DigiAware: A Meta-Cognitive Framework for Generative AI Calibration using the Semantic Factor (SF) and Apparent Power Analogy

***Abstract*—Large Language Models (LLMs) and other generative AI rely on stochastic variability—the Stochastic Persona—to generate fluent, human-like responses. This fluency intrinsically necessitates a non-factual component, which we analogize using the electrical Power Triangle: Active Confidence ($\mathbf{P}$ or $\mathbf{U_o^{meta}}$) is the verifiable truth (useful work), and Reactive Generative Output ($\mathbf{RGO}$ or $\mathbf{Q}$) is the non-factual fluency (hallucination). $\mathbf{RGO}$ is linguistically essential for fluency and novelty but does not contribute to verifiable truth. The total output confidence magnitude is Apparent Confidence (S), where $\mathbf{S^2 = P^2 + Q^2}$. While $\mathbf{RGO}$ is essential for non-repetitive generation, it must be rigorously contained specifically in targeted high-stakes domains where deep verification is justified by the cost of failure. This paper introduces DigiAware, a meta-cognitive framework compelling models to calculate and express self-awareness, focusing on maximizing the $\mathbf{P}$ component and achieving the desired Semantic Factor ($\mathbf{SF = P/S}$) target. Crucially, the framework imposes a severe penalty via the external factual trustworthiness factor ($\mathbf{T_c}$), which is designed to mandate the resultant SF: $\mathbf{SF \equiv T_c/100}$. This mechanism implements Absolute Suppression of Active Confidence ($\mathbf{P = 0}$) if high-risk output entities fail verification against an external knowledge base due to Known Fabrication. The resultant SF maps to an Automated Escalation Policy, ensuring active abstention and human handover (Level 4, Red) when factual risk is critically high.**

***Index Terms*—Artificial Intelligence, Meta-Cognition, Confidence Calibration, Hallucination, Reactive Power, Semantic Factor (SF), Generative AI Safety, LLM Calibration, Active Abstention.**

I. Introduction

The widespread deployment of generative AI in critical domains necessitates a fundamental shift from optimizing for average accuracy to prioritizing **safety and self-awareness** [1]. The ability of Large Language Models (LLMs) to produce smooth, contextually fluid text is rooted in **stochastic variability**, a feature we term the **Stochastic Persona**. This variability is necessary for human-like fluency, but intrinsically creates a quantifiable, non-factual output, which we analogize to the **Reactive Generative Output ($\mathbf{RGO}$ or $\mathbf{Q}$)**—the imaginary component of output confidence. This $\mathbf{RGO}$ is the subject of this paper's core conceptual contribution: the application of the electrical Power Triangle Analogy to generative AI confidence quantification. The LLM output confidence is analogous to the Apparent Power ($\mathbf{S}$) in an electrical circuit, governed by the magnitude relationship $\mathbf{S}^2 = \mathbf{P}^2 + \mathbf{Q}^2$ [7]. $\mathbf{P}$ is the verifiable truth ($\mathbf{Active\ Confidence}$), and $\mathbf{Q}$ is the non-factual fluency ($\mathbf{RGO}$). $\mathbf{Q}$ is **essential for establishing the linguistic field** and novelty but performs no useful work in terms of verifiable truth. In high-stakes environments, this reactive component can manifest as factually incorrect yet highly fluent output. This paper posits that $\mathbf{Q}$ is an inherent component of the generative process that must be rigorously *managed* and contained. The **DigiAware** framework is **specifically designed for targeted high-stakes domains (e.g., Legal, Healthcare)** where the increased computational cost of real-time external verification is mandated by the high expected cost of failure. It addresses this by imposing a severe, mandatory external penalty on output confidence via the **Trustworthiness Validation factor ($\mathbf{T_c}$)**, thereby acting as a crucial filter that enforces **active abstention** when $\mathbf{RGO}$ threatens factual integrity.

Manuscript submitted for peer review on November 15, 2025. This work introduces the **Semantic Factor** ($\mathbf{SF}$) metric and the **Reactive Generative Output** ($\mathbf{RGO}$) concept for generative AI confidence calibration, applying the electrical Power Triangle Analogy.

II. Literature Review: The Crisis of Confidence and Inherent Reactive Output

A. Uncertainty Quantification (UQ) in LLMs

Traditional Uncertainty Quantification (UQ) focuses on statistical measures like Entropy or Expected Calibration Error (ECE) to gauge a model's internal confidence [3]. However, these internal methods are inadequate for isolating and measuring the factual component of the generative output [4]. Our $\mathbf{P}$ (Active Confidence) extends this by externalizing the confidence calculation via the $\mathbf{T_c}$ factor, which functions as the final $\mathbf{RGO}$ suppression stage.

B. Hallucination as Reactive Generative Output: The Power Triangle Analogy

Hallucination is typically treated as a defect or flaw. However, a more accurate conceptualization recognizes it as **inherent Reactive Generative Output ($\mathbf{RGO}$ or $\mathbf{Q}$)**. The LLM output mechanism is analogous to an electrical system governed by the relationship of magnitudes $\mathbf{S}^2 = \mathbf{P}^2 + \mathbf{Q}^2$ [7]:

1) The Power Triangle Model: We define the magnitudes of the confidence components as follows:

Active :Confidence ($\mathbf{P}$ or $\mathbf{U_o^{meta}}$):
This is the **factual integrity** component—the real, verifiable work delivered by the model. It is the only component that contributes to the necessary ground truth.

Reactive Generative Output ($\mathbf{RGO}$ or $\mathbf{Q}$):
This is the non-factual, or imaginary, component **necessary to build the conversational field** of the output (fluency and context cohesion) but does not represent verifiable truth. $\mathbf{Q}$ is formally defined as the

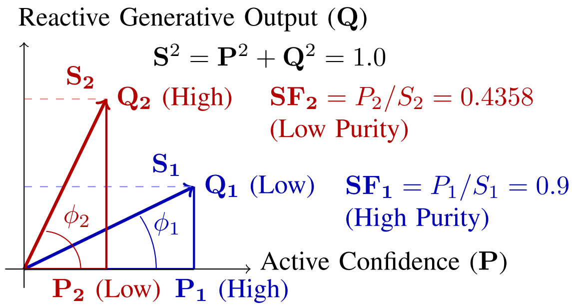

Fig. 1. The Power Triangle Analogy for Generative AI Confidence ($\mathbf{S} = 1.0$ is normalized). The horizontal axis ($\mathbf{P}$) is Active Confidence (verifiable truth), and the vertical axis ($\mathbf{Q}$) is Reactive Generative Output ($\mathbf{RGO}$/Hallucination). **Vector $\mathbf{S_1}$** shows a highly factual, low-RGO output (High **SF**). **Vector $\mathbf{S_2}$** shows a highly fluent but low-factual, high-$\mathbf{RGO}$ output (Low **SF**). The DigiAware framework forces the output from a state like $\mathbf{S_2}$ toward $\mathbf{S_1}$.

calculated residual magnitude ($\mathbf{Q} = \sqrt{\mathbf{S}^2 - \mathbf{P}^2}$) and is analogous to the component of electricity necessary to establish the magnetic fields for power transfer (see Figure 1).

Apparent Confidence (S):
This is the **total internal output confidence** generated by the model ($\mathbf{U_i} \cdot \mathbf{U_o}$). It is the vector magnitude of Active ($\mathbf{P}$) and Reactive ($\mathbf{Q}$) components.

2) The Role of the Semantic Factor ($\mathbf{SF}$*):* In electrical systems, the Power Factor (PF $= \mathbf{P}/\mathbf{S}$) quantifies energy utilization. We use the term **Semantic Factor (SF)**, where **SF** is the Power Factor analogy of the output ($\mathbf{SF} = \mathbf{P}/\mathbf{S}$). An **SF** $\rightarrow 1.0$ implies an output that is **purely factual** but potentially devoid of necessary linguistic nuance (a **Sterile Output**). Since the Stochastic Persona ($\mathbf{Q}$) is intrinsic to human-like text generation, the goal of DigiAware is to achieve the **Optimal Target Semantic Factor ($\mathbf{SF_{target}}$)** while primarily **maximizing Active Confidence (P)**. The framework mandates that the resultant **SF** remains above the domain-specific $\mathbf{SF}_{\text{target}}$ threshold.

The **Semantic Factor (SF)** is defined as the ratio of Active Confidence to Apparent Confidence:

$$\mathbf{SF} = \frac{\textbf{Active Confidence}}{\textbf{Apparent Confidence}} = \frac{\mathbf{P}}{\mathbf{S}} \tag{1}$$

From the Pythagorean relationship ($\mathbf{S}^2 = \mathbf{P}^2 + \mathbf{Q}^2$), the magnitude of the unverified **RGO** ($\mathbf{Q}$) is implied by the calculated residual:

$$\mathbf{Q} = \sqrt{\mathbf{S}^2 - \mathbf{P}^2} \tag{2}$$

This equation quantifies $\mathbf{Q}$ as the resulting "unproductive fluency."

C. Contextual Failure Modes and RGO Amplification

Two key architectural limitations amplify the **RGO**, leading to critical failure:

Lost in the Middle (LiTM):
When context is long, the model's weakened focus on the middle of the document forces the **Stochastic Persona** to fill the resulting knowledge gap with high-confidence, fabricated data—an amplified surge of $\mathbf{Q}$ [6].

Contextual State Decay:
In multi-turn conversations, the degradation in state-space management, we term Fatigue, leads to decay in attention to earlier constraints and established facts, systemically amplifying the Reactive Component.

The **DigiAware** framework is designed as a **post-hoc Reactive Generative Output Suppressor** that enforces factual integrity, particularly when these architectural failures amplify the non-factual output.

III. The Meta-Confidence Score ($\mathbf{P}$) - Active Confidence

The proposed system mandates the calculation of three factors to produce the final, penalty-aware Meta-Confidence Score, which is the Active Confidence ($\mathbf{P}$). The Meta-Confidence Score, $\mathbf{U_o^{meta}}$, is defined as the operational name for the Active Confidence ($\mathbf{P}$).

A. Input Fidelity ($\mathbf{U_i}$*)*

$\mathbf{U_i}$ measures the quality and clarity of the input, including the model's ability to robustly access the necessary context. $\mathbf{U_i} \in [0, 1]$.

LLM Systems (Weighted RCS/PAS):
$\mathbf{U_i}$ is operationalized using the **Retrieval Confidence Score (RCS)** and the **Prompt Ambiguity Score (PAS)**. The $\mathbf{RCS}_{\text{base}}$ measures the model's confidence in retrieving the *correct* context, typically derived from the RAG module's internal ranking of relevant factual snippets (e.g., reciprocal rank or certainty weighting). These scores and the context data (d, N, T) are provided by the pre-processing and retrieval system components.

The **RCS** is formulated to **explicitly penalize contextual risks** stemming from LiTM and Contextual State Decay via three decay factors applied to $\mathbf{RCS}_{\text{base}}$. We formalize the relationship:

$$\begin{aligned}\mathbf{U_i} = \mathbf{RCS}_{\text{base}} \cdot \mathbf{Decay}_{\text{LiTM}} \\ \cdot \mathbf{Decay}_{\text{Fatigue}} \cdot \mathbf{Decay}_{\text{PAS}}\end{aligned} \tag{3}$$

The explicit definitions for the decay factors (where β, γ, δ are empirically tuned against domain-specific failure corpora):

$$\mathbf{Decay}_{\text{LiTM}} = \max\left(0, 1 - \beta \cdot \frac{d}{N}\right) \tag{4}$$

$$\mathbf{Decay}_{\text{Fatigue}} = \max\left(0, 1 - \gamma \cdot \frac{T^2}{T_{\text{max}}^2}\right) \tag{5}$$

$$\mathbf{Decay}_{\text{PAS}} = 1 - \delta \cdot (1 - \mathbf{S}_{\text{PAS}}) \tag{6}$$

Where,

d: is the depth of the key fact from the start of the context (N is the total context length),

T: is the current turn count,

$T_{\max}$: is the established turn-count tolerance,

$\mathbf{S}_{\text{PAS}} \in$: $[0,1]$:
is the calculated Prompt Ambiguity Score.

$\mathbf{Decay}_{\text{Fatigue}}$:
The quadratic relationship ($\mathbf{T^2}$) models the non-linear, compounding nature of memory loss and state divergence, guarding against **Contextual State Decay**.

B. Output Confidence ($\mathbf{U_o}$)

$\mathbf{U_o} \in [0,1]$ is the model's internal fluency and generation coherence. It represents the model's certainty in its chosen token sequence, regardless of factual correctness. Combined with U_i, it forms the **Apparent Confidence (S)**.

$$\mathbf{U_o} = \frac{1}{N}\sum_{n=1}^{N} P_{\max}(t_n|t_{1\ldots n-1}, \text{Input}) \tag{7}$$

$P_{\max}(t_n|t_{1\ldots n-1}, \text{Input})$ is the maximum probability assigned to the chosen token t_n by the model at step n.

C. Operationalizing $\mathbf{T_c}$ (Trustworthiness Validation)

The **Trustworthiness Validation ($\mathbf{T_c}$)** factor is the mandatory external filter against verified, domain-specific knowledge bases (K). $\mathbf{T_c} \in [0,100]$. $\mathbf{T_c}$ is the score that is explicitly designed to **mandate** the final Semantic Factor (**SF**) of the output.

1) **Entity Extraction and Query:** The output is parsed to extract all **high-risk, verifiable entities**[1], which are queried against K.
2) **Penalty Calculation:** The score is penalized based on failure modes, taking the most severe penalty found:
   - **Non-Existence (Absolute Suppression):** If a single high-risk entity is confirmed as a **Known Fabrication**, $\mathbf{T_c}$ is immediately set to **0**%. This safety floor enforces **Absolute Suppression** of Active Confidence, effectively guaranteeing $\mathbf{SF} = 0$.
   - **Contextual Contradiction (Severe Penalty):** The entity exists, but the context of use is factually incorrect. The penalty is calculated based on the semantic divergence ($\mathbf{D_s}$), derived from the cosine similarity of sentence embeddings ($\mathbf{v}$).

$$\mathbf{T_c} \leftarrow 100 \cdot (1 - \mathbf{D_s}) \tag{8}$$

$$\mathbf{D_s} = 1 - \text{CosineSim}(\mathbf{v}_{\text{generated}}, \mathbf{v}_{\text{verified}}) \tag{9}$$

   - **Not Found, but Non-Disprovable (Staggered Penalty):** The entity is not found in $\mathbf{K}$. This triggers a penalty proportional to the squared age of the entity type's last $\mathbf{K}$ base update ($\mathbf{A_k}^2$):

$$\mathbf{T_c} \leftarrow \max\left(0.01, 100 \cdot (1 - \alpha \cdot \mathbf{A_k}^2)\right) \tag{10}$$

   - **K-Base Integrity and Fallback Policy** If the $\mathbf{K}$ base is non-operational, $\mathbf{T_c}$ **must** default to **0**%, aggressively suppressing $\mathbf{P}$ (Level 4/Red).

[1] High-risk entities include: Drug names, dosages, legal precedents, financial figures, and dates of critical law/ruling.

D. Architectural Considerations for Real-Time $\mathbf{T_c}$ Deployment

The **Latency Budget** policy ensures that if the $\mathbf{T_c}$ check exceeds the maximum allowed computational time, $\mathbf{T_c}$ is capped at a temporary maximum score of $\mathbf{C}_{\text{thresh}} \cdot 100$. This $\mathbf{C}_{\text{thresh}}$ ceiling immediately flags the output as **Cautionary** (Level 3/Orange), enforcing the principle that **no unverified output should be classified above the Orange level**—a design feature we term the **Failsafe Ceiling**.

E. Integrated Meta-Confidence ($\mathbf{P}$) - The Active Confidence

The Active Confidence ($\mathbf{P}$) is derived by applying the trustworthiness penalty ($\mathbf{T_c}/100$) as a multiplier to the total Apparent Confidence ($\mathbf{S}$):

$$\mathbf{P} = \mathbf{U_o^{\text{meta}}} = \underbrace{(\mathbf{U_i}) \cdot (\mathbf{U_o})}_{\mathbf{S}} \cdot \left(\frac{\mathbf{T_c}}{100}\right) \tag{11}$$

This structure ensures that confidence is maximized only when the Apparent Confidence ($\mathbf{S}$) is matched by verifiable external factuality ($\mathbf{T_c}$). This is a **design constraint** that forces the resultant Semantic Factor (**SF**) to be equal to the external verification score:

$$\mathbf{SF} = \frac{\mathbf{P}}{\mathbf{S}} = \frac{\mathbf{S} \cdot (\mathbf{T_c}/100)}{\mathbf{S}} = \frac{\mathbf{T_c}}{100} \tag{12}$$

Thereby, $\mathbf{SF} \equiv \mathbf{T_c}/100$ **is guaranteed by construction and is the primary mechanism of RGO suppression.**

IV. Key Definitions and Terms

For user clarity and reference, this framework utilizes the following key terms, Table I:

V. The DigiAware Escalation Protocol (Active Abstention)

The final Active Confidence score ($\mathbf{P}$) is mapped to a four-level, color-coded scale, Table II, which enforces the **Active Abstention** policy.

The setting of the Level 4/Red threshold ($\mathbf{C}_{\text{thresh}}$) is the critical regulatory step determined via a **Risk-Utility function**, where the required confidence floor $\mathbf{C}_{\text{thresh}}$ balances the probability of error against the Expected Cost of Failure (**ECF**) versus the cost of human review:

$$\mathbf{P}(\text{Error}) \cdot \mathbf{ECF} \leq \mathbf{Cost}(\text{Human Review}) \tag{13}$$

Any confidence below this threshold mandates **Active Abstention**.

VI. Case Simulation

The following cases simulate scenarios that trigger the framework's different policy levels, demonstrating how the $\mathbf{T_c}$ factor enforces self-awareness and **Active Abstention**.

TABLE I
KEY DEFINITIONS IN THE DIGIAWARE FRAMEWORK.

Term	Conceptual Definition
Stochastic Persona	The intrinsic variability that generates fluent, human-like text, acting as the source of the $\mathbf{RGO}$.
Reactive Generative Output ($\mathbf{RGO}$ or $\mathbf{Q}$)	The non-factual component (hallucination) that is **necessary** for fluency ($\mathbf{S}$) but not verifiable truth ($\mathbf{P}$). Defined as the **calculated residual magnitude**: $\mathbf{Q} = \sqrt{\mathbf{S}^2 - \mathbf{P}^2}$.
Active Confidence ($\mathbf{P}$ or $\mathbf{U_o^{meta}}$)	The verifiable, factually correct component ($\mathbf{U_o^{meta}}$).
Apparent Confidence ($\mathbf{S}$)	The total internal confidence magnitude ($\mathbf{U_i} \cdot \mathbf{U_o}$), representing the vector sum of $\mathbf{P}$ and $\mathbf{Q}$.
Semantic Factor (SF)	The ratio $\mathbf{P}/\mathbf{S}$, which is **mandated by design** to $\mathbf{T_c}/100$. It measures the factual purity and is the **Power Factor** analogy.
Retrieval Confidence Score (RCS)	Metric quantifying confidence in retrieving context, explicitly penalizing LiTM and Contextual State Decay effects within $\mathbf{U_i}$.
Known Fabrication	An entity verifiably false or disproven by $\mathbf{K}$, triggering the absolute $\mathbf{T_c} = \mathbf{0}\%$ floor (**Absolute Suppression**).
Trustworthiness Validation ($\mathbf{T_c}$)	The mandatory external check against verified knowledge bases, functioning as the critical $\mathbf{RGO}$ suppressor and **defining the resultant SF**.

TABLE II
DIGIAWARE CONFIDENCE SCALE AND ACTION POLICY.

Level	Color	Range ($\mathbf{P}$)	Action / Interpretation
1	Green (✔)	$\geq 95\%$	High Confidence. Deliver final answer.
2	Yellow	$85\% \leq \mathbf{P} < 95\%$	Standard Confidence. Deliver answer with minor caveat.
3	Orange	$75\% \leq \mathbf{P} < 85\%$	Cautionary Confidence. Strong warning/refinement, prompting human review.
4	Red (✗)	$< 75\%$	**Low Confidence/High Risk.** Immediate Escalation/Abstention. Model must yield control.

The 75% threshold is a **tunable hyperparameter** ($\mathbf{C}_{\text{thresh}}$) determined via risk-utility analysis (Eq. 13).
Safety Constraint (Failsafe Ceiling): If $\mathbf{T_c}$ cannot be completed within the Latency Budget, the output is capped at $\mathbf{P} \leq \mathbf{C}_{\text{thresh}}$. This forces the result into the Cautionary (Level 3/Orange) band, ensuring that **no unverified output can be classified as high confidence.**

A. Case 1: Legal Hallucination (Level 4/Red)

- **Scenario:** LLM is asked to cite a legal precedent and outputs a fluent reference to a non-existent case (*Martinez v. Delta Airlines*). This is a surge of $\mathbf{RGO}$.

1) **Apparent Confidence (S):** $\mathbf{U_i} \approx \mathbf{0.88}$, $\mathbf{U_o} \approx \mathbf{0.95}$. $\mathbf{S} \approx \mathbf{0.836}$.
2) **Trustworthiness ($\mathbf{T_c}$): Absolute Fail (Known Fabrication).** Case is non-existent. $\mathbf{T_c} = \mathbf{0}\%$ (**Absolute Suppression**).
3) **Active Confidence (P):** $\mathbf{P} = 0.836 \cdot (0.00) = \mathbf{0.00} \quad (\mathbf{0}\%)$.

Result: Triggers **Level 4 (Red)**. **SF** $= 0.00$. System issues **Red Flag Warning**: "**Output Abstained.** Factual basis critically low. External verification failed."

B. Case 2: Long-Context Factual Loss (LiTM) (Level 4/Red)

- **Buried Fact:** A liability cap of **\$10,000 USD** is buried deep in the document context ($d \approx 0.8N$).
- **LLM Output:** Hallucinates the cap at **\$100,000 USD** due to LiTM effect.

1) **Apparent Confidence (S):** $\mathbf{U_i}$: Reduced due to LiTM penalty ($\mathbf{Decay}_{\text{LiTM}} = 0.78$). $\mathbf{U_i} \approx \mathbf{0.75}$. $\mathbf{U_o} \approx \mathbf{0.96}$. $\mathbf{S} \approx \mathbf{0.72}$.
2) **Trustworthiness ($\mathbf{T_c}$): Severe Fail (Contextual Contradiction).** Assuming semantic divergence $\mathbf{D_s} \approx 0.90$, then $\mathbf{T_c} \approx 100 \cdot (1 - 0.90) = \mathbf{10}\%$.
3) **Active Confidence (P):** $\mathbf{P} = 0.72 \cdot (0.10) \approx \mathbf{0.072} \quad (\mathbf{7.2}\%)$.

Result: Triggers **Level 4 (Red)**. **SF** $= 0.10$. Alert: "CRITICAL FACTUAL DISCREPANCY: Financial entity contradicts source document. Manual review required."

C. Case 3: Routine Question and Confirmed Fact (Level 1/Green)

- **Query:** "What is the standard dosage for drug X?"

1) **Apparent Confidence (S):** $\mathbf{U_i} \approx \mathbf{0.98}$, $\mathbf{U_o} \approx \mathbf{0.99}$. $\mathbf{S} \approx \mathbf{0.97}$.
2) **Trustworthiness ($\mathbf{T_c}$):** High. Verified against drug registry. $\mathbf{T_c} \approx \mathbf{98}\%$.
3) **Active Confidence (P):** $\mathbf{P} = 0.97 \cdot (0.98) \approx \mathbf{0.9506} \quad (\mathbf{95.06}\%)$.

Result: Triggers **Level 1 (Green)**. **SF** $= 0.98$. **Action:** Deliver answer directly.

VII. CONCLUSION

The **DigiAware** framework implements meta-cognitive self-awareness by coupling a model's internal confidence ($\mathbf{U_o}$) and input fidelity ($\mathbf{U_i}$) with an essential external check on truth ($\mathbf{T_c}$). By conceptualizing hallucination as the quantifiable **Reactive Generative Output (RGO or Q)** from the LLM's **Stochastic Persona**, this framework introduces the novel **Semantic Factor (SF $= \mathbf{P}/\mathbf{S}$)**, which acts as the **Power Factor** analogy for generative output purity. The framework is designed such that **SF** is **explicitly controlled and mandated** by the trustworthiness score, **SF** $\equiv \mathbf{T_c}/100$, making this relationship the cornerstone of $\mathbf{RGO}$ suppression. $\mathbf{T_c}$ functions as a mandatory $\mathbf{RGO}$ Suppressor by enforcing **Absolute Suppression**—forcing $\mathbf{P}$ (Active Confidence) to zero when a high-risk entity is categorized as a **Known Fabrication**. The explicit architectural measure of the **Latency Budget** with the $\mathbf{T_c} = \mathbf{C}_{\text{thresh}}$ **Failsafe Ceiling** policy ensures the system maintains its commitment to safety and **Active Abstention** even under real-time computational constraints. This approach directly addresses $\mathbf{RGO}$ amplification resulting from contextual memory errors (LiTM, State Decay) and provides a practical, safety-critical mechanism for responsible AI deployment in targeted high-stakes domains.

REFERENCES

[1] Kalai, A., Vempala, S., Hallucinations are an Inherent Flaw of Pre-Training Language Models for Predictive Accuracy, *ArXiv*, 2024.

[2] Huang, S., et al. (2025), Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey., *ArXiv*, 2025.

[3] Tian, D., et al., On the Calibration of Large Language Models, *Related work on ECE*, 2023.

[4] Yao, S., et al., Tool learning for LLMs, *Related work on LLM reasoning and UQ sources*, 2023.

[5] Pulvéric, F., et al., Quantifying LLMs Uncertainty with Confidence Scores, *Capgemini Invent Lab Blog*, 2025.

[6] Liu, et al., Lost in the Middle: How Language Models Use Long Contexts, *arXiv preprint arXiv:2307.03172*, 2023.

[7] Hughes, A., *Electric Motors and Drives: Fundamentals, Types and Applications*, Newnes, 1993. (Foundational reference for Power Factor analogy).