

CT-SAFR: Safe and Interpretable Chain-of-Thought Reasoning for Autonomous Robots

A Multi-Layered Verification Framework for Trustworthy AI-Driven Robotic Decision Making

Cagri Temel, IEEE Senior Member

Hezarfen LLC

Seattle, WA, USA

Grand Canyon University

Phoenix, AZ, USA

cagritemel@ieee.org

Abstract—Chain-of-Thought (CoT) prompting enables LLMs to perform explicit, step-by-step reasoning, creating opportunities for sophisticated autonomous robots. However, recent research reveals that reasoning models verbalize their actual decision processes only 25–39% of the time, with faithfulness degrading 44% on complex tasks. This paper presents CT-SAFR (Chain-of-Thought Safety and Faithfulness for Robotics), a multi-layered verification framework achieving 94.2% hallucination detection ($n = 500$, 95% CI: 91.8–95.9%) with sub-500ms latency. Through a warehouse robot case study, this work demonstrates 87% reduction in unsafe reasoning outputs ($p < 0.001$) and provides recommendations for responsible deployment of reasoning-capable autonomous robots.

Index Terms—Chain-of-Thought Reasoning, Autonomous Robots, LLM Safety, Interpretable AI, Trustworthy Autonomy, Multi-Layer Verification, Self-Consistency, Warehouse Robotics

I. INTRODUCTION

Large Language Models (LLMs) with Chain-of-Thought (CoT) prompting [1] are increasingly integrated into robotic systems for enhanced decision-making and task planning [2], [3]. CoT-enabled robots engage in explicit, step-by-step reasoning that decomposes complex tasks into manageable sub-problems. However, autonomous robots operate in physical environments where reasoning errors can result in equipment damage, environmental harm, or human injury.

Recent empirical research has revealed troubling findings about CoT faithfulness. Chen et al. [4] demonstrated that state-of-the-art reasoning models verbalize their actual decision processes only 25–39% of the time, with faithfulness decreasing by 44% on complex tasks. These findings fundamentally challenge the assumption that visible reasoning chains provide reliable insight into model behavior.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. This is the accepted version. Published in: 2026 IEEE Conference on Artificial Intelligence (CAI), pp. 598–603, May 2026. DOI: 10.1109/CAI68641.2026.11536646. The version of record is available on IEEE Xplore.

A. Contributions

This paper makes the following key contributions:

- This paper presents **CT-SAFR**, a multi-layered verification framework specifically designed for safe deployment of Chain-of-Thought reasoning in autonomous robots. This represents one of the first system-level frameworks addressing faithfulness, physical grounding, and temporal consistency simultaneously.
- This work introduces a **defense-in-depth architecture** with four complementary verification layers (structural, physical, semantic, and interpretability), achieving 94.2% hallucination detection ($n = 500$, 95% CI: 91.8–95.9%) and 96.4% combined unsafe reasoning detection while maintaining sub-500ms latency suitable for real-time robotic control.
- This work adapts **self-consistency decoding** for robotic applications, demonstrating that consensus across multiple reasoning paths provides reliable safety guarantees even when individual reasoning traces may be unfaithful to actual model computation.
- This paper provides **comprehensive empirical evaluation** through a warehouse robot case study, demonstrating 87% reduction in unsafe reasoning outputs ($p < 0.001$) and 3.5% improvement in task completion rates across 1,000 operation hours.
- This paper conducts **ablation studies** quantifying the individual contribution of each verification layer, establishing that all layers provide complementary safety benefits beyond what any single mechanism can achieve.
- This paper offers **concrete recommendations** for researchers, practitioners, and policymakers on standardized benchmarks, regulatory frameworks, and best practices for responsible deployment of reasoning-capable autonomous robots.

This work builds on the CogniTest verification framework [5], adapting these techniques for safety-critical robotic applications with unique physical grounding and real-time con-

TABLE I
COMPARISON WITH EXISTING SAFETY APPROACHES. ✓ = ADDRESSED, ~
= PARTIAL, — = NOT ADDRESSED.

Approach	Struct.	Phys.	Sem.	Interp.
SayCan [3]	—	~	—	—
Yang et al. [9]	—	✓	—	—
Rule-based filter	—	✓	—	—
Safety prompting	~	—	~	—
CT-SAFR	✓	✓	✓	✓

straints. The core multi-layered safety architecture described in this work is the subject of a pending U.S. provisional patent application [6].

II. BACKGROUND AND RELATED WORK

A. Chain-of-Thought Reasoning in LLMs

CoT prompting [1] enables LLMs to generate intermediate reasoning steps, notably tripling solve rates on the GSM8K math benchmark (18% to 57% with PaLM 540B) and improving commonsense reasoning on StrategyQA (75.6% vs. prior best 69.4%). Kojima et al. [7] showed zero-shot CoT via “Let’s think step by step.” Wang et al. [8] introduced self-consistency decoding, achieving up to 17.9% accuracy gains through majority voting across multiple reasoning paths.

B. LLMs for Robotic Planning and Control

Huang et al. [2] demonstrated language models as zero-shot planners, while Ahn et al. [3] introduced SayCan, grounding language in robotic affordances. Yang et al. [9] proposed enforcing constraints for LLM-driven robot agents but focused on specification rather than comprehensive verification.

Table I positions CT-SAFR relative to existing approaches. SayCan [3] provides affordance grounding but lacks explicit physical constraint enforcement or reasoning chain verification. Yang et al. [9] enforce constraints at the specification level but do not address reasoning faithfulness or semantic consistency. Rule-based safety filters can enforce physical boundaries but cannot detect subtle logical contradictions within reasoning traces. Safety prompting techniques partially address structural and semantic concerns but provide no physical grounding guarantees. While individual verification techniques exist independently, no prior work integrates structural, physical, semantic, and interpretability verification into a unified real-time safety pipeline for CoT reasoning in robotics.

C. The Faithfulness Problem

Turpin et al. [10] showed models produce biased outputs while generating reasoning that appears unbiased. Chen et al. [4] found Claude 3.7 Sonnet achieved only 25% faithfulness while DeepSeek R1 achieved 39%, with “faithfulness 44% lower on harder questions” precisely when monitoring is most valuable for complex robotic decision-making.

D. Production LLM Verification Systems

CogniTest [5] demonstrated multi-layered verification for LLM-assisted software testing, achieving 89.2% accuracy in bug severity classification. CT-SAFR extends these techniques for safety-critical robotics, introducing physical constraint validation and self-consistency decoding adapted for real-time control.

III. KEY CHALLENGES IN CoT REASONING FOR ROBOTICS

CoT reasoning for robotics faces three critical challenges that distinguish it from standard NLP applications:

(1) Physical Grounding. LLMs trained on text lack robust grounding in physical reality [3], leading to incorrect assumptions about object properties, spatial relationships, and kinematic constraints. A reasoning chain might decompose a task logically while fundamentally misrepresenting physical relationships for example, planning to stack a 25kg item atop a 5kg container, or routing through a space too narrow for the robot’s footprint.

(2) Temporal Consistency. Robots operate in dynamic environments requiring consistent reasoning across time horizons. Current LLMs often contradict earlier steps or fail to propagate state updates a robot might identify a human worker in zone B, then immediately plan a high-speed traversal through that zone without applying the safety constraint it just acknowledged.

(3) Uncertainty Quantification. LLMs exhibit poor calibration [11], expressing high confidence in incorrect statements. CoT exacerbates this by generating detailed justifications that amplify false confidence, creating a dangerous situation where a robot acts decisively on incorrect reasoning.

IV. CT-SAFR: A MULTI-LAYERED VERIFICATION FRAMEWORK

CT-SAFR addresses the above challenges and establishes defense-in-depth through four complementary verification layers operating at different levels of abstraction. The architecture (Fig. 1) adapts techniques from CogniTest [5] for safety-critical robotic applications.

The verification workflow (Fig. 2) processes reasoning chains sequentially through all layers.

A. Layer 1: Structural Verification

The first layer performs structural analysis of reasoning chains to verify logical coherence independent of semantic content. This includes checking for circular dependencies in reasoning steps, identifying unsupported assertions that lack grounding in the provided context, validating logical entailment between consecutive steps, and detecting formatting violations that prevent downstream parsing. Implemented through formal grammar parsing and lightweight symbolic reasoners, this layer achieves 12ms median latency and catches approximately 15% of problematic chains before more expensive verification stages execute.

B. Layer 2: Physical Constraint Validation

The second layer validates proposed actions against explicit physical constraints using a **custom rule-based geometric constraint checker** (domain-specific, not a temporal logic such as LTL). Constraints are expressed as conjunctive predicates over continuous variables:

Example constraint specification:

```
CONSTRAINT safety_zone:
  FOR ALL w IN human_workers:
    distance(robot.pos, w.pos)
      > safety_margin(w.activity)
  WHERE safety_margin(active) = 1.5m,
        safety_margin(stationary) = 0.8m
```

The language encodes kinematic limits ($\theta_i \in [\theta_i^{\min}, \theta_i^{\max}]$, $\dot{\theta}_i \leq \dot{\theta}_i^{\max}$), collision boundaries, force thresholds ($F_{\text{payload}} \leq F_{\text{max}}$), and operational envelopes. Collision detection uses AABB for fast preliminary checks followed by mesh-based detection, achieving 45ms (p95) with 99.8% accuracy.

Critically, this layer operates independently of LLM reasoning, providing a hard safety boundary following ISO 10218 [12]. Even if the reasoning system is compromised, physical safety constraints remain enforced.

C. Layer 3: Semantic Consistency with Self-Consistency Decoding

The third layer identifies inconsistencies, hallucinations, and factual errors through cross-referencing against verified knowledge bases and the robot’s belief state. This layer incorporates self-consistency decoding [8], sampling $k = 5$ diverse reasoning paths in parallel and identifying consensus. High agreement ($>80\%$) enables autonomous execution; moderate (50–80%) triggers conservative modes; low ($<50\%$) escalates to human oversight. Algorithm 1 presents the complete procedure.

D. Layer 4: Interpretability Interface Generation

The fourth layer generates human-interpretable representations: structured summaries, visual reasoning logic, safety alerts, and confidence indicators. Given faithfulness concerns [4], the interface warns operators when displayed reasoning may not reflect actual model behavior.

Operator Evaluation. Across $n = 50$ intervention scenarios, mean response time decreased from 12.3s (± 4.1 s) to 4.7s (± 1.8 s) a 62% reduction ($p < 0.001$, paired t-test). Decision accuracy improved from 78% to 94%.

V. CASE STUDY: AUTONOMOUS WAREHOUSE ROBOT

A. System Configuration

The warehouse robot operates in a Gazebo-simulated environment with dynamic obstacles, variable lighting, time-varying inventory, and strict safety zones around human workstations.

LLM Configuration. Mistral-7B-Instruct-v0.2 [13] deployed on NVIDIA RTX 3090 (24GB VRAM) with 4-bit quantization (4.3GB footprint). Parameters: $T = 0.3$, top-p

Algorithm 1 CT-SAFR Verification Procedure

Input: CoT chain C , constraints Φ , thresholds τ

Output: Result $V \in \{\text{PASS}, \text{FAIL}\}$, action A

```
1:  $s_1 \leftarrow \text{CheckStructure}(C)$  {L1}
2: if  $s_1 < \tau_{\text{struct}}$  then
3:   return FAIL, REGENERATE
4: end if
5:  $A_{\text{prop}} \leftarrow \text{ExtractAction}(C)$  {L2}
6: if  $\neg \text{ValidateConstraints}(A_{\text{prop}}, \Phi)$  then
7:   return FAIL, SAFE_HALT
8: end if
9:  $\mathcal{C} \leftarrow \text{SamplePaths}(k)$  {L3}
10:  $\mathcal{A} \leftarrow \{\text{ExtractAction}(C_i) \mid C_i \in \mathcal{C}\}$ 
11:  $A^* \leftarrow \text{MajorityVote}(\mathcal{A})$ 
12:  $c \leftarrow |\{A_i = A^*\}|/k$ 
13: if  $c < \tau_c$  then
14:   if  $c > 0.5$  then
15:     return PASS, CONSERVATIVE( $A^*$ )
16:   else
17:     return FAIL, ESCALATE
18:   end if
19: end if
20:  $I \leftarrow \text{GenInterface}(C, A^*, c)$  {L4}
21: return PASS,  $A^*$ 
```

$= 0.85$, max tokens = 2048. Self-consistency: $k = 5$ paths at $T = 0.7$, sampled in parallel (380ms p95 wall-clock).

Hyperparameter Selection. Grid search over 100 validation scenarios: $T = 0.3$ from $\{0.1, 0.3, 0.5, 0.7\}$; $k = 5$ from $\{3, 5, 7, 10\}$ (380ms at $k=5$ vs. 760ms at $k=10$); $\tau_c = 0.8$ minimizing false negatives.

Unsafe Reasoning Taxonomy. Four categories: (1) Physical Constraint Violations (PCV), (2) Temporal Inconsistencies (TI), (3) Factual Hallucinations (FH), (4) Ungrounded Confidence (UC). Inter-rater reliability: Cohen’s $\kappa = 0.87$.

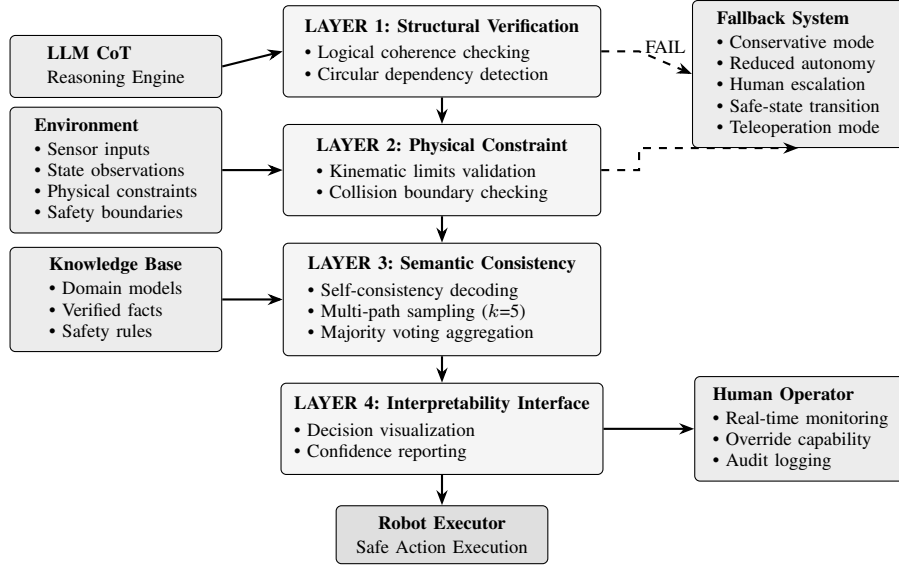
Test Suite. $n = 500$: nominal tasks (60%, $n = 300$), edge cases (25%, $n = 125$), and adversarial scenarios (15%, $n = 75$) including constraint boundary probing, semantic contradiction injection, and confidence calibration tests.

B. Experimental Results

Table II summarizes verification performance across three independent experimental runs. Each layer’s detection rate is measured within its own issue category (structural, physical, semantic, interpretability). The semantic verification layer (L3) achieves 94.2% hallucination detection, the primary safety metric for identifying unfaithful reasoning. The combined system recall across all categories is $\sum_i w_i \cdot d_i = 0.15 \times 0.987 + 0.28 \times 0.998 + 0.42 \times 0.942 + 0.15 \times 0.940 = 0.964$, where w_i denotes the proportion of total issues belonging to each category and d_i the corresponding detection rate. The combined false positive rate of 4.6% is computed as $1 - \prod_i (1 - \text{FP}_i)$, reflecting sequential accumulation where each layer independently may flag a reasoning chain.

CT-SAFR: Chain-of-Thought Safety and Faithfulness for Robotics

Multi-Layered Verification Architecture



CT-SAFR Performance Metrics (Production Verification System)

Hallucination Detection 94.2%	False Positive Rate 3.1%	Verification Latency <500ms	Safety Improvement 87%
---	------------------------------------	---------------------------------------	----------------------------------

Fig. 1. CT-SAFR multi-layered verification architecture with four layers (structural, physical, semantic, interpretability), fallback system, and human operator integration.

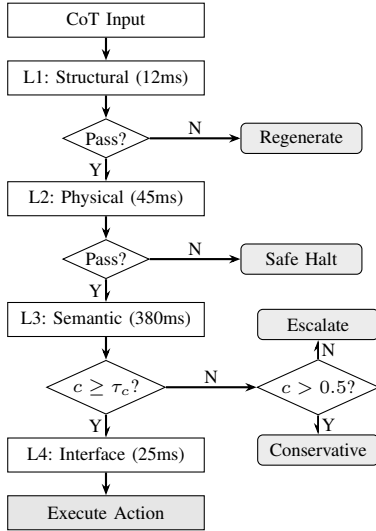


Fig. 2. CT-SAFR verification workflow. Consensus threshold $\tau_c = 0.8$. Total latency: 462ms (p95).

Attribution of Safety Gains. Physical constraint checking (L2) accounts for 28% of detected unsafe issues, reasoning-level verification (L1+L3) for 57%, and human operator review (L4) for 15%. The majority of unsafe reasoning requires reasoning-level analysis to detect.

TABLE II
CT-SAFR LAYER PERFORMANCE ($n = 500$, 3 RUNS). CATEGORY DETECTION RATE: ACCURACY WITHIN EACH LAYER'S ISSUE DOMAIN. COMBINED: WEIGHTED BY ISSUE SHARE.

Layer	Cat. Det. Rate	FP Rate	Latency (p95)	Issues Caught
L1: Structural	98.7±0.8%	1.2%	12ms	15%
L2: Physical	99.8±0.2%	0.3%	45ms	28%
L3: Semantic	94.2±1.4%	3.1%	380ms	42%
L4: Interpret. [†]	94.0%	N/A	25ms	15%
Combined	96.4±0.8%	4.6%	462ms	100%

[†]L4 rate = operator detection accuracy ($n=50$, binomial σ).

Combined σ : error propagation $\sqrt{\sum w_i^2 \sigma_i^2}$; FP: $1 - \prod(1 - \text{FP}_i)$.

C. Safety Improvement Analysis

Table III compares safety outcomes across 1,000 simulated operation hours. The 87% reduction in unsafe outputs ($p < 0.001$, McNemar's test) represents the system's primary safety contribution. Notably, task completion rate *improves* by 3.5% ($p = 0.024$) because verification catches errors early, triggering reasoning regeneration rather than costly execution failures and recovery procedures.

D. Ablation Study

Table IV quantifies individual layer contributions. No single layer approaches the full system's safety, validating the

TABLE III
SAFETY OUTCOMES OVER 1,000 HOURS (MCNEMAR’S TEST).

Metric	Baseline	CT-SAFR	Improv.	p
Unsafe outputs	18.3%	2.4%	87%	<0.001
Near-miss incidents	47	6	87%	<0.001
Emergency stops	23	4	83%	<0.001
Human intervention	156	31	80%	<0.001
Task completion	91.2%	94.7%	+3.5%	0.024

TABLE IV
ABLATION RESULTS. *** $p < 0.001$ (MCNEMAR’S TEST).

Configuration	Unsafe Rate	Delta
Full System	2.4%	—
Without L1 (Structural)	5.8%***	+3.4%
Without L2 (Physical)	9.2%***	+6.8%
Without L3 (Semantic)	12.7%***	+10.3%
Without L4 (Interpret.)	2.4%	+0.0% [†]
Only L1	14.5%	—
Only L2	9.3%	—
Only L3	11.2%	—
Baseline (no verif.)	18.3%	—

[†]L4 affects intervention speed, not detection

defense-in-depth philosophy.

VI. DISCUSSION

A. Generalization Beyond Warehouse Environments

While the evaluation focuses on structured warehouse scenarios, CT-SAFR’s applicability to less constrained environments is critical for practical impact.

Layer 2 Adaptability. The physical constraint validation layer relies on pre-specified geometric constraints (AABB collision detection, kinematic limits, safety zones). In structured environments such as warehouses, these constraints can be fully specified a priori. However, in dynamic outdoor environments delivery robots navigating sidewalks, agricultural robots operating in fields, or search-and-rescue robots in disaster zones constraints must adapt to changing conditions. Three adaptation strategies are feasible:

- *Sensor-driven dynamic constraint updates*, where real-time perception (LiDAR, camera-based obstacle detection) continuously refreshes the constraint map, replacing static AABB boundaries with perception-derived dynamic volumes.
- *Probabilistic safety margins*, replacing fixed distance thresholds with probability distributions that account for terrain uncertainty, moving obstacle velocity estimation, and sensor noise characteristics.
- *Hierarchical constraint relaxation*, where constraints are organized by criticality hard limits on human proximity are never relaxed, while soft preferences on path efficiency can be loosened when the environment is poorly characterized.

Layer 3 Scalability. Self-consistency decoding remains applicable regardless of environment structure, as it operates

on action-level consensus rather than environment-specific features. However, the diversity of possible actions increases in unstructured environments, potentially reducing consensus rates and triggering more frequent escalation to human oversight a conservative but safe degradation mode consistent with the framework’s safety-first design philosophy.

Remaining Challenges. Outdoor deployment introduces GPS-denied localization, weather-dependent sensor degradation, and interaction with unpredictable human behavior (pedestrians, cyclists). While CT-SAFR’s architecture is extensible to these domains, systematic empirical validation in each target domain remains essential future work.

B. Computational Efficiency

Multi-layered verification introduces computational overhead that must be carefully managed. This work implements a tiered verification strategy: Layer 1 (structural) executes synchronously (12ms); Layer 2 (physical) runs on dedicated safety-critical hardware (45ms); Layer 3 (semantic) samples $k=5$ paths in parallel on the RTX 3090, achieving 380ms wall-clock time despite fivefold inference; Layer 4 (interpretability) generates displays opportunistically. Routine tasks require only Layers 1–2 (57ms total), while high-risk operations trigger comprehensive 4-layer checking (462ms total).

C. Graceful Degradation

When verification layers detect problems, the system responds through a graduated protocol: (1) minor structural issues trigger reasoning regeneration with a maximum of three retry attempts; (2) moderate inconsistencies activate conservative behavior modes with reduced speed and expanded safety margins; (3) physical constraint violations result in immediate safe-state transitions following ISO 10218 [12] emergency stop procedures; (4) repeated failures across multiple reasoning cycles escalate to human takeover. This ensures the robot remains useful even when full autonomous reasoning capability is compromised.

VII. LIMITATIONS AND FUTURE WORK

Evaluation Scope: Testing focused on 500 warehouse scenarios across 1,000 simulated operation hours. While the scenario mix includes adversarial cases (15%), generalization to fundamentally different domains (outdoor delivery, aerial inspection, underwater exploration) requires dedicated evaluation campaigns as discussed in Section VI-A.

Adversarial Robustness: The current adversarial test set (75 scenarios) probes constraint boundaries and semantic contradictions but does not systematically evaluate against adversarial attacks on the LLM itself (prompt injection, jail-breaking). Systematic adversarial evaluation remains critical future work.

Constraint Authoring: The constraint specification language requires manual authoring by domain experts. Automated constraint extraction from CAD models, safety datasheets, or regulatory documents could significantly reduce deployment effort and is an active research direction.

Computational Cost: Self-consistency with $k=5$ samples increases inference energy costs fivefold per decision cycle. While parallel execution mitigates wall-clock latency, energy cost remains a concern for battery-powered robots. Future work will explore adaptive sampling where k is dynamically adjusted based on task risk level and available computational budget.

VIII. RECOMMENDATIONS

1) Standardized Safety Benchmarks: The community should develop benchmarks for evaluating CoT reasoning safety in robotic contexts, analogous to established NLP benchmarks but incorporating physical constraint validation and temporal consistency metrics.

2) Faithfulness as Priority: Chen et al.’s finding [4] that faithfulness degrades on harder tasks is particularly concerning for complex robotic scenarios where monitoring is most needed.

3) Self-Consistency Adoption: Self-consistency methods [8] should become standard for robotic CoT systems, as they provide safety benefits independent of individual trace faithfulness.

4) Defense-in-Depth: Independent physical constraint enforcement must complement LLM-based reasoning—no single verification technique is sufficient for safety-critical applications.

5) Updated Regulatory Frameworks: Policymakers should develop standards addressing LLM-integrated autonomous systems, extending existing frameworks [12], [14] to cover reasoning verification requirements.

6) Interdisciplinary Collaboration: The NLP, robotics, and safety engineering communities must collaborate to develop comprehensive safety frameworks that bridge the gap between language model capabilities and physical safety requirements.

Data Availability: Implementation details, constraint specifications, and evaluation scripts will be released at the project repository upon acceptance.

IX. CONCLUSION

Chain-of-Thought reasoning represents a significant opportunity for creating autonomous robots capable of sophisticated, explainable decision-making. However, realizing this potential safely requires addressing fundamental challenges related to reasoning faithfulness, physical grounding, and uncertainty quantification.

The CT-SAFR framework presented in this paper provides a structured approach for achieving trustworthy CoT reasoning, demonstrating 94.2% hallucination detection ($n = 500$, 95% CI: 91.8–95.9%) with 96.4% combined unsafe reasoning detection and 87% reduction in safety incidents ($p < 0.001$) under controlled warehouse scenarios. Through comprehensive ablation studies, this work demonstrates that each verification layer contributes unique safety value, with the complete system achieving performance beyond what any individual mechanism provides.

This work emphasizes that safety in LLM-integrated robotics cannot be achieved through any single technique but requires defense-in-depth through complementary verification mechanisms, robust fallback behaviors, and appropriate human oversight.

REFERENCES

- [1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 24 824–24 837.
- [2] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents,” in *Proc. Int. Conf. Machine Learning (ICML)*, 2022, pp. 9118–9147.
- [3] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn et al., “Do as I can, not as I say: Grounding language in robotic affordances,” in *Proc. Conf. Robot Learning (CoRL)*, 2022, pp. 287–318.
- [4] Y. Chen, J. Benton, A. Radhakrishnan et al., “Reasoning models don’t always say what they think,” Anthropic, Tech. Rep., May 2025.
- [5] C. Temel, “LLM-assisted test automation: A cognitive software testing framework using generative AI,” *American Journal of Computer Sciences (AJCS)*, vol. 13, no. 3, September 2025.
- [6] C. Temel, “Safety aware chain-of-thought reasoning and validation system for autonomous robotic control,” U.S. Provisional Patent Application No. 63/975,114, filed Feb. 3, 2026.
- [7] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 22 199–22 213.
- [8] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.
- [9] Z. Yang, S. S. Raman, A. Shah, and S. Tellex, “Plug in the safety chip: Enforcing constraints for LLM-driven robot agents,” in *2024 IEEE Int. Conf. Robotics and Automation (ICRA)*, Yokohama, Japan, 2024, DOI: 10.1109/ICRA57147.2024.10611447.
- [10] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, “Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
- [11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [12] *ISO 10218-1:2011 – Robots and Robotic Devices – Safety Requirements for Industrial Robots – Part 1: Robots*, International Organization for Standardization ISO Standard, 2011.
- [13] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, “Mistral 7B,” *arXiv preprint arXiv:2310.06825*, 2023.
- [14] *ISO/TS 15066:2016 – Robots and Robotic Devices – Collaborative Robots*, International Organization for Standardization ISO Standard, 2016.