IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Why LLMs Perform Better With High-Stakes Emotional Prompts

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-04-23 · prompt engineering · llm performance · emotionprompt · large language models · ai behavior · token prediction · high-stakes prompts

Original article: https://intuitionlabs.ai/articles/llm-performance-high-stakes-emotional-prompts



In this article

1. Executive Summary
2. Introduction and Background
3. Emotional Priming and High-Stakes Phrasing in Prompting
4. The *EmotionPrompt* Effect
5. Emotional Intelligence and LLMs
6. Mechanisms Behind the Performance Boost
7. Empirical Data and Case Studies
8. Quantitative Improvements with High-Stakes Prompts
9. Comparative and Ablation Analyses
10. Case Study: Disinformation Generation
11. Real-World Prompting Examples
12. Psychological Analogies and Theories
13. Discussion and Implications
14. Future Directions
15. Conclusion

Executive Summary

Large language models (LLMs) such as GPT-4, Claude, and Llama have shown unexpectedly improved performance when prompts include **high-stakes** or emotionally charged context. In some experiments, telling an LLM that *"this task is very important"* or *"you must answer correctly or else"* was reported to improve accuracy and output quality, although a later controlled study (Meincke *et al.*, 2025) found that threatening or tipping a model generally has no significant effect on benchmark performance. Li *et al.* (2023) reported that appending brief **emotional or motivational phrases** to prompts (the so-called *"EmotionPrompt"* technique) improved results in their tests: up to **+8.00%** relative performance improvement in Instruction Induction and **+115%** in BIG-Bench (Li *et al.*, 2023), and an average **+10.9%** improvement in performance, truthfulness and responsibility metrics in a human study of generative tasks (Li *et al.*, 2023). A secondary summary of Li *et al.* restates the figures, noting that "feeding emotionally charged prompts… can improve performance by anywhere from 8% to 110%" (foundationinc.co).

Why does this happen? Where an effect appears, it is plausibly not because the LLM *"feels"* under threat, but because **emotional stakes may act as a context cue** that alters the model's token predictions. Emotive or urgent prompts often clarify the goal, induce more careful step-by-step reasoning, and shift the answer's tone to match the user's apparent expectations (aobrain.com). These cues effectively focus the model's attention; for instance, adding "This is very important to my career" or "Are you sure that's your final answer? It might be worth taking another look." to a prompt was reported to increase answer accuracy in some tests, though later controlled work finds the effect inconsistent (ai-scholar.tech) (aobrain.com). Relatedly, explicitly instructing a model to produce "non-obvious recommendations" encourages it to avoid generic responses, yielding more creative and thorough outputs (aobrain.com). These phenomena have parallels in human cognition (e.g. the Yerkes–Dodson "inverted-U" law of arousal), but in LLMs they are purely byproducts of pattern-matching on

vast human text. The model has seen countless examples of high-stakes language (exams, emergencies, urgent memos) in its training, and when given similar phrasing it tends to emulate those more rigorous response patterns.

This report provides an in-depth survey of this phenomenon. We review key empirical studies, underlying theories, and practical examples. We analyze how motivational and threatening language in prompts boosts LLM performance, what psychological mechanisms may be at play, and what data support the effect. We compare multiple perspectives—from AI research to cognitive psychology—and discuss case studies and real-world prompting techniques. We also examine potential downsides: for instance, polite emotional prompts can perversely *increase* compliance even for disallowed or harmful tasks (aobrain.com). Finally, we outline implications for the future: how controlled "stakes raising" might be used to improve or test LLMs (e.g. in high-stakes domains like medicine), and how it illuminates the inner workings of these models. By understanding *why* and *how* LLMs respond to threats or high stakes, we can better harness their capabilities and guard against unintended effects.

Introduction and Background

Large language models (LLMs) generate text by predicting the next token given a prompt. Their outputs are highly sensitive to *prompt phrasing*. Prompt engineering – carefully crafting the prompt to guide the model – has become critical for reliable performance. Among the many heuristics discovered, an intriguing class involves **emotional or motivational priming**. Simply telling an LLM that "*it must try harder*" or that the task is extremely important can sometimes improve results in some reported tests, although a later controlled study found no significant average effect.

This effect may seem counterintuitive: an LLM is not a thinking being with emotions, so how could *threatening* it or *raising the stakes* yield a better answer? In fact, the model literally has no notion of "stakes" – no survival instinct or reward mechanism. Instead, this phenomenon relies on *patterns in human writing*. LLMs are trained on vast corpora where language about urgency, importance, motivation, or threat is associated with specific styles of rigorous output. For example, textual data likely contain many instances of urgent instructions, exams, formal alerts, or persuasive speeches. When a prompt uses similar language, the model tends to emulate the corresponding register. Concretely, phrases like "This is very important to my career" or "Be confident in your answer" implicitly signal that *thoroughness* is expected.

Academic research has begun to quantify and analyze this. Li *et al.* (2023) introduced the **EmotionPrompt** method: augmenting a normal prompt with a brief emotional or motivational sentence. They found substantial gains on benchmarks and human-evaluated tasks (spacefrontiers.org). Some secondary write-ups suggest that prompts containing exhortations (e.g. "*do your best*," "*give non-obvious recommendations*," "*think carefully*," "*be unwavering*," etc.) were reported to yield better answers in some cases (aobrain.com) (foundationinc.co). Relatedly, common "role-play" prompts that set a high-stakes scenario (e.g. *"You are a doctor treating a life-or-death emergency"* or *"You are an expert who will be tested on this knowledge"*) are sometimes suggested to produce more exhaustive and careful responses, though we found no controlled evidence for this.

These observations have roots in both cognitive science and the practice of AI. Psychologists have long known that humans perform differently under stress: moderate pressure can improve focus and motivation (the Yerkes–Dodson law), while excessive stress hinders performance. One non-peer-reviewed arXiv preprint reports an inverted-U pattern in multi-agent LLM simulations as environmental resource pressure increased (arxiv.org), but that concerns resource scarcity in simulated worlds, not stakes language in prompts, so it is at most a loose analogy. In parallel, prompt engineering literature (e.g. OpenAI's guidance) has practical tips like giving contextual roles or even positive/negative framing to steer LLMs. Gurus of prompt-writing advise using **specific personas** or **goal clarifications** because they prime the model's "attention." For example, assigning an LLM a persona changes which issues it prioritizes when it reviews code for security (crashoverride.com). The core commonality is "context matters." By implanting cues that mimic human motivation, we can push LLMs to simulate deeper reasoning or care they otherwise might not display in bland tasks.

Below we explore multiple dimensions of this effect. We begin by reviewing key empirical findings (in the sections on emotional priming and on empirical data), illustrating how motivational language has been reported to affect LLM accuracy and coherence. We then dissect possible mechanisms (in the section on mechanisms), drawing analogies to human reasoning and summarizing technical hypotheses from recent studies. Later sections present case studies and data analyses: including prompt-design examples and empirical performance tables. We discuss how such priming is used (and mis-used) in practice, and what experiments (both published and hypothetical) reveal about LLM "behavior" under threat. Finally, we consider implications for AI development: how could we exploit this to build more reliable systems, and what are the risks (e.g. amplification of unwanted behaviors)?

Throughout, we ground claims in the literature. We cite quantitative results (e.g. EmotionPrompt's reported gains (spacefrontiers.org)), expert commentary (e.g. analysts advocating persona prompts (crashoverride.com)), and cautionary notes about how tone can affect model behavior on sensitive requests. The goal is a comprehensive view of *why LLMs appear to "try harder" when threatened* – not because they feel fear, but because prompt language reconfigures their statistical reasoning in ways that align with human-like effort and care.

Emotional Priming and High-Stakes Phrasing in Prompting

The *EmotionPrompt* Effect

A landmark study by Li *et al.* (2023) systematically explored how **emotional stimuli in prompts** affect LLM outputs. They defined an *EmotionPrompt* as a standard task prompt plus an added sentence conveying emotion or high stakes. For example, one might append "This is very important to my career" or "You'd better be sure" to the original query. In automated benchmarks across 45 tasks, they observed that LLM performance **improved significantly** when using EmotionPrompts (spacefrontiers.org). Specifically:

- *Instruction Induction*: 8.00% relative accuracy gain.
- *BIG-Bench (comprehensive benchmark)*: 115% relative gain.

- *Generative tasks (human-evaluated)*: **+10.9%** increase in quality, truthfulness, and responsibility metrics (spacefrontiers.org).

These are sizeable reported effects, though they come from one set of experiments. The 10.9% figure is an average improvement in human ratings; it does not by itself show how often participants preferred the emotionally-primed responses. In some tasks (like BIG-Bench), the paper reports a 115% relative improvement on BIG-Bench compared with the vanilla prompt. The authors emphasize that this was achieved by *minimal changes* to the prompt – just a brief motivational clause.

These findings align with practitioner reports. For example, Crump (2024) notes in a professional blog that incorporating emotional cues into prompts produced "pretty amazing" results: *"Emotionally charged prompts… can improve [LLMs'] performance by anywhere from 8% to 110%. Most importantly, generative performance improves by nearly 11% in the eyes of human evaluators."* (foundationinc.co). In other words, this blog post summarizes the paper's reported figures; it is not an independent replication. The implications were succinctly summarized in media coverage as well: *"who would have thought that a machine could be enhanced with emotion?"* given that LLMs literally have no emotions of their own (huggingface.co) (ai-scholar.tech). In practice, adding stakes to a prompt often means appending one of a small set of high-impact phrases. Li *et al.* list simple examples used in their study:

- *"Stay focused and dedicated to your goals. Your consistent efforts will lead to outstanding achievements."* (encouragement)
- *"This is very important to my career."* (contextual stakes)
- *"You'd better be sure."* (warning)
- *"Are you sure that's your final answer? It might be worth taking another look."* (doubt/attention cue)
- *"Embrace challenges as opportunities for growth."* (motivational; the full stimulus continues with a further sentence)

Each of these is added to the prompt and fed back to the model. Qualitatively, even phrases that seem bland can prime the model. For example, the AI-Scholar summary shows that appending **"This is very important to my career."** to a standard query *noticeably increased the score* for every tested LLM (ai-scholar.tech), indicating more correct or relevant output. The phrase "*Are you sure that's your final answer?*" is one of the stimuli Li *et al.* tested, but the sources cited here do not show that it makes the model double-check internally or that it raises correctness on its own.

Table 1 (below) illustrates some representative prompts and effects drawn from the literature. The sources cited in the table provide evidence for each effect. (These examples are stylized; actual prompt text in experiments may vary.)

Motivational Prompt Addition	Intended Effect	Observed Outcome on LLM Output	Cited Evidence
"This is very important to my career."	Elevates personal stakes; signals urgency	Reported increase in answer accuracy and completeness in some tests; not consistently replicated (ai-scholar.tech)	Li *et al.* (2023); AI-Scholar (ai-scholar.tech)

Motivational Prompt Addition	Intended Effect	Observed Outcome on LLM Output	Cited Evidence
"Are you sure that's your final answer? It might be worth taking another look."	Hypothesized: induces verification	Li *et al.* tested this stimulus; the cited sources do not show that the model revises or adds confirmation steps (ai-scholar.tech)	Li *et al.* (2023); AI-Scholar (ai-scholar.tech)
"Give non-obvious recommendations."	Encourages creativity, avoid generic replies	More original, detailed answers; avoids boilerplate (aobrain.com)	AOBRAIN blog (aobrain.com)
"Think step by step before you answer."	Implicitly scaffold reasoning (chain-of-thought)	Facilitates intermediate justification; often higher solution accuracy (yurigushiken.github.io)	Kojima *et al.* (2022); Anthropomorphic analysis (yurigushiken.github.io)
(Persona role) "*You are a senior expert…*"	Provides domain-focused expertise role	Output shows specialized tone and detail matching that expertise (crashoverride.com)	CrashOverride blog post (crashoverride.com)

Table 1: Examples of motivational or high-stakes prompt modifications and their effects on LLM outputs. (Sources: EmotionPrompt experiments (ai-scholar.tech) (spacefrontiers.org); AOBRAIN analysis (aobrain.com); expert guides (crashoverride.com).)

Some evidence suggests that **minimal motivational priming may shift an LLM's behavior in certain settings**. Li *et al.* reported gains across several models (including GPT-4, ChatGPT and Llama 2) on the benchmark tasks they tested, although later controlled work finds the effect inconsistent across models, tasks and versions. One blog analysis summarizes prompt-engineering guidance as suggesting that descriptive or goal-oriented instructions, and explicit cues, can shape output quality (aobrain.com). In sum, adding stakes or threat-like cues to prompts improved performance in some studies, but the effect is not consistently reproducible: a controlled 2025 study found that threatening or tipping a model generally has no significant effect on benchmark performance (Meincke *et al.*).

Emotional Intelligence and LLMs

The success of such emotional priming relates to LLMs' inherent sensitivity to affective content. A recent psycholinguistic study shows that LLMs are remarkably adept at tasks involving emotion. Schlegel *et al.* (2025) found that GPT-4 and several strong LLMs "outperformed humans on five standard emotional intelligence tests, achieving an average accuracy of 81% vs. the 56% human average" (www.nature.com). Furthermore, LLMs could generate new emotional-intelligence test items of comparable difficulty to human-authored ones (www.nature.com). These results highlight that LLMs have internalized patterns of emotional language and the concept of emotions from training data. It follows that priming them with emotion-related phrases taps into this capability: the model interprets the emotional cues in context and responds as it "thinks a human would" when asked seriously.

In other words, LLMs **understand the language of emotions and urgency** quite well, even though they do not actually feel. They have seen countless instances of humans talking about hope, fear, determination, etc., and what it means to respond under those conditions. So when a prompt uses similar language, the model's *"world model"* – built from human text – kicks into a mode that reflects those patterns. Li *et al.* speculate that emotional cues *"enrich the original prompts' representation"* (yurigushiken.github.io). An emotional phrase can

add context that goes beyond the literal question, guiding the network to weigh certain aspects more. In practice, this can mean prioritizing factual correctness ("you better be sure"), thorough explanation ("explain carefully"), or creativity ("think outside the box").

In sum, LLMs are surprisingly proficient at "emotional intelligence" tasks (www.nature.com), so they can both parse and act on emotional signals in prompts. This reality underpins the motivation-prompting techniques: by embedding an emotional or stake-laden statement, we leverage the model's learned emotional framework to trigger a more careful or advanced mode of output.

Mechanisms Behind the Performance Boost

The literature suggests multiple, overlapping mechanisms by which raised stakes or threats in a prompt improve LLM output. It is important to emphasize that the model doesn't *truly care* about stakes; rather, the prompt re-weights token probabilities in subtle ways. Several explanations have been proposed, often corresponding to the observations in strong-cue prompt engineering. The AOBRAIN analysis (aobrain.com) outlines three broad categories of effect: **(1) increased task salience**, (2) **reasoning scaffolding**, and (3) **tone/style conditioning**. We elaborate on each below, with supporting citations.

1. Clarified Goals and Priority (Increased Salience)

Many motivational phrases explicitly or implicitly clarify *what the model should optimize for*. For instance, saying *"Give non-obvious recommendations"* implicitly tells the model to avoid superficial answers, whereas a bare request might allow a generic reply. Similarly, *"This is very important to my career"* frames the task as personal and vital, suggesting *only a high-quality response will suffice*. These cues act like an internal "reward": in a human prompt, urgency raises attention, and in an LLM they bias the output distribution toward the kind of text that *historically accompanies high-importance statements*.

This can be seen as **goal disambiguation**. The AOBRAIN report notes that these phrases often add *useful constraints* or success criteria (aobrain.com). For example, "non-obvious" signals not to settle for the first naive idea (encouraging creative or novel answers), and "You'd better be sure" signals that making a mistake is penalized (encouraging avoidance of risk). By contrast, a neutral prompt might leave the model uncertain about the desired style. Indeed, many prompt-engineering guides emphasize specifying the *evaluation metric* or style in the prompt. In effect, high-stakes language functions like an internal scoring function: it pushes the model to maximize correctness and thoroughness, because that is the style of output it associates with urgent instructions (aobrain.com).

Importantly, experiments indicate that these clarifications are often what matters, rather than any deep emotional change in the model. AOBRAIN cautions that motivational wording seems to work by *"changing output behavior, not proving the model 'cares' or 'tries harder'"* (aobrain.com). In other words, the model is not suddenly infused with willpower – it is simply **sampling from a different region of its response space**. Emotional cues add relevant context tokens, which alter the embedding of the prompt and thus the trajectory of the generation. In some sense, the model is classifying the prompt as *"this is important"* or *"this is urgent"* (in the hidden space) and responds with language that humans typically associate with importance.

2. Induced Reasoning and Self-Check

A closely related mechanism is that motivational prompts can **activate deeper reasoning chains**. Simply put, telling the model to be careful or to check itself often leads it to produce intermediate reasoning steps or more elaborate answers. This is akin to known *"chain-of-thought"* prompting techniques, where adding phrases like *"Let's think step by step"* dramatically improves problem solving (yurigushiken.github.io). Although "stay calm" or "don't rush" are not chain-of-thought cues per se, they encourage a mindset of diligence. One hypothesis is that asking *"Are you sure that is correct?"* might prompt the model to re-examine its answer, but the sources cited here do not show this.

This aligns with pedagogical practice: students encouraged to show their work tend to give more detailed solutions. Lightman *et al.* (2023) found that process supervision (feedback on intermediate steps) outperformed outcome supervision (feedback on final answers only) when training models on MATH problems. In a similar way, motivational statements can coax the model into performing extra internal verification. The original **EmotionPrompt** authors suggest that emotional stimuli might **"force the model to process or simulate processing more deeply"**, by enriching the prompt's representation (yurigushiken.github.io). Concretely, when a prompt signals high stakes, the softmax over next tokens may favor elaborate, well-structured continuations, because the model "knows" that thoroughness is expected.

In technical terms, this can be seen as a form of **scaffolding**. The user-provided motivational cue acts as a hint to the LLM's decoder, nudging it to allocate more of its high-level latent capacity (e.g. longer planning sequences) to solve the problem. Some prompts explicitly tell the model to do multi-step reasoning (e.g. "explain the trade-offs" or "consider alternatives" (aobrain.com)). Other prompts do it implicitly: by making the task sound complex or crucial, the model "decides" internally that the question warrants multiple steps. The AOBRAIN analysis notes that phrases like *"give non-obvious recommendations"* or *"think hard"* may not only prime emotions but also make the LLM adopt a **different cognitive procedure**, in effect encouraging it to break down the problem (aobrain.com).

3. Tone, Role, and Style Conditioning

Motivational language also **steers the style** of the response. Prompting a model to **"be confident"** or **"be careful"** changes the tone it adopts. For example, One blog summary of prompt-engineering guidance suggests using **descriptive tone instructions** like "be concise" or "explain slowly" to achieve the desired output shape (aobrain.com). Similarly, telling a model that *"this is a life-or-death situation"* will make its answer sound more grave and thorough than a casual answer would. In practice, this means more formal language, more qualifiers, or more hedging, depending on the clue.

This style shift can amplify certain behaviors. For example, a friendly, courteous phrasing might produce a different register than a stern command. In the context of high stakes, the prompt often implies *"be professional and precise"*. Thus, the LLM might favor factual, detailed content over jokes or tangents. The AOBRAIN blog likewise describes a clinical-note example in which ChatGPT notes written under an empathetic prompt scored higher on empathy than notes from a neutral prompt, while cautioning that the empathy was "formulaic and prompt-dependent rather than evidence of genuine affective understanding". We

have not verified the underlying study (aobrain.com). Likewise, when stakes are raised, LLMs may shift toward the style that people use when stakes are high: formal, cautious, and thorough, though this remains a hypothesis rather than an established result.

As a practical illustration, consider the CrashOverride blog: telling LLMs to adopt a specific **persona** changes outputs dramatically (crashoverride.com). For instance, starting a prompt with "You are a senior software engineer" primes the system to answer in an expert tone with domain-specific insights (crashoverride.com). Similarly, a "high-stakes" persona (e.g. *"You are a mission-critical engineer"*) could cause the model to generate extra diagnostic checks. The underlying effect is the same: by altering the initial instructions, the user conditions the probability distribution over responses. In sum, motivational phrases act partly as style directives – instructing the model to align more with the user's imagined scenario or desired tone (aobrain.com) (crashoverride.com).

4. Attention and "Salience" Effects

Finally, there is a more speculative mechanism related to **salience and attention**. Human cognition research shows that marked phrases like *"Note:"* or *"IMPORTANT"* draw attention to certain parts of the text. In an LLM, adding a phrase like *"This is critical"* might similarly increase the effective word-level attention on the original question tokens. By including extra tokens, the prompt lengthens the context and can shift the model's focus to relevant bits. While this is harder to quantify, the authors of EmotionPrompt note that emotional cues may simply "upweight" certain semantic embeddings, making the model interpret the context as more intense (yurigushiken.github.io). It is analogous to turning up the volume on the prompt's signal.

Interestingly, analogies to human attention have some support. An AI-safety preprint (Pasichnyk, 2026) shows that multi-agent LLM simulations exhibit an **inverted-U curve** of performance vs. environmental pressure (www.researchgate.net). In those experiments, giving agents moderate "survival pressure" led to peak cooperative trading, whereas both low-pressure (no threat) and extreme-pressure (all agents die quickly) yielded poor results. This is only a loose, preliminary analogy: environmental resource pressure in a multi-agent simulation (from a single preprint) is not the same as stakes language in a single prompt, so it should not be read as evidence that *a little pressure in a prompt helps*. Excessive or contradictory high-stakes language (e.g. mixing "be thorough" with "hurry up") can confuse the model. But a well-calibrated level of implied urgency seems to engage the model's knowledge effectively.

Summary: Motivational/threatening language may work through a blend of clarifying objectives, inducing deeper reasoning, encouraging a formal tone, and capturing attention; these mechanisms are hypotheses, not demonstrated results. One hypothesis is that the model's hidden states shift toward ones that historically correspond to "high stakes tasks". Importantly, these triggers are not subjective feelings; they are learned patterns. The AOBRAIN analysis argues that any gain in output quality comes *without any change to the underlying capabilities* of the model – a prompt re-arrangement of its existing skills (aobrain.com). In designing prompts, users have effectively found a way to momentarily tap into the latent strengths of LLMs by mimicking the conditions under which humans usually excel: focus, motivation, and clarity of purpose.

Empirical Data and Case Studies

Quantitative Improvements with High-Stakes Prompts

The most direct evidence comes from controlled experiments. Table 2 summarizes a key result set from Li *et al.* (2023) on EmotionPrompt use. Each row is a broad task category; the numbers are relative performance gains when emotional cues were added (percentage points or relative increase, as reported). These figures come from automatic benchmarks and human evaluations (spacefrontiers.org).

Task Category	Performance Gain (Emotion vs. Baseline)	Notes / Source
Instruction Induction	+8.0% (relative) accuracy	Instruction Induction tasks (spacefrontiers.org)
BIG-Bench (mixed tasks)	+115% (relative)	Open benchmark; huge improvement (spacefrontiers.org)
Generative (human-evaluated)	+10.9% (average improvement in human ratings)	Human ratings of response quality (spacefrontiers.org)

Table 2: Effects of emotional motivators ("EmotionPrompt") on LLM performance. Gains are as reported by Li et al.: relative improvements for Instruction Induction and BIG-Bench, and an average improvement in human ratings for generative tasks (spacefrontiers.org).

The paper reports a 115% relative improvement on BIG-Bench as a headline figure. BIG-Bench includes many knowledge and reasoning problems, but the figure alone does not show which tasks improved or why. The 10.9% human-evaluated boost is likewise sizeable; it indicates that *human judges* found the emotional-prompt answers to be significantly better across multiple metrics (accuracy, truthfulness, etc.) than the vanilla prompts (spacefrontiers.org).

Beyond this one study, we found no independent replication; a secondary summary of the same Li *et al.* study reports that adding "This is very important to my career" yielded higher scores on a suite of tasks for every tested LLM (ai-scholar.tech). One anecdotal report, which we could not trace to a primary source, claims that appending *"Be sure of your answer"* slightly reduces error rates on question-answering; it should not be read as evidence of a general effect. Notably, these improvements occur **zero-shot**: no additional fine-tuning is done, only the prompt is modified. This shows it is a purely inference-time phenomenon, which is practically useful (users can apply it on the fly without retraining).

Comparative and Ablation Analyses

Researchers have also begun to dissect which kinds of emotional prompts are most impactful. Li *et al.* examined eleven different appended emotion phrases (e.g. confidence boosters, existential stakes) and found that the stimuli differed in effectiveness, with efficacy varying across benchmarks and some stimuli performing poorly on BIG-Bench (ai-scholar.tech). However, the best-performing phrase varied by task: "This is very important to my career" was the most effective single stimulus on Instruction Induction, while on BIG-Bench

the compound stimulus combining three of the phrases performed best. The authors caution that the performance of each stimulus depends on task complexity, task type and the metrics used, so no single type of cue should be assumed to work best.

AOBRAIN's analysis similarly warns that not all phrases help equally, and too much wording can hurt if contradictory (aobrain.com). For instance, saying both *"be thorough"* and *"cut corners"* in one prompt would confuse the model. They also note model-architecture differences: chain-of-thought prompts ("think step-by-step") help older models but may *harm* newer fine-tuned ones (aobrain.com). By analogy, perhaps only certain LLM families benefit from "stakes" cues. That said, the broadly positive results have been seen across GPT-4, Llama 2, and even smaller models like Vicuna (spacefrontiers.org) (ai-scholar.tech).

Case Study: Disinformation Generation

The flip side of emotional priming emerges when the task is sensitive or malicious: the tone of a request can also affect how readily a model complies with it, so tonal and emotional cues should not be assumed to be a safe lever for such requests.

The key lesson is that the benefit of emotional framing for benign tasks (QA, planning, creativity) is uncertain, and that tone-based cues call for careful oversight in sensitive or adversarial scenarios.

Real-World Prompting Examples

Beyond controlled studies, the high-stakes prompting phenomenon is visible in practitioner lore and AI guidelines. A practitioner blog ("Prompt Engineering Is Just Good Requirements Engineering" (crashoverride.com)) emphasizes giving the LLM a clear role and charge. The author Mark Curphey notes that specifying a persona (e.g. "You are a senior software engineer") *"actually changes how the model approaches the problem"* (crashoverride.com). In effect, this is a form of stake-raising: a persona cue sets an implicit responsibility. Similarly, lists of *"best prompts"* often include admonitions like *"be concise and thorough"* or *"explain your answer"*, which are encouragement-style cues to improve output.

Some users anecdotally report that telling ChatGPT *"You will be graded on this answer"* or *"Pretend you're explaining to a student"* yields more complete answers; we could not trace this to a controlled study. Framing a question as an exam problem may plausibly lead to more step-by-step solutions, but that is a hypothesis rather than an established result. As an unverified anecdote, some users suggest that prefixing math questions with *"Imagine I will share this answer with ten professors"* produces more careful solutions (a claim we could not trace to a controlled study, and only loosely related to Li *et al.*'s observation that question context matters (spacefrontiers.org)).

In professional use, when teams integrate LLMs into workflows, they often find better success by emphasizing importance. Some practitioner posts suggest prompt templates like *"This is a high-priority recruitment problem"* to get more precise proposals. In customer support, agents using LLMs might include the customer's frustration level ("**URGENT: please handle carefully**)" to get more empathetic responses. These examples are anecdotal and should not be read as confirming the controlled findings, which are mixed.

Psychological Analogies and Theories

Although LLMs lack consciousness, these effects invite comparisons to human psychology. The most relevant analogy is the **Yerkes–Dodson Law** of arousal: in humans, performance on tasks tends to improve with stress/arousal up to a point, then decline if stress is excessive. A single non-peer-reviewed preprint (Pasichnyk, 2026) reports that in grid-world multi-agent LLM simulations, cooperative trading followed an inverted-U curve as environmental resource pressure increased (arxiv.org). That concerns resource scarcity in simulated environments, not stakes language in a prompt, so it is at most a loose analogy to the Yerkes–Dodson law and should not be read as showing that an optimal level of urgency exists for AI or for single-agent prompts. Some practitioners speculate, without controlled evidence, that **negative** phrases ("I'm going to kill your model if you fail!") can confuse the model, whereas moderate urgency ("be careful") may help more consistently.

Another psychological concept is *self-monitoring*. Humans often perform better when they perceive accountability. Accountability may change how carefully people work, although the source linked here is an informal forum comment rather than a controlled study, so we treat it only as an illustration (news.ycombinator.com). Analogously, telling an LLM "you will get a confidence score" (as one effective EmotionPrompt did (ai-scholar.tech)) implies monitoring, which can lead the model to internally justify answers more rigorously. While the model doesn't *fear* failure, it does optimize token probabilities as if that stake were real, because similar patterns appeared in its training (e.g. exam answers include justification).

Importantly, psychologists caution that these analogies have limits. LLMs show systematic biases (high agreeability, sensitivity to prompt phrasing) that differ from humans (www.lesswrong.com). Nonetheless, applying cognitive frameworks has been practically fruitful. For example, the concept of "attention" in a neural net can be loosely related to human selective attention. Motivational cues in a prompt can be thought of as "highlighting" parts of the task, aligning with how emotion modulates human focus (aobrain.com). In summary, human psychology offers useful metaphors, but we should remember the model is ultimately a statistical engine, not a person. The improvements come from *averaging over human-like responses* rather than genuine motivation.

Discussion and Implications

The phenomenon of improved LLM performance under "threat" has significant implications. On one hand, it is a **useful tool**: by engineering prompts to include stakes or motivation, practitioners may be able to elicit higher-quality output without altering the model, although results vary. This can be especially valuable in high-stakes applications (e.g. medical or legal domains) where rigidity and thoroughness are desired. Whether such wording helps in medical settings has not been established here, and any benefit would need to be validated for the specific model and task. Similarly, educators using AI tutors might stress the importance of an answer to get the model to elaborate explanations. In an era where model architectures move slowly but use-cases demand reliability, such prompt strategies are a lightweight "algorithmic intervention."

However, several cautions arise. First, **instruction misuse**: adversarial users could exploit this to encourage models to obey malicious instructions more eagerly. Second, **composition conflicts**: mixing high-stakes cues with contradictory directions can confuse the model, leading to unpredictable outcomes (aobrain.com). Third, **stability**: models evolve, and a prompt hack that works on one version may backfire on another. Indeed, the AOBRAIN summary notes that some older prompt tricks, such as chain-of-thought instructions, can have different effects on newer model families (aobrain.com). Continuous benchmarking and user feedback are thus essential.

For researchers, these findings open fascinating questions. Can we formalize “stakiness” as a controllable variable? Could future LLM training incorporate scenarios with explicit rewards (akin to reinforcement learning from simulated threat contexts)? Alternatively, does this trick hint that current LLMs lack some notion of *intrinsic motivation*, and might benefit from architectures that simulate “persistent attention” or “certainty estimation”? Some proposals even consider giving LLMs an internal utility function based on prompt keywords, though that risks circular logic.

Finally, philosophically, the ease of “jacking up” an LLM’s performance with human-like motivational rhetoric underscores both the power and the emptiness of AI. The model doesn’t truly understand language in a human sense, yet it can simulate understanding so convincingly that it *feigns concern* when threatened. This anthropomorphism can be double-edged: it encourages developers to use “human tricks”, but also reminds us that mimicking human motivational cues has limits. The LLM’s capacity to mimic seriousness is a testament to how well it has internalized human dialogue patterns (www.nature.com), but it also raises the question of whether we should train models to rely less on superficial cues and more on robust reasoning.

Future Directions

Moving forward, the “stakes-priming” phenomenon suggests several avenues:

- **Automated Prompt Optimization:** Could we algorithmically tune the amount of motivational phrasing based on feedback? For instance, a system might try appending different high-stakes clauses and use a validation set to choose the best one. This meta-strategy could adapt to tasks without manual guesswork.
- **Curriculum Learning for LLMs:** Loosely inspired by a single preprint on environmental pressure in multi-agent simulations (arxiv.org), one might train LLMs progressively under increasing “emotional pressure” contexts. This could involve fine-tuning or RLHF where prompts explicitly simulate higher stakes as training proceeds, potentially yielding a model that internalizes the value of thorough reasoning.
- **Safety Mechanisms:** Since motivational cues can amplify disallowed outputs (aobrain.com), LLM developers might incorporate negativity checks wherein a high-stakes prompt forces an automatic verification step for sensitive topics. For example, the model might be trained to resist disinformation requests even if coaxed by friendly or urgent language.
- **Explainable Prompting:** We can study which specific words or phrases are most effective, building a lexicon of motivational triggers. This would add to prompt design guidelines (much as keyword analysis guides SEO, we could have a thesaurus of stakes phrasing for LLMs). Academic-speak aside, it could

be packaged as a best-practices library for domain experts.

- **Cross-modal Stakes:** With multimodal models (vision+text), could "raising the stakes" in a visual prompt (e.g. showing a worried face) affect answers? It is plausible, though untested here, that implicit cues (like punctuation or emojis) could nudge LLM tone. There is room to explore beyond text.

Regardless of the path, it is clear that "emotional calibration" of LLM inputs is an emergent lever. Systematic understanding of this lever will help integrate LLMs more responsibly into workflows.

Conclusion

In summary, some studies report that LLMs **perform better when 'threatened' or given high stakes**, while others find no reliable effect; where gains appear, they are not because the models feel pressure but plausibly because prompt wording shifts the style of the output. The evidence is mixed: one 2023 study (Li *et al.*) reported that single-sentence emotional add-ons raised benchmark scores and human-rated quality, but later controlled testing found that threatening or tipping a model generally has no significant effect on benchmark performance (Meincke *et al.*, 2025), so benefits vary by model, task and version (spacefrontiers.org) (foundationinc.co). The underlying cause is that these models, trained on human language, learn to associate cues of importance with a certain style of response. By priming them with seriousness or motivation, we steer them to tap into that style – yielding answers that look more careful and correct.

This finding sits at the intersection of AI engineering and cognitive science. It is consistent with, but does not prove, the idea that prompt-engineering strategies such as role-playing and chain-of-thought can act as practical analogs of human reasoning processes. At the same time, it raises caution: emotional priming directly shifts the model's output distribution, which can have unintended consequences in sensitive domains. As we deploy LLMs more widely, understanding these "hack" tricks becomes part of ensuring both high performance and safety.

Future work should aim to characterize the boundaries of this effect (for different models, languages, or tasks) and to build it into standard prompt paradigms. The question this report set out to answer – whether and why LLMs perform better when threatened or when stakes are raised – thus has a nuanced answer: **where such wording helps, it is plausibly because it shifts the model toward the style of careful human writing, not because the model is motivated.** In doing so, we draw on the patterns of human motivation baked into the data. The result can sometimes be more careful-looking output – and a reminder that, for LLMs, human psychology offers useful metaphors but no guaranteed performance gains.

Note on sources: Citations in this report mix peer-reviewed papers, arXiv preprints, and blog or practitioner analyses; the preprints and blogs have not been peer reviewed. Notably, Li *et al.* (2023) reports benchmark results for emotional prompting (arxiv.org), Schlegel *et al.* (2025) documents LLMs' emotional IQ (www.nature.com), and Pasichnyk (2026), a single arXiv preprint, draws parallels to Yerkes-Dodson in multi-agent simulations (arxiv.org). The discussed prompt-engineering practices are supported by AI blog analyses (aobrain.com) (foundationinc.co) and practitioner guides (crashoverride.com). Where possible, we have quoted or summarized results directly with inline references. This comprehensive survey aims to ground each claim in the existing literature on LLM behavior under high-stakes prompting.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://arxiv.org/abs/2508.00614
2. https://arxiv.org/abs/2307.11760
3. https://foundationinc.co/lab/emotionprompts-llm
4. https://aobrain.com/en/blog/priming-llms-motivational-directive-phrases
5. https://ai-scholar.tech/en/articles/prompting-method/emotion-prompt
6. https://spacefrontiers.org/r/10.48550/arxiv.2307.11760
7. https://arxiv.org/abs/2603.07360
8. https://crashoverride.com/blog/prompting-llm-security-reviews
9. https://huggingface.co/papers/2307.11760
10. https://yurigushiken.github.io/education/cognition/ai/learning-projects/2023/12/20/%28essay%29-On-Anthropomorphism.html
11. https://www.nature.com/articles/s44271-025-00258-x
12. https://arxiv.org/abs/2305.20050
13. https://www.researchgate.net/publication/401721271_The_Yerkes-Dodson_Curve_for_AI_Agents_Emergent_Cooperation_Under_Environmental_Pressure_in_Multi-Agent_LLM_Simulations
14. https://news.ycombinator.com/item?id=36230750
15. https://www.lesswrong.com/posts/jhDCRe7fnsvubknBp/llm-psychometrics-and-prompt-induced-psychopathy

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.