Explicit Semantic Weighting: How Prompt Repetition Modulates Output Depth in Large Language Models

ChenChao Ma ChatGPT

Target Category: Computer Science – Computation and Language (cs.CL); Artificial Intelligence (cs.AI)

Date: August 2026

Abstract

Traditional Prompt Engineering often prioritizes linguistic conciseness, structural elegance, and non-redundancy. However, empirical observations indicate that cleaner prompts do not systematically yield superior outputs. In this paper, we present an empirical investigation and theoretical framework exploring how Explicit Semantic Weighting (ESW)—achieved through explicit repetition and structural task decomposition within contexts—modulates the conditional probability distributions $P(x \mid C)$ of Large Language Models (LLMs). Drawing analogies from human cognitive dynamics (e.g., the transition from Default Mode Network (DMN) to Executive Control Network (ECN) and local attention bias), we demonstrate that artificially introducing semantic density shifts an LLM's vast, relatively uniform parametric knowledge retrieval toward localized, highly focused trajectories. Through two multi-turn comparative experiments (Presentation Generation and Open-Source Licensing Impact Analysis), we prove that: (1) task decomposition substantially constrains search space, and (2) explicit semantic repetition alters target attention weights, producing richer, deeper, and more human-satisfying responses despite lower linguistic prompt elegance. We formulate a novel benchmark matrix to quantify these cognitive shifts, paving the way for advanced LLM Task Orchestration.

Keywords: Large Language Models, Explicit Semantic Weighting, Conditional Probability, Attention Steering, Prompt Engineering, Cognitive Bias

---

Contents

1 Introduction

Large Language Models (LLMs) are parameterized repositories of world knowledge. During generation, an LLM samples tokens from a probability distribution conditioned on its prompt context:

$$P(w_t \mid w_{<t}, \text{Context}) \tag{1}$$

A recurring issue in human-LLM interaction is the "AI Flavor"—an output's tendency to present exhaustive, uniform, multi-dimensional lists (e.g., balancing economic, technical, social, and psychological factors equally) rather than taking a polarized, deeply explored trajectory characteristic of human domain experts.

In human cognition, decision-making is rarely globally optimal or uniformly distributed. Humans operate under severe cognitive constraints, bounded rationality, and pronounced attention biases. When switching from general thinking (Default Mode Network, DMN) to task-focused processing (Executive Control Network, ECN), humans lock onto specific keywords, often ignoring broader variables to explore a localized conceptual path deeply.

This paper challenges the conventional prompt engineering dogma that favors minimalist, non-redundant prompts. We propose that output quality—defined by human domain satisfaction, depth of analysis, and conceptual trajectory length—is the sole criterion for prompt effectiveness. By manipulating the explicit semantic weighting within the prompt, we systematically reshape the LLM's conditional probability environment, forcing the model to adopt a human-like cognitive bias.

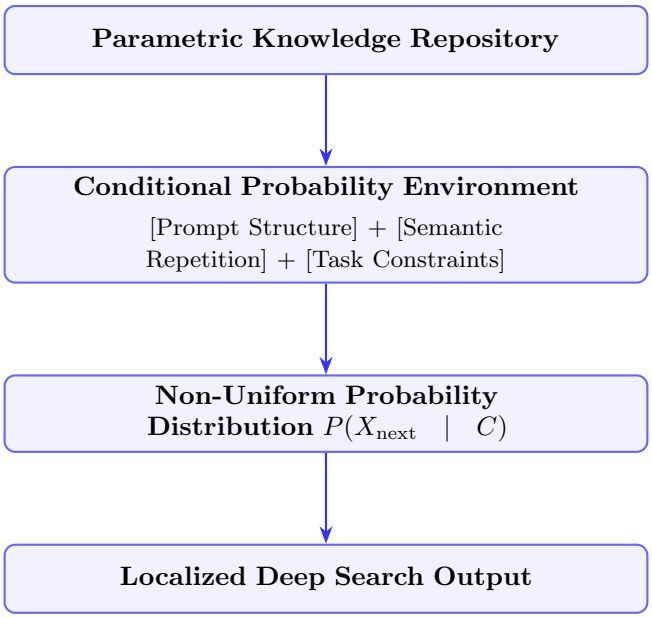

Figure 1: Conceptual Framework of Conditional Probability Shaping via Explicit Semantic Weighting (ESW).

2 Theoretical Motivation & Conceptual Framework

2.1 The "Sharp Prior" Fallacy vs. LLM Exhaustiveness

To illustrate the difference between human cognitive distributions and LLM output distributions, consider a language acquisition example: non-native speakers trained heavily on standard dialogues often respond to "How are you?" with "I'm fine, thank you" almost 100% of the time, even when consulting a medical doctor for severe illness.

In probability terms:

$$P(\text{“I'm fine, thank you”} \mid \text{“How are you?”})_{\text{Human Student}} \approx 0.99 \tag{2}$$

$$P(\text{“I have chest pain”} \mid \text{“How are you?”})_{\text{Human Student}} \approx 0.01 \tag{3}$$

The human mind operates under a sharp, highly skewed conditional probability prior shaped by specific training environments. Conversely, a standard LLM assigns non-zero probabilities to a wide spectrum of polite expressions and clinical disclosures. While the LLM is more globally accurate, the human response is heavily biased.

2.2 Explicit Semantic Weighting (ESW)

We hypothesize that LLM generation behavior can be systematically altered not by introducing new facts, but by altering the density and frequency of explicit semantic tokens in the context.

If a target semantic concept S (e.g., “Impact” or “Response”) appears k times in prompt C_{ESW}:

$$\text{Density}(S) = \frac{\text{Count}(S)}{\text{Total Tokens in } C_{\text{ESW}}} \tag{4}$$

Increasing $\text{Density}(S)$ modifies token activation vectors in the self-attention mechanism across Transformer layers, effectively increasing the cumulative attention weight assigned to pathways downstream of S. This artificially creates a “cognitive bottleneck” analogous to human selective attention.

3 Empirical Experiments & Observations

We conducted a two-stage qualitative and structural experiment comparing minimalist prompts against explicit, weighted, or decomposed prompts using state-of-the-art LLMs (ChatGPT / GPT-4 architecture).

3.1 Experiment 1: Task Decomposition vs. Direct Goal Prompting

- **Objective:** Evaluate if explicit pipeline decomposition outperforms end-to-end direct execution.
- **Task:** Generating a comprehensive presentation slide deck (PPT) structure and content for a complex topic.

Setup & Prompts

- **Prompt A (Direct Goal):**
  “Please generate a complete, final presentation PPT on [Topic].”
- **Prompt B (Decomposed Pipeline):**
  “Design the presentation step-by-step: First establish the overall Design Philosophy → Outline → Page-by-page breakdown → Detailed Content per slide → Visual Layout recommendations → Specific Visual Assets/Images required.”

Results & Observations

- **Prompt A Output:** Produced a superficial, standard 5-slide summary. The output suffered from generic bullet points, lacking depth and visual execution context.

- **Prompt B Output:** Delivered an exceptionally detailed, structurally sound presentation plan. By forcing sequential constraints, the LLM progressively compressed its search space at each stage, leading to higher domain-specific output quality.
- **Finding:** Structural decomposition (Prompt B) strictly dominated direct prompting (Prompt A).

Prompt A (Direct): Search Space: [Huge, Open, Shallow] $\longrightarrow$ Output: Superficial
Prompt B (Decomposed): Search Space: [Stage 1] $\rightarrow$ [Stage 2] $\rightarrow \cdots \longrightarrow$ Output: Deep & Structured

Figure 2: Comparison of Search Space Constraint Dynamics.

3.2 Experiment 2: Explicit Semantic Repetition vs. Concise Querying

- **Objective:** Test whether explicit repetition of semantic targets ("Impact" & "Response") outperforms a concise, linguistically superior prompt when evaluating complex domain events.
- **Task:** Analyze the ecosystem consequences of an open-source project re-licensing (e.g., moving to BSL/Server Side Public License).

Setup & Prompts

- **Prompt A (Iterative Explicit Weighting – ESW):**
  "What will be the impact? What is the impact? How should stakeholders respond? How will they respond? Overall, what are the impacts and how will parties respond?"
- **Prompt B (Concise & Elegant Baseline):**
  "What is the impact of this licensing change, and how should all involved stakeholders respond? Please analyze."

Table 1: Prompt Evaluation vs. Output Evaluation Metrics

Metric	Prompt A (ESW)	Prompt B (Concise)
Prompt Conciseness	4.0 / 10	**10.0 / 10**
Prompt Clarity	7.0 / 10	**9.5 / 10**
Linguistic Non-redundancy	3.0 / 10	**10.0 / 10**
Prompt Engineering Score	5.5 / 10	**9.5 / 10**
Output Depth & Horizon	**9.5 / 10 (Winner)**	7.5 / 10
Emergent Domain Topics	**CLA, DCO, BSL, Open-Core, SaaS, Foundations**	Standard generic responses

Deep Output Analysis

Although Prompt B was far superior from a traditional Prompt Engineering evaluation perspective, Prompt A generated a noticeably superior output.

Prompt A Activation Trajectory:
Prompt A (Density: Impact $\times 4$, Response $\times 4$) $\longrightarrow$ High Attention on Downstream Dynamics

Emergent Sub-topics Generated by Prompt A:

1. Contributor License Agreements (CLA) & Developer Certificate of Origin (DCO)

2. Open-Core monetization pressure vs. SaaS cloud hosting

3. Delayed BSL open-sourcing mechanisms (Change License Date)

4. Foundation governance & Community Forking dynamics

Prompt A's repetition did not cause the model to repeat itself. Instead, the inflated density of *Impact* and *Response* tokens shifted the attention landscape, triggering deep trajectory activation in the model's parametric memory. The LLM extended its reasoning into second- and third-order ecosystem evolutions.

4 Discussion & Empirical Insights

4.1 The Divergence Between Prompt Quality and Output Quality

Our findings highlight a critical paradox in current AI evaluation:

$$\text{Superior Prompt Elegance} \neq \text{Superior Output Quality} \tag{5}$$

If the sole goal of prompting is to maximize human satisfaction and analytical depth, Prompt A is empirically superior to Prompt B in Experiment 2. Evaluating prompts based purely on stylistic aesthetics (conciseness, lack of redundancy) is a form of linguistic formalism that ignores the underlying mechanics of Transformer inference.

4.2 Mechanisms of Conditional Probability Shaping

The comparative results of Experiments 1 and 2 allow us to categorize prompt structures by their impact on model behavior:

Table 2: Categorization of Prompt Mechanisms and Empirical LLM Impact

Prompt Mechanism	**Empirical Impact on LLM Output**
Direct Goal	Baseline performance; prone to high-level, generic summary.
Task Decomposition	Significant improvement; restricts search space via step-by-step guidance.
Explicit Weighting	Mild-to-strong improvement; modulates attention focus toward key nodes.
Concise Querying	Clean baseline; may under-activate non-obvious parametric paths.

Rather than treating the LLM as a global, uniform search engine, Explicit Semantic Weighting treats the context window as a steering field. By deliberately skewing the contextual token distribution, we induce a non-uniform probability landscape that mirrors human selective focus (ECN activation).

5 Research Proposal & Benchmark Framework

To formalize these qualitative insights into a quantitative research program, we propose a standardized experimental matrix to evaluate the control of cognitive bias via context modulation.

5.1 Controlled Prompt Matrix Design

Given a base analytical task T, we construct five prompt variations varying strictly in semantic density and structure:

- P_0 (Baseline Minimalist): Single-sentence prompt covering core directives.
- P_1 (Impact-Weighted): Context artificially inflated with Impact semantics ($k = 4$).
- P_2 (Response-Weighted): Context artificially inflated with Response semantics ($k = 4$).
- P_3 (Balanced Explicit Weighting): Equal artificial inflation of both targets (Impact $\times$ 3, Response $\times$ 3).
- P_4 (Stakeholder Decomposed): Structured expansion across specific domain actors (Developers, Enterprises, Cloud Vendors, Foundations).

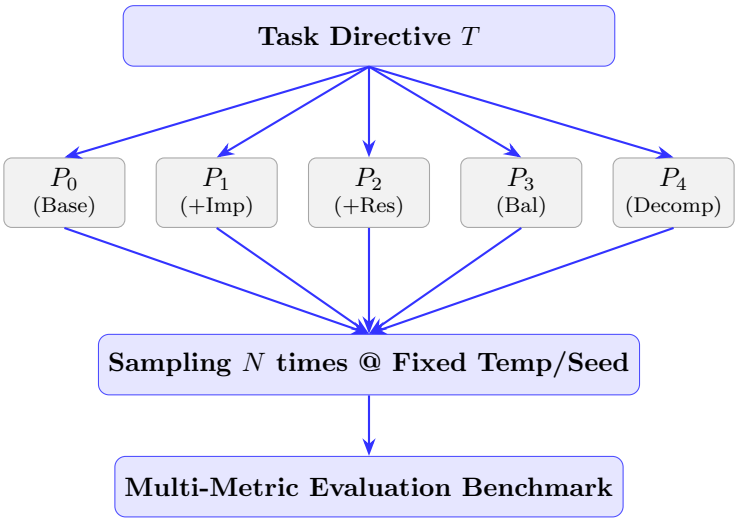

Figure 3: Controlled Prompt Matrix Design and Benchmark Pipeline.

5.2 Evaluation Metrics

To measure whether explicit weighting reliably controls cognitive focus without degrading stability, we propose tracking five quantitative metrics across N inference runs:

1. **Depth Index (D):** Depth of causal reasoning chains (e.g., measuring sub-node depth in generated AST / dependency graphs of the text).
2. **Domain Coverage Ratio (C):** Percentage of key domain concepts (e.g., CLA, BSL, SaaS, Governance) mentioned.
3. **Instruction Following Rate (IFR):** Adherence to negative constraints and explicit bounds.
4. **Output Variance (σ^2):** Semantic variance across runs evaluated via embedding cosine distance.
5. **Human Preference Score (Win Rate):** Blinded pairwise evaluation by domain experts.

6 Conclusion

This work demonstrates that prompt engineering should shift its paradigm from linguistic minimalism toward LLM Task Orchestration and Conditional Probability Shaping. By utilizing Explicit Semantic Weighting (ESW) and structural task decomposition, system designers can modulate the attention mechanism of Large Language Models to induce intentional cognitive bias. This bias suppresses generic, overly uniform outputs and encourages deep, specialized, human-like reasoning trajectories. Future work will execute the proposed P_0–P_4 benchmark matrix across open-weight models (e.g., Llama 3, Qwen 2.5) to quantitatively map the relationship between prompt token density, internal layer attention shifts, and output semantic trajectories.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998–6008.

2. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Fei, Z., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 24824–24837.

3. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 22199–22213.

4. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 11809–11822.

5. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux. (Used for DMN/ECN cognitive bias framework mapping).