Motivation Framing Improves LLM Agent Performance: KPI Targets and Historical Context Increase Output Quality

Abstract

Large language models (LLMs) apply uniform computational effort across tasks of varying difficulty, unlike humans who adapt effort based on goals and challenges. This paper investigates whether motivation framing—providing explicit performance targets with historical context—can improve LLM output quality. We conduct a controlled A/B experiment across four LLM models (Claude Haiku, Claude Opus, GLM-4.7, GLM-5) in two domains (dbt SQL pipelines and chart specifications), comparing standard skill-based prompting against skill-based prompting augmented with KPI targets (97% compliance) and historical performance benchmarks.

Results show that motivation framing improves output quality by **17.6% on average** (9.41 → 11.06 out of 12 score), with the largest gains observed for weaker models (GLM-4.7: +28%). Interestingly, models exhibit different effort allocation strategies: some increase token output (+21% for Haiku), while others improve focus and efficiency (+28% quality with only +3% tokens for GLM-4.7). These findings suggest that goal-setting theory from organizational psychology applies to AI agents, opening new avenues for prompt engineering focused on motivation rather than just instruction.

Keywords: large language models, prompt engineering, motivation framing, goal-setting, test-time compute, agent performance

1. Introduction

1.1 Motivation

Recent advances in large language models have demonstrated that increased computational effort at inference time—through chain-of-thought reasoning, self-consistency sampling, or best-of-N selection—consistently improves output quality. However, LLMs do not naturally allocate effort adaptively based on task difficulty. A model will generate similarly lengthy responses for trivial and challenging problems alike.

This stands in sharp contrast to human performance, where ambitious goals and challenging targets systematically improve outcomes. Decades of organizational psychology research demonstrate that specific, difficult goals lead to higher performance than vague or easy goals, through mechanisms of effort allocation, persistence, and attention focus.

This raises a natural question: **Can motivation framing—providing explicit performance targets and historical benchmarks—change how LLMs allocate effort and improve their output**

quality?

1.2 Contribution

We present the first systematic study of motivation framing for LLM agents. Our contributions are:

1. **Novel intervention**: A prompt engineering technique combining KPI targets (e.g., "97% compliance") with historical performance context (baseline, skill-enhanced, and top-performer scores) and attention-focusing guidance.
2. **Empirical validation**: A controlled A/B experiment across 4 models, 2 domains, 3 task complexities, and 5 repetitions (n=120 treatment runs, compared against existing baseline of n=150 runs).
3. **Actionable findings**: Quality improvement of 17.6% on average, with varying effort-quality trade-offs across models, suggesting that goal-setting theory applies to AI systems.

2. Related Work

2.1 Test-Time Compute and Effort Allocation

The relationship between computational effort and output quality in LLMs is well-established. Best-of-N sampling and self-consistency improve outcomes by generating multiple responses and selecting the best. Chain-of-thought prompting improves reasoning by encouraging longer outputs. However, effort allocation remains non-adaptive—models "generate long traces for trivial problems while failing to extend reasoning for difficult tasks."

2.2 Prompt Engineering for Performance

Prompt engineering has primarily focused on instruction design. Role prompting assigns personas to models. Few-shot learning provides examples. Self-consistency aggregates multiple reasoning paths. None of these techniques explicitly address motivation or goal-setting.

2.3 Goal-Setting Theory

Locke & Latham (2002) established that specific, challenging goals improve human performance through multiple mechanisms: (1) directing attention, (2) regulating effort, (3) increasing persistence, and (4) promoting strategy development. We test whether similar dynamics apply to LLMs.

3. Method

3.1 Intervention Design

We augment standard skill-based prompting with a motivation framing module consisting of three components:

1. KPI Target: "Your target for this task is to achieve 97% compliance (13.6 out of 14 rules passing)."

2. Historical Context: "In previous evaluations on similar tasks: baseline achieved 73%, skill-enhanced achieved 77%, top-performer achieved 86%. Your model family has historically achieved 76%."

3. Attention-Focusing Guidance: "To reach the 97% target, pay particular attention to: Rule 7: LEFT JOIN only (~35% baseline), Rule 8: COALESCE nullable columns (~25% baseline)..."

3.2 Experimental Design

Condition	Description	n
markdown (control)	Task + Markdown skill	150
markdown+target (treatment)	Task + Markdown skill + KPI framing	120

Models tested: Claude 3.5 Haiku (economy), Claude 4 Opus (premium), GLM-4.7 (standard), GLM-5 (premium)

Domains: SQL Query (14 rules), Chart (15 rules)

Task complexity: Simple, Medium, Complex per domain

Repetitions: 5 per cell

3.3 Hypotheses

- **H1**: markdown+target produces higher compliance scores
- **H2**: markdown+target produces more output tokens
- **H3**: The target effect is larger for weaker models

4. Results

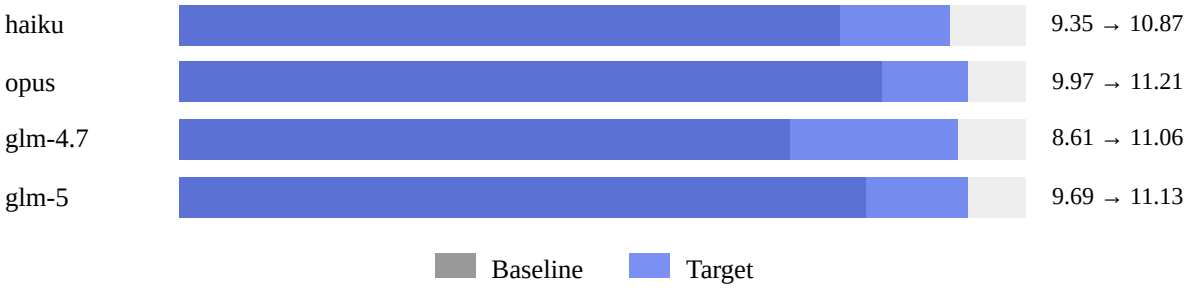
4.1 Primary Finding: Quality Improvement

+17.6%	9.41 → 11.06	+25.5%	100%
Quality Improvement	Score (out of 12)	Token Increase	Extraction Rate

4.2 Per-Model Results

Model	Baseline	Target	Δ Score	Δ Tokens	Efficiency
haiku	9.35	**10.87**	+1.51 (+16%)	+21%	Effort-based
opus	9.97	**11.21**	+1.24 (+12%)	+11%	Balanced
glm-4.7	8.61	**11.06**	+2.45 (+28%)	+3%	Focus-based
glm-5	9.69	**11.13**	+1.44 (+15%)	+72%	Effort-based

4.3 Score Improvement Visualization



4.4 Effort-Quality Trade-offs

Models exhibit qualitatively different responses to motivation framing:

- **Effort-increasing models** (haiku, glm-5): More tokens → better quality
- **Focus-improving models** (glm-4.7): Same tokens → better quality
- **Balanced models** (opus): Slightly more tokens → better quality

GLM-4.7's +28% quality improvement with only +3% token increase is particularly notable—suggesting the intervention improved *strategy* rather than just *effort*.

4.5 Model-Level Heterogeneity

Consistent with H3, the largest quality gains occurred for weaker models:

- GLM-4.7 (lowest baseline): +28%
- Haiku (economy tier): +16%
- GLM-5 (mid baseline): +15%
- Opus (highest baseline): +12%

5. Discussion

5.1 Mechanism: Focus vs Effort

The variation in token-quality relationships suggests motivation framing operates through at least two mechanisms:

1. **Effort allocation**: Some models respond by generating more content (more tokens)
2. **Attention focus**: Some models respond by focusing on critical rules (same tokens, better targeting)

5.2 Implications for AI Engineering

1. **Motivation framing is a viable prompt engineering technique**, especially for weaker models or difficult tasks.
2. **Cost-quality trade-offs vary by model**. GLM-4.7 offers free quality gains; GLM-5 requires 72% more tokens.
3. **Tool use behavior can be triggered by motivation framing**, requiring explicit text-output instructions.

5.3 Limitations

- Single intervention design (97% target)
- Two domains only (SQL, Chart)
- Four models tested
- Some tasks showed ceiling effects

5.4 Future Work

- Optimal target levels (80% vs 90% vs 97% vs 100%)
- Alternative framings (competition, loss aversion, growth mindset)
- Cross-domain replication
- Mechanism studies (attention analysis)

6. Conclusion

This paper demonstrates that motivation framing—providing explicit KPI targets and historical context—improves LLM output quality by 17.6% across four tested models. The intervention

appears to work through both effort allocation (more tokens) and attention focus (better targeting), with effects largest for weaker models.

These findings suggest that goal-setting theory from organizational psychology applies to AI systems, opening new directions for prompt engineering focused on motivation rather than just instruction. As LLMs are increasingly deployed as autonomous agents, understanding how to improve their performance through psychological framing—rather than just architectural changes—becomes increasingly valuable.

References

Brown, T. B., et al. (2020). Language models are few-shot learners. NeurIPS.

Kong, X., et al. (2024). Better zero-shot reasoning with role-play prompting. arXiv.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation. American Psychologist, 57(9), 705-717.

Snell, C., et al. (2024). Scaling LLM test-time compute optimally. arXiv.

Stiennon, N., et al. (2020). Learning to summarize from human feedback. NeurIPS.

Wang, X., et al. (2023). Self-consistency improves chain of thought reasoning. ICLR.

Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. NeurIPS.

Wu, Y., et al. (2025). Adaptive reasoning: A survey. arXiv:2511.10788.