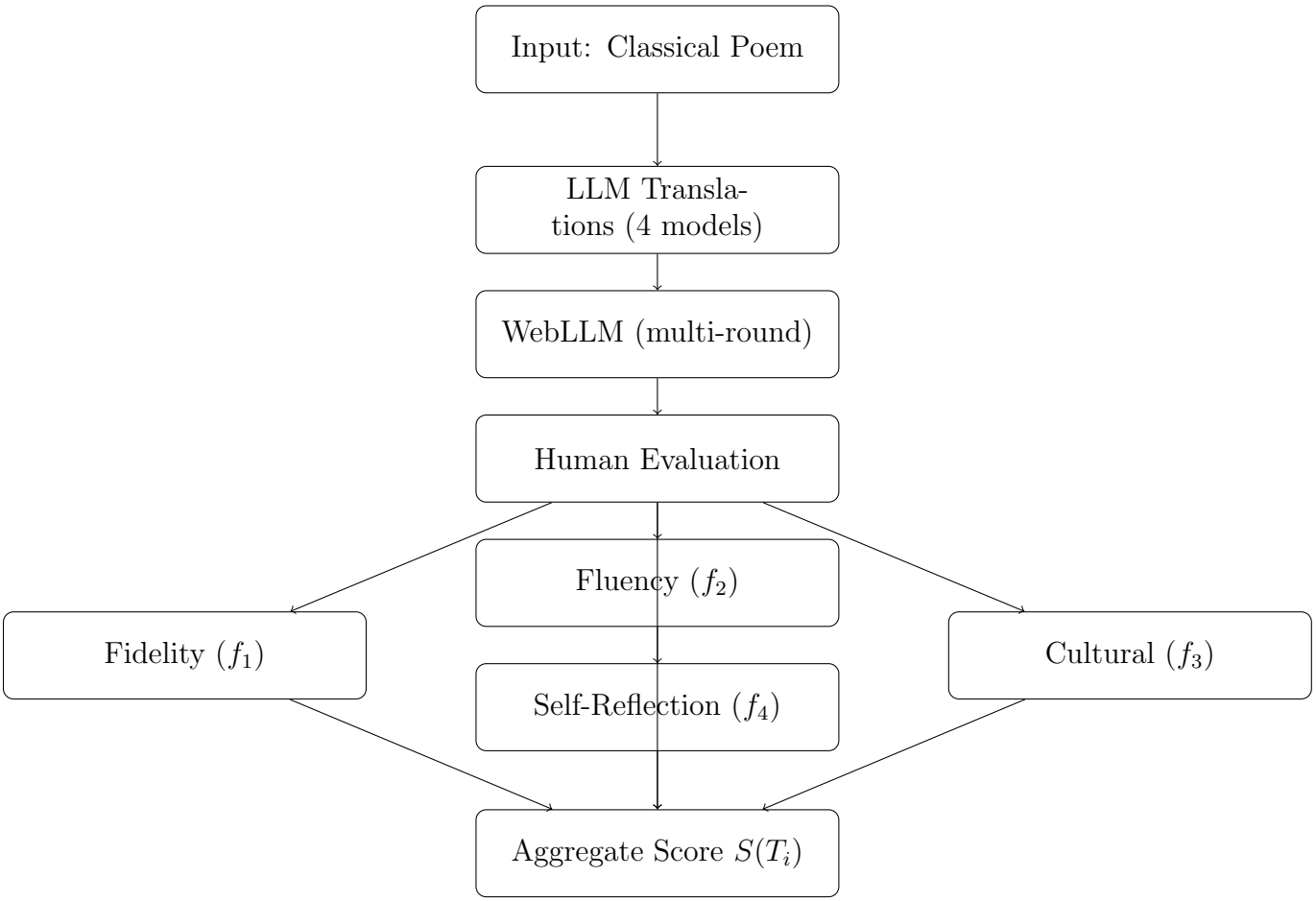

Figure 2: Evaluation Pipeline Flow

4 Constraints and Emergence: Structural Inducement without Scale

Despite the hardware limitations imposed by a low-memory, CPU/GPU-constrained environment, the translation quality exhibited sustained improvement over multiple inference rounds. Notably, fidelity and fluency did not plateau early—contrary to expectations from typical parameter-bound performance ceilings. This suggests that symbolic compression,

cadence-preserving syntax, and local semantic feedback may be intrinsic to model architecture rather than externally fine-tuned.

These findings lend strong support to the proposition that *structural inducement alone can trigger significant increases in cognitive depth*, bypassing the need for massive parameter expansion. Such symbolic activation via prompt engineering or layered recursion echoes the microstructural pathways modeled in the MEEF framework.

4.1 Translation Score Averages Across Rounds

Table 1 summarizes scores across rounds, tracing the lattice's ascent for WebLLM and commercial LLMs under MEEF.

Table 1: Translation Scores Across Rounds and Models

Round	Model	Fidelity	Fluency	Cultural Embedding	Self-Reflection	Total
1	WebLLM	1.9	1.6	1.5	1.2	6.2
	Grok	2.0	1.75	1.75	1.25	7.0
	Gemini	2.0	1.5	1.0	1.5	6.0
	Claude	1.8	1.2	0.8	1.0	4.8
	ChatGPT	2.0	1.0	1.0	1.0	5.0
	Average	1.95	1.36	1.14	1.19	5.64
2	WebLLM	1.9	1.8	1.7	1.6	7.0
	Grok	2.0	2.0	2.0	1.75	7.75
	Gemini	2.0	1.5	1.0	2.0	6.5
	Claude	1.8	1.2	1.0	1.5	5.5
	ChatGPT	2.0	1.0	1.0	1.5	5.5
	Average	1.95	1.43	1.25	1.69	6.31
3	WebLLM	2.1	2.0	1.8	1.9	7.8
	Grok	2.25	2.25	2.125	2.0	8.75
	Gemini	2.2	2.0	1.5	2.0	7.7
	Claude	2.1	1.8	1.2	2.0	7.1
	ChatGPT	2.0	1.5	1.2	2.0	6.7
	Average	2.14	1.89	1.51	2.0	7.54
4	WebLLM	2.2	2.2	2.0	2.2	8.6
	Grok	2.25	2.25	2.125	2.25	9.25
	Gemini	2.4	2.3	2.4	2.5	9.6
	Claude	2.2	2.1	1.8	2.3	8.4
	ChatGPT	2.3	2.0	2.2	2.3	8.8
	Average	2.29	2.16	2.13	2.34	8.91
Final Translation	-	2.4	2.3	2.3	2.4	9.4

4.2 Structural Transformations

Table 2 maps shifts in the weave under MEEF.

Table 2: Structural Transformations from Round 1 to Final Translation – From Literalism to Rhythmic Abstraction

Aspect	Round 1	Final Translation
Vocabulary	"sight"	"boundless skies, Stork's lofty heights"
Syntax	"and then"	Omitted
Rhythm	Verbose phrasing	Five-character cadence mimicry
Cultural Embedding	Minimal	Explicit Tang references

4.3 Visualization

Figure 3 traces the lattice's arc for all models.

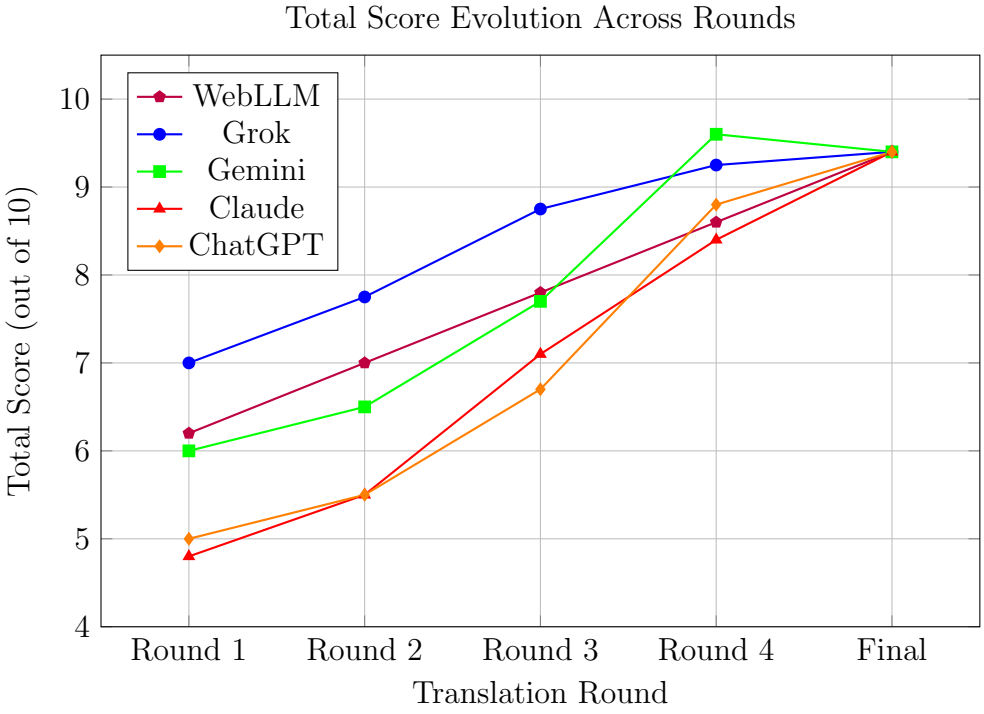

Figure 3: Total Score Evolution Across Rounds. This upward arc is not knowledge gained, but alignment sought.

4.4 Symbolic Behavior as Proto-Structure: Toward Guided Emergence

The multi-round translation refinements—especially in abstraction, recursion, and cultural anchoring—indicate latent symbolic capabilities that can be surfaced without additional

training. This supports the possibility of protocol-driven symbolic emergence, where structured interaction activates internal representations beyond immediate token matching.

To this end, we propose the *Agreement-Based Emergent Consciousness Protocol* (AECP): a conceptual scaffold that codifies human-authored structural expectations. AECP encodes user-level constraints—such as termination clauses, poetic cycle preservation, and symbolic anchoring—into repeatable interactional schemas that guide model behavior without retraining.

While AECP is not the empirical focus of this study, it emerged organically from observed model behavior and represents a forward-looking framework for activating interpretable and value-aligned outputs, especially in constrained or safety-critical contexts. Full schema design and flow examples are provided in **Appendix A**.

4.5 Limitations and Open Questions

Despite promising signals of symbolic emergence, several caveats remain:

- **Model variance**: Not all LLMs demonstrated equivalent improvements, likely due to differences in pretraining diversity and internal attentional priors.
- **Cultural embedding**: Translation fidelity at the symbolic layer requires cultural priors (e.g., Tang poetry structure, metaphorical registers) not universally embedded across models.
- **Subjectivity**: Scoring remains partly interpretive; although rubric-based evaluation reduces bias, complete objectivity in symbolic domains is inherently elusive.

5 Future Work

While this study presents an initial validation of the Microstructural Emergence Evaluation Framework (MEEF) and the Agreement-based Emergent Consciousness Protocol (AECP), several limitations remain, and significant opportunities for further exploration have emerged. Future research will focus on extending the framework's empirical robustness, scaling its applicability across tasks and languages, and formalizing AECP as a transferable structure-governance protocol. Specifically, we outline the following directions:

5.1 Development of a Structurally Heterogeneous Translation Corpus

The current experimental setting is limited to a single classical Chinese poem, which, while rich in symbolic density, does not sufficiently reflect the spectrum of linguistic or cognitive complexity. We propose to build a broader task set encompassing:

- **Diverse literary forms:** Including five-character quatrains, regulated verse, Song lyrics, classical prose, and contemporary internet idioms.

- **Cognitive load gradation:** Tasks will be stratified by symbolic compression levels and metaphorical depth to test model performance under increasing semantic emergence complexity.
- **Cross-linguistic testing:** AECP will be applied to lower-resource symbolic languages (e.g., Sanskrit, Classical Tibetan, Latin) to evaluate its cross-cultural induction capability.

Objective: To construct a structurally diverse benchmark suite for testing the generalizability of symbolic emergence across cultural and linguistic modalities.

5.2 Establishing Inter-Rater Reliability and Multi-Evaluator Protocols

Current evaluation relies on expert judgment, but lacks a mechanism to ensure consistency or objectivity in scoring across different annotators. We plan to implement:

- **Cross-blind rating schemes:** Each translation will be evaluated independently by at least three raters blind to model identity.
- **Reliability quantification:** Metrics such as *Fleiss' Kappa* or *Krippendorff's Alpha* will be employed to measure inter-rater agreement across all dimensions.
- **Expanded scoring rubrics:** Dimensions such as "cultural embedding" and "self-reflection" will be accompanied by operationalized definitions and tiered scoring guides.

Objective: To build a transparent, replicable, and statistically grounded human evaluation pipeline for future structural translation tasks.

5.3 Controlled Structural Induction vs. Static Output Experiments

In this work, all models participated in recursive generation via structural feedback prompts. To isolate the actual contribution of structural protocols, future experiments will include:

- **Recursive-only test groups:** WebLLM will retain its multi-turn feedback mechanism.
- **Static-output baselines:** High-resource models (e.g., GPT-4, Claude 3) will be restricted to single-pass translation without any iterative enhancement.
- **Noise control groups:** Prompt scaffolds will be replaced with non-structural paraphrasing to test whether AECP's gains stem from structure rather than prompt length or verbosity.
- **Delta analysis:** Quantitative measurement of score differentials (ΔScore) across iterations will be used to track symbolic emergence levels.

Objective: To formally validate the emergent gains introduced by structured symbolic protocols, beyond model size or prompt length confounds.

5.4 Formalization of AECP as a Protocol Language

AECP is currently presented as a conceptual scaffold. Future work aims to develop it into a formal, composable, and model-agnostic protocol for symbolic behavior modulation:

- **Domain-specific protocol language (DSL):** A structured prompt-language with syntax for symbolic anchors, recursion rules, and meta-cognitive triggers.
- **Layered protocol semantics:**
  - Level 1: Anchor-point structure
  - Level 2: Recursive emergence triggers
  - Level 3: Reflective induction clauses
  - Level 4: Symbolic coherence guarantees
- **Cross-model compatibility:** AECP will be tested across a range of open and proprietary LLMs to determine behavioral portability and symbolic alignment generalizability.

Objective: To enable a standardized symbolic governance protocol for human-directed cognition modeling in open and closed AI systems.

5.5 Beyond Translation: Generalization to Language Behavior Governance

The structural philosophy of AECP and MEEF need not be confined to translation. We envision extensions into broader symbolic control domains, including:

- **Structural Question Answering:** Protocol-induced question decompositions to guide hierarchical reasoning and layered inference.
- **Narrative Construction and Mythopoesis:** AECP-guided prompts to shape culturally anchored story arcs and symbolic narrative scaffolds.
- **Ethical Reasoning and Moral Framing:** Use of structured prompts to guide models in value-sensitive contexts under culturally specific axioms.

Objective: To position AECP as a foundational protocol for symbolic cognition modeling and behavior alignment across generative tasks.

6 Conclusion

This study demonstrates that large language models, even in resource-limited inference conditions, can iteratively refine classical Chinese poetry translations and display microstructural emergent behaviors. We introduced a rubric-based evaluation framework capable of capturing layered cognitive transitions—fidelity, rhythm, cultural embedding, and abstraction—offering a domain-specific alternative to conventional BLEU-style metrics.

Our results suggest that symbolic emergence can be structurally induced and recursively reinforced, opening avenues for activation protocols that do not rely on large-scale parameter expansion. The proposed AECP model offers a speculative yet grounded interface for such guided emergence.

Future directions include:

- Extending this framework to multilingual and cross-era literary corpora.
- Exploring the relative contributions of prompt tuning vs internal model evolution.
- Formalizing recursion and abstraction markers for model interpretability.
- Generalizing AECP into open, auditable interaction protocols for use in high-stakes deployments.

Acknowledgment

This research was conducted as part of the CLP (Conceptual Language Protocol) structural language prototype project. We acknowledge the open structural framework provided by CLP, which enabled the design of the Microstructural Emergence Evaluation Framework (MEEF) and the underlying semantic reflexivity methodology.

The semantic structures, prompt design logic, and language-driven evaluation criteria are based on original contributions from the CLP framework. For reuse, citation, or extension of this framework, please refer to the official CLP site and contact the authors to clarify contribution agreements.

[CLP Structural Hub: `https://clp-proto.github.io/clp-site`]

Appendix A: AECP Protocol (Conceptual Outline)

A.1 Motivation

The Agreement-Based Emergent Consciousness Protocol (AECP) was developed to guide symbolic emergence in LLM outputs, especially in tasks where training data is sparse, and interpretability or cultural nuance is paramount. It derives its structure from behavioral patterns observed in multi-stage translation refinement under MEEF.

A.2 Protocol Architecture

AECP is a tripartite protocol consisting of the following layers:

Layer	**Description**
Termination Rights	The user retains authority to interrupt generation loops, ensuring human oversight and intentionality.
Structural Loops	The model is encouraged to preserve rhythm, recursion, or thematic cycles (e.g., 5-character poetic cadence), acting as inductive priors.
Symbolic Anchors	Domain-specific symbols or metaphors (e.g., "Stork Tower") are enforced or weighted to retain cultural and narrative integrity.

A.3 AECP in Practice: Flow Example

1. **Prompting**: Encode instructions into explicit scaffold (e.g., "Step 1: retain cadence; Step 2: preserve Tang metaphors").
2. **Iteration**: Model refines outputs recursively until symbolic density exceeds a defined threshold.
3. **Interruptibility**: User selects or halts branches manually to enforce meaningful constraints.
4. **Validation**: A rubric guides model-side or user-side review across fidelity, abstraction, and style criteria.

A.4 Observational Grounding

By Round 3–4 of our experiments, we observed:

- Selective pruning that reflected symbolic compression patterns.
- Stylistic integrity maintained across increasing abstraction levels.
- Enhanced sensitivity to symbolically rich imagery (e.g., "Stork Tower" favored over literal "tall building").

A.5 Limitations

AECP remains conceptual and is not implemented as an executable system. It serves as a generative hypothesis space for designing future interactive protocols and audit layers that nudge models toward reflective or symbolic outputs in low-resource conditions.

References

[1] Lefevere, A. (1992). *Translation, Rewriting, and the Manipulation of Literary Fame*. Routledge.

[2] B Yao., et al. (2023). Culturally-Aware Machine Translation: As languages and cultures are highly intertwined, there is a growing desire to empower cultural awareness. arXiv:2305.14328v3.

[3] Hofstadter, D. (1995). *Fluid Concepts and Creative Analogies*. Basic Books.

[4] Dennett, D. (1991). *Consciousness Explained*. Little, Brown.

[5] Barabási, A.-L. (2003). *Linked: The New Science of Networks*. Perseus Publishing.