Novelty Assessment Report

Paper: Agentic Confidence Calibration
PDF URL: https://openreview.net/pdf?id=6YMFsGFabM
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

AI agents are rapidly advancing from passive language models to autonomous systems executing complex, multi-step tasks. Yet their overconfidence in failure remains a fundamental barrier to deployment in high-stakes settings. Existing calibration methods, built for static single-turn outputs, cannot address the unique challenges of agentic systems, such as compounding errors along trajectories, uncertainty from external tools, and opaque failure modes. To address these challenges, we introduce, for the first time, the problem of \emph{Agentic Confidence Calibration} and propose \textbf{Holistic Trajectory Calibration (\htcnospace)}, a novel diagnostic framework that extracts rich process-level features ranging from macro dynamics to micro stability across an agent's entire trajectory. Powered by a simple, interpretable model, \htc consistently surpasses strong baselines in both calibration and discrimination, across eight benchmarks, multiple LLMs, and diverse agent frameworks. Beyond performance, \htc delivers three essential advances: it provides \emph{interpretability} by revealing the signals behind failure, enables \emph{transferability} by applying across domains without retraining, and achieves \emph{generalization} through a \emph{General Agent Calibrator} (\gacnospace) that {achieves the best calibration (lowest ECE)} on the out-of-domain GAIA benchmark. Together, these contributions establish a new process-centric paradigm for confidence calibration, {\color{blue}providing a framework} for diagnosing and enhancing the reliability of AI agents.

Core Task Landscape

This paper addresses: **Confidence Calibration for Autonomous AI Agents**

A total of **50 papers** were analyzed and organized into a taxonomy with **32 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multi-Agent Deliberation and Collective Calibration**
- **Human-AI Collaborative Calibration**
- **Agentic and Trajectory-Level Calibration**
- **Domain-Specific Calibration Applications**
- **Neural Network Calibration Foundations**
- **Uncertainty-Aware Learning and Self-Improvement**
- **Reliability and Robustness in Agent Systems**
- **Policy, Governance, and Ethical Frameworks**
- **Cross-Disciplinary Perspectives and Reviews**

Complete Taxonomy Tree

- Confidence Calibration for Autonomous AI Agents Survey Taxonomy
- Multi-Agent Deliberation and Collective Calibration
  - LLM-Based Multi-Agent Debate for Calibration (2 papers)
  - [1] Confidence Calibration and Rationalization for LLMs via Multi-Agent Deliberation (Yang Rui-xin, 2024) View paper
  - [8] ConfidenceCal: Enhancing LLMs Reliability through Confidence Calibration in Multi-Agent Debate (Yilin Bai, 2024) View paper
  - Vision-Language Multi-Agent Calibration (2 papers)
  - [21] Refine and Align: Confidence Calibration through Multi-Agent Interaction in VQA (Ayush Pandey, 2025) View paper
  - [27] AlignVQA: Debate-Driven Multi-Agent Calibration for Vision Language Models (A Pandey, 2026) View paper
  - Model-Agnostic Multi-Agent Perception (1 papers)
  - [23] Model-Agnostic Multi-Agent Perception Framework (Runsheng Xu, 2023) View paper
- Human-AI Collaborative Calibration
  - Trust Calibration in AI-Assisted Decision Making (2 papers)
  - [2] Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making (Zhang Yunfeng, 2020) View paper
  - [12] Calibration of Trust in Autonomous Vehicle (Seul Chan Lee, 2022) View paper
  - Confidence-Guided Human-AI Collaboration (3 papers)
  - [9] Towards intelligent decision support systems in robotics: Investigating the role of self-confidence calibration in joint decision-making (R Bhattacharyya, 2024) View paper
  - [30] Confidence-Guided Human-AI Collaboration: Reinforcement Learning with Distributional Proxy Value Propagation for Autonomous Driving (Zeqiao Li, 2025) View paper
  - [39] Enhancing Joint Human-AI Inference in Robot Missions: A Confidence-Based Approach (Nguyen Duc An, 2025) View paper
  - Trust Dynamics and Adjustment in Human-AI Interaction (2 papers)
  - [43] Humans Adjust Trust Based on Experience, AI Prefers Humans Regardless (Xiaosong He, 2025) View paper
  - [45] Human-AI teaming co-learning in military operations (Clara, 2025) View paper

- Agentic and Trajectory-Level Calibration
  - Holistic Trajectory Calibration Frameworks ★ (1 papers)
  - [0] Agentic Confidence Calibration (Anon et al., 2026) View paper
  - Embodied Agent Confidence Elicitation (1 papers)
  - [3] Uncertainty in Action: Confidence Elicitation in Embodied Agents (Yu Tianjiao, 2025) View paper
  - Bayesian Meta-Learning for Self-Improving Agents (1 papers)
  - [18] Self-Improving Agentic AI through Bayesian Meta-Learning (Kaushik Bar, 2025) View paper
  - Metacognitive Self-Confidence Frameworks (2 papers)
  - [10] â□□A Good Bot Always Knows Its Limitationsâ□□: Assessing Autonomous System Decision-Making Competencies through Factorized Machine Self-Confidence (Brett Israelsen, 2025) View paper
  - [31] Towards self-confidence in autonomous systems (Nicholas Sweet, 2016) View paper
- Domain-Specific Calibration Applications
  - Autonomous Driving Calibration (2 papers)
  - [22] Uncertainty quantification and calibration of imitation learning policy in autonomous driving (MÃ¼ller Christian, 2020) View paper
  - [47] Optical aberrations in autonomous driving: Physics-informed parameterized temperature scaling for neural network uncertainty calibration (Wolf, 2024) View paper
  - Sensor Network Calibration for Robotics (2 papers)
  - [6] Self-Calibration of a Network of Radar Sensors for Autonomous Robots (Timo Grebner, 2023) View paper
  - [35] Joint Calibration of a Multimodal Sensor System for Autonomous Vehicles. (Jon MuhoviÄ□, 2023) View paper
  - Simulation and Scenario-Based Calibration (2 papers)
  - [4] Advanced Scenario Generation for Calibration and Verification of Autonomous Vehicles (Xuan Li, 2023) View paper
  - [26] Efficient calibration of multi-agent simulation models from output series with bayesian optimization (Bai, 2022) View paper
  - Energy and Smart Grid Forecasting Calibration (1 papers)
  - [36] Estimating Energy Forecasting Uncertainty for Reliable AI Autonomous Smart Grid Design (Maher Selim, 2021) View paper
- Neural Network Calibration Foundations
  - Calibration Metrics and Theoretical Frameworks (2 papers)
  - [5] Quantifying calibration error in modern neural networks through evidence based theory (Krontiris Ioannis, 2024) View paper
  - [19] Calibrating machine behavior: a challenge for AI alignment (Erez Firt, 2023) View paper
  - Preference Distribution Calibration (1 papers)
  - [24] Judging with Confidence: Calibrating Autoraters to Preference Distributions (Li, 2025) View paper
  - Impossibility Results and Fundamental Limits (1 papers)
  - [37] Disproving the Feasibility of Learned Confidence Calibration Under Binary Supervision: An Information-Theoretic Impossibility (AS Nair, 2025) View paper
- Uncertainty-Aware Learning and Self-Improvement
  - Uncertainty-Driven Self-Improvement (1 papers)
  - [25] Uncertainty-Aware Self-Improving Framework for Depth Estimation (X. Nie, 2021) View paper
  - Efficient Uncertainty Estimation via Distillation (1 papers)
  - [32] Self-Distribution Distillation: Efficient Uncertainty Estimation (Fathullah, 2022) View paper
  - Uncertainty-Aware Autonomous Exploration (1 papers)
  - [28] SPACESHIP: synthesizable parameter acquisition via closed-loop exploration and self-directed, hardware-aware intelligent protocols for autonomous lab (Nayeon Kim, 2025) View paper
  - Self-Directed Learning Under Uncertainty (1 papers)
  - [50] Selfâ□□directed learning favors local, rather than global, uncertainty (Douglas Markant, 2016) View paper
- Reliability and Robustness in Agent Systems
  - Byzantine Fault Tolerance in Multi-Agent Systems (1 papers)
  - [46] Rethinking the Reliability of Multi-agent System: A Perspective from Byzantine Fault Tolerance (Lifan Zheng, 2025) View paper
  - Hallucination Reduction in Multilingual Agents (2 papers)
  - [13] Detecting and classifying llm hallucinations: A framework for skill-specific error analysis (T Duenas, 2024) View paper
  - [48] Reducing Hallucination in Multilingual Voice Agents Using Instruction-Tuned Models (Mahmoud Abdelhadi Mahmoud Safia, 2024) View paper
  - Modular Trust-Aware Agent Architectures (1 papers)
  - [44] Agentic AI with Orchestrator-Agent Trust: A Modular Visual Classification Framework with Trust-Aware Orchestration and RAG-Based Reasoning (Sapkota, 2025) View paper
- Policy, Governance, and Ethical Frameworks
  - AI Governance and Risk Management Frameworks (1 papers)
  - [11] Framework for Government Policy on Agentic and Generative AI: Governance, Regulation, and Risk Management (Joshi, 2025) View paper
  - Ethical AI and Dynamic Equilibrium Theory (1 papers)
  - [42] Dynamic Equilibrium Theory for Ethical AI: Balancing Epistemic Uncertainty, Human Autonomy, and Social Equity in High-Stakes Fluctuational Decision â□¦ (Tan, 2025) View paper
  - AI-Enabled Policy Scenario Planning (1 papers)
  - [40] AI4Policy: AI-Enabled Scenario Planning for Policy-Making in the Age of AI (J Kgomo, 2025) View paper
- Cross-Disciplinary Perspectives and Reviews
  - Neuroscience-AI Interface and Uncertainty (1 papers)
  - [20] Recent advances at the interface of neuroscience and artificial neural networks (Y Cohen, 2022) View paper
  - Metacognition in Humans and LLMs (1 papers)
  - [38] Metacognition and Uncertainty Communication in Humans and Large Language Models (Steyvers, 2025) View paper
  - AI in Education and Self-Directed Learning (4 papers)
  - [16] Artificial Intelligence and ChatGPT in Medical Education: A Cross-Sectional Questionnaire on studentsâ□□ Competence (L. MaaÃ□, 2024) View paper
  - [17] Artificial Intelligence and Healthcare Simulation: The Shifting Landscape of Medical Education (Allan Hamilton, 2024) View paper

- [29] Self-directed learning during the COVID-19 pandemic: Perspectives of South African final-year health professions students (S, 2022) View paper
- [41] Neural network technologies in the self-development of adultsâ life skills (Olena Pehota, 2025) View paper
- Broad AI Surveys and Emerging Ecosystems (4 papers)
- [7] Introduction to the Research Handbook on Public Management and Artificial Intelligence (Yannis Charalabidis, 2024) View paper
- [14] Emerging materials intelligence ecosystems propelled by machine learning (Rohit Batra, 2021) View paper
- [33] Smart Robotic Assistants for Manufacturing Applications through Advances in Artificial Intelligence (Gupta, 2019) View paper
- [34] Self-Learning Algorithms: History, Advancements, Applications, Challenges, and Future Directions (Virk, 2022) View paper
- Domain-Specific AI Applications Without Calibration Focus (2 papers)
- [15] Bridging Coaching Knowledge and AI Feedback to Enhance Motor Learning in Basketball Shooting Mechanics Through a Knowledge-Based SOP Framework (Jian-Jia Weng, 2025) View paper
- [49] Self-Directed Machine Learning Based Prediction OfMultifaceted Gaze-Based Interactions For Extended Reality In Immersive Environments (MA Akey Sungheetha, 2024) View paper

Narrative

Core task: confidence calibration for autonomous AI agents. The field addresses how AI systems can accurately assess and communicate their own uncertainty when making decisions or predictions. The taxonomy reveals a rich landscape spanning nine major branches. Neural Network Calibration Foundations[47] and Uncertainty-Aware Learning and Self-Improvement[22] provide the technical underpinnings, focusing on methods that ensure model outputs reflect true probabilities and enable systems to learn from their own uncertainty. Multi-Agent Deliberation and Collective Calibration[1] and Human-AI Collaborative Calibration[2] explore how calibration emerges through interaction—either among multiple agents or between humans and machines. Agentic and Trajectory-Level Calibration examines calibration across entire decision sequences rather than isolated predictions, while Domain-Specific Calibration Applications[6] tailors these techniques to specialized contexts like autonomous vehicles[12], healthcare[16], and robotics[9]. Reliability and Robustness in Agent Systems[46] and Policy, Governance, and Ethical Frameworks[7] address the broader implications of deploying calibrated agents in safety-critical and socially sensitive settings.

Several active lines of work reveal key tensions and open questions. One strand emphasizes holistic, trajectory-aware approaches that calibrate confidence over multi-step agent behaviors, contrasting with traditional per-prediction calibration methods. Another explores the interplay between self-assessment and external validation, as seen in works on metacognition[38] and human trust dynamics[43]. Agentic Confidence Calibration[0] sits squarely within the Holistic Trajectory Calibration Frameworks cluster, emphasizing end-to-end confidence assessment across agent decision sequences. This positions it closely with efforts like Uncertainty Embodied Agents[3] and Bot Knows Limitations[10], which similarly focus on agents that recognize and communicate their epistemic boundaries throughout task execution. Compared to ConfidenceCal[8], which targets calibration at individual decision points, Agentic Confidence Calibration[0] adopts a more integrated view of how uncertainty propagates and compounds across an agent's operational trajectory, reflecting a shift toward calibration as an ongoing, context-sensitive process rather than a static property.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Sibling Subtopics

- **Bayesian Meta-Learning for Self-Improving Agents** (leaves: 1, papers: 1)
- Scope: Bayesian approaches enabling agents to self-improve through meta-learning with well-calibrated confidence estimates.
- Exclude: Non-Bayesian self-improvement and static calibration methods belong in other categories.
- **Embodied Agent Confidence Elicitation** (leaves: 1, papers: 1)
- Scope: Methods for eliciting and calibrating confidence in embodied agents navigating multimodal environments with perception and action uncertainty.
- Exclude: Disembodied language agents and static perception systems belong elsewhere.
- **Metacognitive Self-Confidence Frameworks** (leaves: 1, papers: 2)
- Scope: Factorized frameworks for autonomous systems to assess their own competency and knowledge through meta-reasoning.
- Exclude: External calibration methods and human-supervised confidence belong elsewhere.

Contributions Analysis

Overall novelty summary. The paper introduces Agentic Confidence Calibration as a novel problem formulation and proposes Holistic Trajectory Calibration (HTC) to address multi-step agent uncertainty. It resides in the 'Holistic Trajectory Calibration Frameworks' leaf, which currently contains only this work as its sole member. This positioning suggests the paper occupies a relatively sparse research direction within the broader taxonomy, distinguishing itself from single-turn calibration methods and multi-agent deliberation approaches that populate neighboring branches.

The taxonomy reveals that the paper sits at the intersection of several active research areas. Its closest neighbors include 'Embodied Agent Confidence Elicitation' and 'Metacognitive Self-Confidence Frameworks' within the same parent branch, both addressing agent-level uncertainty but through different mechanisms. The broader 'Agentic and Trajectory-Level Calibration' branch contains only four leaf nodes, indicating this process-centric perspective on calibration remains less explored than foundational neural network calibration methods or domain-specific applications, which collectively account for over half the taxonomy's papers.

Among thirty candidates examined through semantic search, the contribution-level analysis reveals mixed novelty signals. The problem formulation for Agentic Confidence Calibration shows one refutable candidate among ten examined, suggesting some conceptual overlap with prior work on agent uncertainty. In contrast, both the HTC framework and General Agent Calibrator (GAC) components encountered no clear refutations across their respective ten-candidate searches, indicating these technical contributions may offer more distinctive methodological advances within the limited scope examined.

Based on the top-thirty semantic matches analyzed, the work appears to introduce a relatively novel perspective on trajectory-level calibration, though the limited search scope and single refutable candidate for the problem formulation suggest caution. The sparse population of its taxonomy leaf and the absence of refutations for its core technical components hint at meaningful differentiation from existing approaches, but a more exhaustive literature review would be needed to confirm the full extent of its originality.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Agentic Confidence Calibration problem formulation

Description: The authors formally define the novel problem of calibrating confidence in agentic AI systems by diagnosing entire execution trajectories rather than only final outputs. This formulation addresses unique challenges such as compounding errors, multi-source uncertainty from tools and environments, and opaque failure modes across multi-step reasoning processes.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A Survey on Joint Embedding Predictive Architectures and World Models

URL: View paper

Brief Assessment

Joint Embedding Architectures[79] is a survey on world models and joint embedding predictive architectures, focusing on model-based RL and control with uncertainty-aware imagination. It does not address confidence calibration in multi-step agentic AI systems or trajectory-level diagnostic frameworks for LLM-based agents.

---

2. Deep Hidden Cognition Facilitates Reliable Chain-of-Thought Reasoning

URL: View paper

Brief Assessment

Deep Hidden Cognition[80] focuses on calibrating confidence in chain-of-thought reasoning steps within a single model's internal reasoning process, not on diagnosing multi-step agentic execution trajectories involving external tools and environments.

---

3. TUNER-compliant error estimation for MIPAS

URL: View paper

Brief Assessment

TUNER-Compliant Error[77] addresses error estimation for atmospheric sounding instrument retrievals, focusing on measurement noise, spectroscopic uncertainties, and calibration errors in a physical sensing context. This is fundamentally different from calibrating confidence in multi-step AI agent reasoning trajectories with compounding errors from tool use and LLM generation.

---

4. Don't Think Twice! Over-Reasoning Impairs Confidence Calibration

URL: View paper

Brief Assessment

Over-Reasoning Impairs Calibration[78] focuses on confidence calibration in question-answering tasks using expert-labeled datasets (climate science, health), not on diagnosing multi-step agentic execution trajectories with compounding errors from tool interactions and environments.

---

5. Uncertainty-aware decision transformer for stochastic driving environments

URL: View paper

Brief Assessment

Uncertainty-Aware Decision Transformer[75] addresses uncertainty in stochastic driving environments for offline RL, not confidence calibration in multi-step agentic AI systems with compounding errors across diverse tool-using trajectories.

---

6. Comprehensive Evaluation of AI Hallucination and Novel UV-Oriented Framework toward Safe and Trustworthy AI

URL: View paper

Brief Assessment

UV-Oriented Framework[74] focuses on hallucination detection and mitigation in LLMs through uncertainty calibration at the output level, not on calibrating confidence across multi-step agentic trajectories with tool interactions and compounding errors.

---

7. Uncertainty Quantification for Scientific Machine Learning using Sparse Variational Gaussian Process Kolmogorov-Arnold Networks (SVGP KAN)

URL: View paper

Brief Assessment

SVGP KAN[73] addresses uncertainty quantification in scientific machine learning using Gaussian processes and Kolmogorov-Arnold networks. It does not address confidence calibration in multi-step agentic AI systems with compounding errors from tool interactions and sequential reasoning trajectories.

---

8. Multivariate Bayesian predictive synthesis in macroeconomic forecasting

URL: View paper

Brief Assessment

Bayesian Predictive Synthesis[76] addresses multivariate macroeconomic forecasting through synthesis of multiple forecast densities, not confidence calibration in multi-step agentic AI systems with compounding errors and tool-based uncertainty.

---

9. UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making

URL: View paper

Prior Art Analysis

UProp[72] demonstrates that prior work exists on formulating uncertainty quantification problems for multi-step agentic systems. The candidate paper explicitly addresses the same core challenges identified in the original contribution: compounding errors across trajectories, multi-source uncertainty from tools and environments, and opaque failure modes in sequential decision-making. UProp[72] provides a formal information-theoretic framework that decomposes uncertainty in sequential decision-making into intrinsic and extrinsic components, directly addressing the trajectory-level diagnosis problem that the original paper claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce frameworks for diagnosing uncertainty across entire agent trajectories. UProp[72] explicitly decomposes uncertainty into components that propagate through sequential decisions, addressing the same trajectory-level diagnosis problem. - **Original**: we introduce, for the first time, the problem ofagentic confidence calibration and proposeholistic trajectory calibration (htc), a novel diagnostic framework that extracts rich process-level features ranging from macro dynamics to micro stability across an agent's entire trajectory. - **Candidate**: in this paper, we introduce a principled, information-theoretic framework that decomposes llm sequential decision uncertainty into two parts: (i) internal uncertainty intrinsic to the current decision, which is focused on existing uq methods, and (ii) extrinsic uncertainty, a mutual-information (mi)...

Evidence 2 - **Rationale**: Both papers identify compounding uncertainty across sequential steps as a fundamental challenge. UProp[72] provides a formal mathematical framework for how uncertainty propagates through decision trajectories, addressing the same problem of accumulated uncertainty. - **Original**: this shift from a static generator to a dynamic actor fundamentally alters the nature of the reliability challenge. first, uncertainty is no longer an isolated property of a single output, but a compounding factor that accumulates and propagates throughout a sequential trajectory - **Candidate**: in the llm multi-step decision-making process, uq quantifies the uncertainty within the predictive distribution pθ(y|x). without loss of generality, we quantify the uncertainty at the t-th step predictive distribution yt ∼ pθ(yt|x). by marginalizing preceding decisions, we obtain the following decom...

Evidence 3 - **Rationale**: Both papers address uncertainty from external environment interactions. UProp[72] explicitly models agent-environment interactions in its framework, showing prior work on multi-source uncertainty in agentic systems. - **Original**: agents introduce new, external sources of uncertainty through their interaction with tools and environments (gao et al., 2024; levy & yih, 2024). api failures, noisy data returned by tools, or the misuse of a tool's functionality create new reliability bottlenecks independent of the model's internal... - **Candidate**: llm multi-step agentic decision-making liu et al. (2023); duan et al. (2024b) is usually modeled as a stochastic markov decision process (mdp) (fθ, o, y, t), where llm fθ interacts with the environment continuously. o and y are observation space and decision space, respectively. t : y∗ → ois the de...

Evidence 4 - **Rationale**: The original paper claims to be the first systematic framework for trajectory-level calibration. UProp[72] explicitly provides such a framework with formal mathematical foundations, demonstrating prior work exists. - **Original**: there is a lack of a systematic framework for effectively calibrating the confidence of an agent's final output by diagnosing its entire execution trajectory. - **Candidate**: we provide an information-theoretic framework that decomposes the uncertainty of llm sequential decision into intrinsic and extrinsic uncertainty. we highlight the necessity of propagating extrinsic uncertainty along the llm decision chain for more accurate uncertainty quantification.

---

10. Self-evaluation guided beam search for reasoning

URL: View paper

Brief Assessment

Self-Evaluation Beam Search[71] focuses on calibrating multi-step reasoning chains through stepwise self-evaluation during decoding, not on diagnosing entire execution trajectories in agentic systems with tool interactions and environmental uncertainty.

Contribution 2: Holistic Trajectory Calibration (HTC) framework

Description: The authors introduce HTC, a feature-based calibration framework that transforms raw confidence traces into process-diagnostic features (cross-step dynamics, intra-step stability, positional indicators, structural attributes) and maps them through a simple interpretable model to produce calibrated confidence estimates. The framework is decoupled from specific agent architectures and provides interpretability, transferability, and generalization.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Calibrating uncertainties in human trajectory forecasting

URL: View paper

Brief Assessment

Human Trajectory Forecasting[55] focuses on calibrating spatial trajectory predictions for human movement forecasting, not on calibrating confidence estimates for AI agent execution trajectories across multi-step reasoning tasks.

---

2. Calibrating car-following models by using trajectory data: Methodological study

URL: View paper

Brief Assessment

Car-Following Calibration[56] focuses on calibrating car-following models in traffic simulation using vehicle trajectory data, not on confidence calibration for AI agents. The domains, objectives, and methodologies are fundamentally different.

---

3. CCTR: calibrating trajectory prediction for uncertainty-aware motion planning in autonomous driving

URL: View paper

Brief Assessment

CCTR[58] focuses on calibrating trajectory predictions in autonomous driving using vehicle motion data, while the original paper addresses confidence calibration for AI agents executing multi-step reasoning tasks. These are fundamentally different application domains with distinct technical challenges.

---

4. Multi-agent reachability calibration with conformal prediction

URL: View paper

Brief Assessment

Conformal Prediction Reachability[54] focuses on multi-agent trajectory forecasting in autonomous driving with conformal prediction for spatial reachable sets, not on general agentic AI systems' process-level confidence calibration across diverse reasoning tasks.

---

5. Temporal early exiting with confidence calibration for driver identification based on driving sensing data

URL: View paper

Brief Assessment

Driver Identification Exiting[59] addresses temporal confidence calibration for driver identification in vehicles using driving sensing data, not general AI agent trajectory calibration. The domains (automotive driver ID vs. multi-step AI agent reasoning) and applications are fundamentally different.

---

6. TAU: trajectory data augmentation with uncertainty for next POI recommendation

URL: View paper

Brief Assessment

TAU[57] focuses on trajectory data augmentation for next POI recommendation in location-based services, not on confidence calibration for AI agents. The trajectories refer to user mobility patterns (check-ins at locations), not agent execution traces with confidence estimation.

---

7. Interpretable self-aware neural networks for robust trajectory prediction

URL: View paper

Brief Assessment

Self-Aware Neural Networks[51] focuses on trajectory prediction for autonomous vehicles with epistemic uncertainty estimation using evidential deep learning over interpretable latent concepts (agent behavior, road structure, social context). This differs from HTC's process-diagnostic calibration framework for AI agent confidence across diverse tasks using cross-step dynamics and intra-step stability features.

---

8. Confidence-Based Fusion of AC-LSTM and Kalman Filter for Accurate Space Target Trajectory Prediction

URL: View paper

Brief Assessment

Space Target Trajectory[53] focuses on physical space object trajectory prediction using dual-model fusion (AC-LSTM + Kalman Filter) for aerospace applications, not on calibrating confidence estimates for AI agent reasoning trajectories across diverse cognitive tasks.

---

9. Dynamics of postdecisional processing of confidence.

URL: View paper

Brief Assessment

Postdecisional Confidence Dynamics[60] studies human confidence judgments in perceptual decision-making through postdecisional evidence accumulation, not AI agent trajectory calibration using process-level features from LLM execution traces.

---

10. Degradation Pattern Recognition and Features Extrapolation for Battery Capacity Trajectory Prediction

URL: View paper

Brief Assessment

Battery Capacity Prediction[52] focuses on battery degradation trajectory prediction using health indicators and LSTM networks for capacity forecasting. This is a completely different domain (battery health monitoring) with different technical objectives (capacity prediction vs. confidence calibration) and does not address agentic AI systems or confidence estimation.

Contribution 3: General Agent Calibrator (GAC)

Description: The authors develop GAC, a pretrained universal calibrator trained on diverse datasets that generalizes to unseen tasks without retraining. GAC achieves the best calibration performance on challenging out-of-domain benchmarks, demonstrating that pretraining captures a transferable uncertainty grammar that serves as a plug-and-play reliability layer for agentic systems.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. UMFC: Unsupervised multi-domain feature calibration for vision-language models

URL: View paper

Brief Assessment

UMFC[68] addresses vision-language model calibration for domain shifts in image classification, not agentic trajectory calibration. The candidate focuses on visual and text encoder biases in CLIP for multi-domain image tasks, while the original develops a calibrator for sequential agent trajectories with tool use and multi-step reasoning.

---

2. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks

URL: View paper

Brief Assessment

Image Foreign Language[63] focuses on vision and vision-language pretraining using masked data modeling on images and text, not on calibrating confidence for agentic systems or uncertainty quantification.

---

3. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

URL: View paper

Brief Assessment

ViLBERT[65] focuses on pretraining vision-and-language representations for multimodal tasks like VQA and image retrieval, not on calibrating confidence or uncertainty in agentic systems.

---

4. Frozen Pretrained Transformers as Universal Computation Engines

URL: View paper

Brief Assessment

Frozen Pretrained Transformers[70] focuses on transferring pretrained language models to non-language modalities (vision, numerical computation, proteins) by freezing self-attention layers. This is fundamentally different from GAC, which is a calibrator for uncertainty quantification in agentic systems trained on diverse agent trajectory datasets.

---

5. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models

URL: View paper

Brief Assessment

Diff-instruct[62] addresses knowledge transfer from pre-trained diffusion models to other generative models, not pretrained universal calibrators for agent confidence estimation across unseen tasks.

---

6. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks

URL: View paper

Brief Assessment

Uni-perceiver[64] focuses on pre-training a unified architecture for generic perception tasks (image/video classification, VQA, retrieval) across modalities, not on calibrating confidence or uncertainty in agentic systems. The candidate addresses zero-shot and few-shot task transfer via unified representations, whereas GAC addresses trajectory-level confidence calibration for agent reliability.

---

7. On calibration of pre-trained code models

URL: View paper

Brief Assessment

Code Models Calibration[69] focuses on calibrating pretrained code models for software engineering tasks (code understanding), not on general AI agents executing multi-step trajectories with tool use. The domains and problem formulations are fundamentally different.

---

8. On calibration and out-of-domain generalization

URL: View paper

Brief Assessment

Calibration Out-of-Domain[67] focuses on multi-domain calibration for static classifiers under distribution shift, not on pretrained universal calibrators for agentic systems with sequential trajectories and tool interactions.

---

9. Integrating Task-Specific and Universal Adapters for Pre-Trained Model-based Class-Incremental Learning

URL: View paper

Brief Assessment

Task-Specific Universal Adapters[66] focuses on class-incremental learning with adapter fusion for continual learning tasks, not on pretrained universal calibrators for agent confidence estimation across unseen tasks.

---

10. Thermometer: Towards universal calibration for large language models

URL: View paper

Brief Assessment

Thermometer[61] focuses on calibrating LLMs for question-answering tasks using a recognition network trained on multiple datasets, not on calibrating agentic systems with multi-step trajectories involving tool use and environmental interaction.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Agentic Confidence Calibration View paper
- [1] Confidence Calibration and Rationalization for LLMs via Multi-Agent Deliberation View paper
- [2] Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making View paper
- [3] Uncertainty in Action: Confidence Elicitation in Embodied Agents View paper
- [4] Advanced Scenario Generation for Calibration and Verification of Autonomous Vehicles View paper
- [5] Quantifying calibration error in modern neural networks through evidence based theory View paper
- [6] Self-Calibration of a Network of Radar Sensors for Autonomous Robots View paper
- [7] Introduction to the Research Handbook on Public Management and Artificial Intelligence View paper
- [8] ConfidenceCal: Enhancing LLMs Reliability through Confidence Calibration in Multi-Agent Debate View paper
- [9] Towards intelligent decision support systems in robotics: Investigating the role of self-confidence calibration in joint decision-making View paper
- [10] â□□A Good Bot Always Knows Its Limitationsâ□□: Assessing Autonomous System Decision-Making Competencies through Factorized Machine Self-Confidence View paper
- [11] Framework for Government Policy on Agentic and Generative AI: Governance, Regulation, and Risk Management View paper
- [12] Calibration of Trust in Autonomous Vehicle View paper
- [13] Detecting and classifying llm hallucinations: A framework for skill-specific error analysis View paper
- [14] Emerging materials intelligence ecosystems propelled by machine learning View paper
- [15] Bridging Coaching Knowledge and AI Feedback to Enhance Motor Learning in Basketball Shooting Mechanics Through a Knowledge-Based SOP Framework View paper
- [16] Artificial Intelligence and ChatGPT in Medical Education: A Cross-Sectional Questionnaire on studentsâ□□ Competence View paper
- [17] Artificial Intelligence and Healthcare Simulation: The Shifting Landscape of Medical Education View paper
- [18] Self-Improving Agentic AI through Bayesian Meta-Learning View paper
- [19] Calibrating machine behavior: a challenge for AI alignment View paper
- [20] Recent advances at the interface of neuroscience and artificial neural networks View paper
- [21] Refine and Align: Confidence Calibration through Multi-Agent Interaction in VQA View paper
- [22] Uncertainty quantification and calibration of imitation learning policy in autonomous driving View paper
- [23] Model-Agnostic Multi-Agent Perception Framework View paper
- [24] Judging with Confidence: Calibrating Autoraters to Preference Distributions View paper
- [25] Uncertainty-Aware Self-Improving Framework for Depth Estimation View paper
- [26] Efficient calibration of multi-agent simulation models from output series with bayesian optimization View paper
- [27] AlignVQA: Debate-Driven Multi-Agent Calibration for Vision Language Models View paper
- [28] SPACESHIP: synthesizable parameter acquisition via closed-loop exploration and self-directed, hardware-aware intelligent protocols for autonomous lab View paper
- [29] Self-directed learning during the COVID-19 pandemic: Perspectives of South African final-year health professions students View paper
- [30] Confidence-Guided Human-AI Collaboration: Reinforcement Learning with Distributional Proxy Value Propagation for Autonomous Driving View paper
- [31] Towards self-confidence in autonomous systems View paper
- [32] Self-Distribution Distillation: Efficient Uncertainty Estimation View paper
- [33] Smart Robotic Assistants for Manufacturing Applications through Advances in Artificial Intelligence View paper
- [34] Self-Learning Algorithms: History, Advancements, Applications, Challenges, and Future Directions View paper
- [35] Joint Calibration of a Multimodal Sensor System for Autonomous Vehicles. View paper
- [36] Estimating Energy Forecasting Uncertainty for Reliable AI Autonomous Smart Grid Design View paper
- [37] Disproving the Feasibility of Learned Confidence Calibration Under Binary Supervision: An Information-Theoretic Impossibility View paper
- [38] Metacognition and Uncertainty Communication in Humans and Large Language Models View paper
- [39] Enhancing Joint Human-AI Inference in Robot Missions: A Confidence-Based Approach View paper
- [40] AI4Policy: AI-Enabled Scenario Planning for Policy-Making in the Age of AI View paper
- [41] Neural network technologies in the self-development of adultsâ□□ life skills View paper

- [42] Dynamic Equilibrium Theory for Ethical AI: Balancing Epistemic Uncertainty, Human Autonomy, and Social Equity in High-Stakes Fluctuational Decision â View paper
- [43] Humans Adjust Trust Based on Experience, AI Prefers Humans Regardless View paper
- [44] Agentic AI with Orchestrator-Agent Trust: A Modular Visual Classification Framework with Trust-Aware Orchestration and RAG-Based Reasoning View paper
- [45] Human-AI teaming co-learning in military operations View paper
- [46] Rethinking the Reliability of Multi-agent System: A Perspective from Byzantine Fault Tolerance View paper
- [47] Optical aberrations in autonomous driving: Physics-informed parameterized temperature scaling for neural network uncertainty calibration View paper
- [48] Reducing Hallucination in Multilingual Voice Agents Using Instruction-Tuned Models View paper
- [49] Self-Directed Machine Learning Based Prediction OfMultifaceted Gaze-Based Interactions For Extended Reality In Immersive Environments View paper
- [50] Selfâdirected learning favors local, rather than global, uncertainty View paper
- [51] Interpretable self-aware neural networks for robust trajectory prediction View paper
- [52] Degradation Pattern Recognition and Features Extrapolation for Battery Capacity Trajectory Prediction View paper
- [53] Confidence-Based Fusion of AC-LSTM and Kalman Filter for Accurate Space Target Trajectory Prediction View paper
- [54] Multi-agent reachability calibration with conformal prediction View paper
- [55] Calibrating uncertainties in human trajectory forecasting View paper
- [56] Calibrating car-following models by using trajectory data: Methodological study View paper
- [57] TAU: trajectory data augmentation with uncertainty for next POI recommendation View paper
- [58] CCTR: calibrating trajectory prediction for uncertainty-aware motion planning in autonomous driving View paper
- [59] Temporal early exiting with confidence calibration for driver identification based on driving sensing data View paper
- [60] Dynamics of postdecisional processing of confidence. View paper
- [61] Thermometer: Towards universal calibration for large language models View paper
- [62] Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models View paper
- [63] Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks View paper
- [64] Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks View paper
- [65] ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks View paper
- [66] Integrating Task-Specific and Universal Adapters for Pre-Trained Model-based Class-Incremental Learning View paper
- [67] On calibration and out-of-domain generalization View paper
- [68] UMFC: Unsupervised multi-domain feature calibration for vision-language models View paper
- [69] On calibration of pre-trained code models View paper
- [70] Frozen Pretrained Transformers as Universal Computation Engines View paper
- [71] Self-evaluation guided beam search for reasoning View paper
- [72] UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making View paper
- [73] Uncertainty Quantification for Scientific Machine Learning using Sparse Variational Gaussian Process Kolmogorov-Arnold Networks (SVGP KAN) View paper
- [74] Comprehensive Evaluation of AI Hallucination and Novel UV-Oriented Framework toward Safe and Trustworthy AI View paper
- [75] Uncertainty-aware decision transformer for stochastic driving environments View paper
- [76] Multivariate Bayesian predictive synthesis in macroeconomic forecasting View paper
- [77] TUNER-compliant error estimation for MIPAS View paper
- [78] Don't Think Twice! Over-Reasoning Impairs Confidence Calibration View paper
- [79] A Survey on Joint Embedding Predictive Architectures and World Models View paper
- [80] Deep Hidden Cognition Facilitates Reliable Chain-of-Thought Reasoning View paper