SAGE: A Theory-Informed LLM-Based Multi-Agent Recommendation System for Grounded AI Solutions Across Domains

Jiawei Tong[1, 2], Guangyu Wang[3], Yan Wang[4], Ruoxi Liao[4], Shuihua Wang[1*], John Moraros[1†]

[1]Department of Biosciences and Bioinformatics
Suzhou Municipal Key Lab AI4Health, School of Science
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

[2]Institute of Systems, Molecular & Integrative Biology
University of Liverpool
Liverpool, United Kingdom

[3]Department of Computer Science, Data Science, and Engineering
New York University Shanghai
Pudong, Shanghai, China

[4]School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Abstract

Large Language Model-based multi-agent systems (LaMAS) show promise for solving complex problems, yet lack systematic mechanisms for responsible integration of established domain knowledge—a critical gap that risks producing technically sophisticated but contextually inappropriate solutions. To address this fundamental challenge, we present **SAGE** (**S**pecialized **A**gent **G**roup for **E**xpert-grounded recommendations), a novel multi-agent framework that addresses this challenge through five specialized agents coordinating via asynchronous protocol to integrate domain theories, AI methods, and data sources. Our architecture implements three key innovations: (1) *theoretical grounding mechanisms* that prioritize established domain wisdom over algorithmic capabilities, (2) *coordinated specialization* where agents handle distinct reasoning tasks—scenario analysis, theory retrieval, algorithm matching, data selection, and validation—rather than monolithic processing, and (3) *robustness validation* through Monte Carlo simulation ($N = 1000$ runs) with human-in-the-loop oversight ensuring $R > 0.85$ stability across perturbed scenarios. We validate SAGE's effectiveness in urban planning and safety, a complex domain requiring integration of social, environmental, and spatial theories. Empirical analysis of 1,123 AI applications reveals that only 1.16% demonstrate explicit theoretical integration, with 47.3% following technology-driven rather than problem-driven approaches. Through three representative case studies, we demonstrate how SAGE transforms narrow technical solutions into comprehensive, theory-grounded interventions that generate value beyond algorithmic performance metrics. Our framework establishes generalizable design principles for responsible agentic systems—agent specialization strategies, coordination protocols for knowledge integration, and validation mechanisms—with demonstrated potential for adaptation to other knowledge-intensive domains requiring systematic integration of theoretical frameworks with computational approaches.

Introduction

LLM-based multi-agent systems (LaMAS) demonstrate promising capabilities for complex problem-solving (Qian et al. 2023; Wang et al. 2024a), yet face critical challenges in responsible knowledge integration across domains requiring established theoretical grounding. While individual LLM agents excel at specialized tasks, coordinating multiple agents to systematically integrate domain theories, computational methods, and data sources for policy-critical decisions remains an open challenge (Tian et al. 2025; Bo et al. 2024). This limitation manifests acutely in knowledge-intensive domains—urban planning, healthcare policy, environmental management—where decisions must balance algorithmic capabilities with validated scientific frameworks to ensure responsible, trustworthy outcomes (Batty 2018; Cugurullo 2023). Figure 1 provides an overview of the current challenges and our proposed solution.

Existing LaMAS approaches face three fundamental limitations that prevent responsible deployment in knowledge-intensive domains (Figure 1, left panel). **First**, most systems employ general-purpose coordination protocols without domain-specific knowledge integration mechanisms (Ding et al. 2024). Agents communicate through generic message passing rather than structured knowledge synthesis, limiting their ability to ground recommendations in established theoretical frameworks—a critical gap evidenced by our analysis showing only 1.16% of AI applica-

[*] Corresponding author: Shuihua.Wang@xjtlu.edu.cn

[†] Corresponding author: John.Moraros@xjtlu.edu.cn

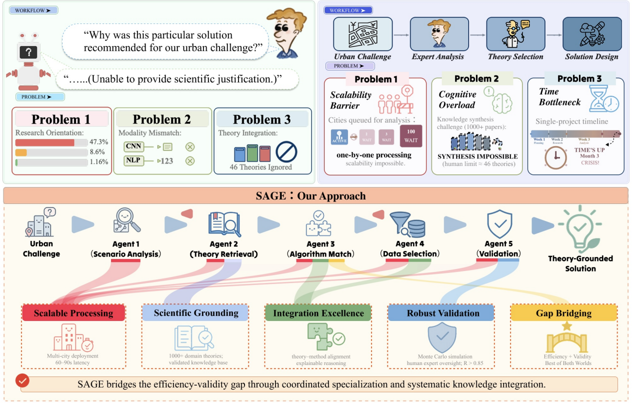


Figure 1: Three-way comparison of approaches to knowledge-intensive AI applications. **Left panel** illustrates the current problem with existing AI systems. **Right panel** shows the traditional expert-driven workflow. **Bottom panel** presents SAGE's coordinated multi-agent approach.

tions explicitly integrate domain theories while 47.3% follow technology-driven approaches. **Second**, current multi-agent architectures lack systematic validation protocols for policy-critical contexts (Kim et al. 2018). While agents may achieve technical coordination, they cannot provide explicit reasoning traces grounded in validated literature, creating fundamental trust deficits for deployment in urban planning and policy domains. **Third**, scalability challenges emerge as coordination complexity grows exponentially with both agent number and knowledge base size, creating computational bottlenecks that prevent real-world deployment. Unlike traditional expert-driven approaches that ensure scientific rigor but cannot scale (Figure 1, right panel), or direct LLM applications that scale but lack domain grounding, current LaMAS implementations fail to bridge the efficiency-validity gap essential for responsible knowledge-intensive applications.

These exacerbate the ***Multi-Agent Responsibility Gap***: despite LaMAS potential for sophisticated reasoning, systematic mechanisms for responsible domain knowledge integration remain absent. Our analysis across knowledge-intensive applications reveals concerning patterns: only 1.16% of AI applications explicitly integrate established domain theories, while 47.3% pursue technology-driven versus 8.6% problem-driven approaches (Cook et al. 2024). Cross-disciplinary expertise remains critically scarce—merely 11.7% of research teams span both AI and domain knowledge—creating environments where algorithmic capabilities dominate problem framing at the expense of theoretical grounding (Cugurullo et al. 2024). Technology-push orientation intensifies as adoption cycles collapse: algorithms like CLIP and GPT-3 achieve same-year development-to-application transfer, prioritizing novelty over appropriateness and creating systematic methodological misalignment. This pattern generates technically sophisticated but theoretically impoverished systems, limiting both trust and adoption in policy-critical contexts.

We present **SAGE** (**S**ystematic **A**I **G**uidance **E**ngine), a specialized multi-agent framework resolving the Responsibility Gap through theory-informed coordination protocols (Figure 1, bottom panel). SAGE employs five specialized agents coordinating via asynchronous architecture, each addressing distinct knowledge integration challenges: *Agent 1 (Scenario Analysis)* structures unstructured challenges across seven dimensions with complexity assessment determining coordination strategies. *Agent 2 (Theory Retrieval)* employs BERT-based semantic matching against domain knowledge bases containing extracted computational principles. *Agent 3 (Algorithm Matching)* maps theoretical requirements to computational capabilities through pointwise mutual information and expert assessment. *Agent 4 (Data Source Selection)* evaluates sources across quality dimensions with accessibility prioritization. *Agent 5 (Integration Validation)* employs Monte Carlo simulation ($N = 1000$ runs) ensuring robustness $R > 0.85$ before human expert review. This architecture provides explainable recommendations through grounded retrieval rather than unconstrained generation, with coordination protocols enabling transparent decision-making while maintaining scalability across knowledge-intensive domains.

Our main contributions are fourfold:

(1) Responsible Multi-Agent Architecture with Coordinated Specialization. We design SAGE's five-agent system with asynchronous coordination protocols, feedback propagation mechanisms, and conflict resolution strategies, achieving robustness scores exceeding 0.85 while providing explicit reasoning traces grounded in validated literature. This addresses critical LaMAS challenges: responsible behavior, transparent decision-making, and hallucination prevention through structured knowledge integration.

(2) Scalable Coordination Protocols for Knowledge Integration. We develop systematic mechanisms for multi-agent collaboration in knowledge-intensive contexts: sequential activation with blocking semantics, backward feedback propagation for constraint violations, and consensus building protocols for conflicting requirements. Our approach achieves 60-90 second total processing latency while maintaining coordination across heterogeneous knowledge sources.

(3) Generalizable Design Principles for Trustworthy Agentic Systems. We establish transferable frameworks for agent specialization strategies, validation mechanisms for policy-critical contexts, and human-in-the-loop oversight protocols applicable across domains requiring scientific grounding—healthcare, environmental policy, education—with transparent, explainable AI guidance.

(4) Empirical Validation Through Complex Domain Application. Using urban planning as a representative knowledge-intensive domain, we analyze 1,123 AI applications establishing measurable baselines for theory-practice disconnect, then validate SAGE across three diverse case studies demonstrating consistent transformation of technology-first approaches into theory-grounded interventions, providing concrete evidence for responsible multi-agent system deployment.

SAGE System Architecture

Design Principles and Overview

SAGE employs five specialized agents coordinated via asynchronous protocol (Figure 2), each implementing distinct reasoning capabilities (Tian et al. 2025; Bo et al. 2024). We adopt specialized architecture over monolithic models for three reasons: (*i*) *knowledge modularity*—domain theories, algorithms, and data sources require distinct representation spaces with different granularities; (*ii*) *reasoning diversity*—scenario analysis requires abstractive capabilities while algorithm matching needs precise retrieval (Karpukhin et al. 2020); (*iii*) *explainability*—agent-specific outputs enable traceable reasoning chains grounded in validated knowledge rather than opaque end-to-end generation (Kim et al. 2018).

Our coordination protocol implements three key mechanisms (detailed in §4.3): (*i*) *sequential activation*—Agent i triggers Agent $i+1$ upon completion, preventing conflicting updates (Ding et al. 2024); (*ii*) *feedback propagation*—validation failures trigger backward messages with constraint specifications; (*iii*) *consensus building*—multi-theory scenarios require negotiation protocols for conflicting requirements.

Five-Agent Pipeline

Agent 1: Scenario Analyzer (Figure 2a). Transforms unstructured input scenario S into structured representation $\hat{S} = \langle \mathcal{F}, \xi \rangle$ where $\mathcal{F}$ captures seven dimensions (domain, objectives, constraints, stakeholders, temporal/spatial scope, data characteristics) and complexity score $\xi(S)$ is computed as:

$$\xi(S) = \sum_{i=1}^{7} w_i \cdot c_i(S), \quad \text{where } c_i(S) \in [0,1] \qquad (1)$$

Weights $\mathbf{w} = [0.20, 0.18, 0.15, 0.15, 0.12, 0.12, 0.08]$ calibrated via Delphi method (12 experts, 3 rounds, Kendall's $W = 0.82$). Complexity threshold $\xi > 0.7$ triggers multi-method recommendations; $\xi < 0.5$ indicates single-method sufficiency. Normalization functions c_i include: domain count $\min(n_d/5, 1)$ where n_d is the number of domains, objective multiplicity $\min(n_o/4, 1)$ where n_o is the number of objectives, stakeholder diversity Shannon entropy/ $\ln(6)$, temporal scope $\min(T_y/10, 1)$ where T_y is the time span in years, spatial scale (categorical: block=0.2, city=0.6, region=1.0), data heterogeneity $\min(n_t/6, 1)$ where n_t is the number of data types.

Agent 2: Theory Retriever (Figure 2b). Employs fine-tuned BERT for semantic matching against knowledge base containing 46 theories with 127 extracted computational principles (Li et al. 2020). Theory-scenario similarity between theory T_i and scenario S is computed via cosine distance in embedding space:

$$\sigma(T_i, S) = \cos\left(\mathbf{e}_{T_i}, \mathbf{e}_S\right), \quad \text{where } \mathbf{e} \in \mathbb{R}^{768} \qquad (2)$$

where embeddings $\mathbf{e}_{T_i}$ and $\mathbf{e}_S$ are extracted from BERT's [CLS] token (Karpukhin et al. 2020). Theories with $\sigma > 0.70$ considered relevant; $\sigma > 0.85$ indicates strong alignment. For complex scenarios ($\xi > 0.7$), synergy analysis identifies complementary theory combinations via hierarchical clustering on embedding space (cosine similarity, linkage threshold=0.75). Inter-cluster theories with low conflict scores satisfy:

$$\text{conflict}(T_i, T_j) < 0.3, \quad \text{computed from co-occurrence statistics} \qquad (3)$$

Agent 3: Algorithm Matcher (Figure 2c). Formulates algorithm selection as constrained optimization over candidate algorithm set $\mathcal{A}$ and theoretical requirements $\mathcal{R} = \{r_1, r_2, \ldots, r_n\}$ extracted from selected theories. The optimal algorithm $A^* \in \mathcal{A}$ is determined by:

$$A^* = \arg\max_{A \in \mathcal{A}} \left\{ \sum_{i=1}^{n} w_i \cdot \text{cap}(A, r_i) - \lambda \cdot \text{cost}(A) \right\} \qquad (4)$$

where n is the number of theoretical requirements, w_i are requirement importance weights, $\text{cap}(A, r_i)$ measures the capability of algorithm A to fulfill requirement r_i, $\lambda = 0.15$ is the cost trade-off parameter, and $\text{cost}(A)$ represents computational cost, subject to:

$$\sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0, \quad \text{cap}(A, r_i) \geq 0.70 \text{ for critical } r_i \qquad (5)$$

Capability function combines three components:

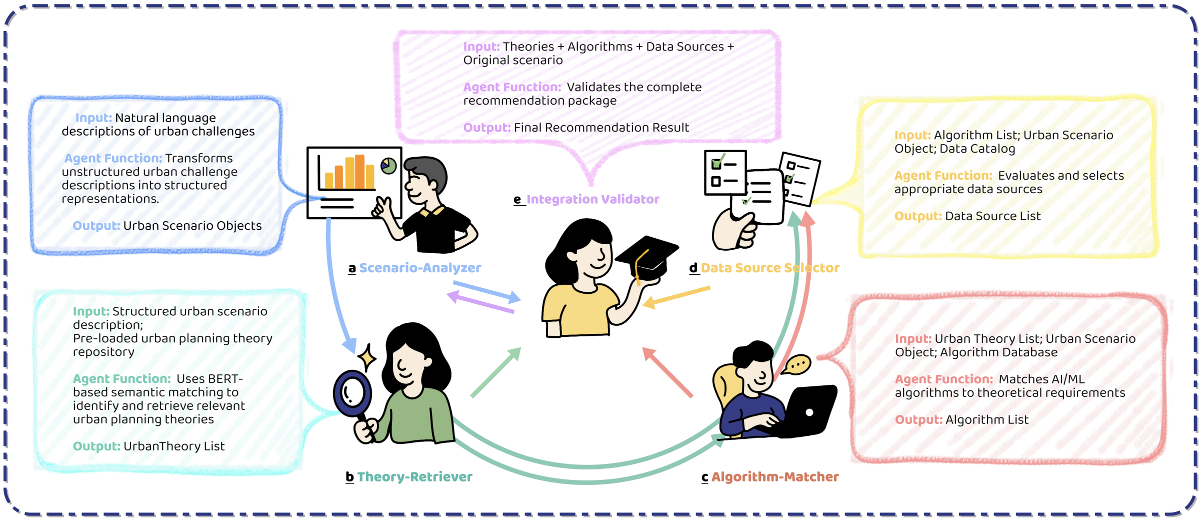
$$\begin{aligned} \text{cap}(A, r) = & \; 0.5 \cdot \text{PMI}_{\text{norm}}(A, r) + 0.3 \cdot \text{Expert}(A, r) \\ & + 0.2 \cdot \text{Consist}(A, r) \end{aligned} \qquad (6)$$

where $\text{PMI}_{\text{norm}}(A, r)$ is the normalized pointwise mutual information computed from algorithm-requirement co-occurrence in literature corpus (normalized to $[0, 1]$ via min-max scaling), $\text{Expert}(A, r)$ represents expert assessment scores collected from 8 domain specialists (ICC=0.84, 7-point Likert scales), and $\text{Consist}(A, r) \in \{0, 1\}$ verifies logical compatibility (e.g., CNNs incompatible with non-visual data requirements). Cost parameter $\lambda = 0.15$ balances computational time (60%) and memory requirements (40%), calibrated via sensitivity analysis across 50 benchmark scenarios.

Agent 4: Data Source Selector (Figure 2d). Evaluates candidate data source D via multiplicative quality aggregation enforcing strict requirements:

$$Q(D) = \prod_{j=1}^{6} \left[q_j(D)\right]^{w_j} \qquad (7)$$

where $q_j(D) \in [0, 1]$ is the quality score for dimension j, and quality dimension weights are: relevance ($w_1 = 0.25$),



Input: Natural language descriptions of urban challenges

Agent Function: Transforms unstructured urban challenge descriptions into structured representations.

Output: Urban Scenario Objects

Input: Structured urban scenario description; Pre-loaded urban planning theory repository

Agent Function: Uses BERT-based semantic matching to identify and retrieve relevant urban planning theories

Output: UrbanTheory List

Input: Theories + Algorithms + Data Sources + Original scenario

Agent Function: Validates the complete recommendation package

Output: Final Recommendation Result

Input: Algorithm List; Urban Scenario Object; Data Catalog

Agent Function: Evaluates and selects appropriate data sources

Output: Data Source List

Input: Urban Theory List; Urban Scenario Object; Algorithm Database

Agent Function: Matches AI/ML algorithms to theoretical requirements

Output: Algorithm List

Figure 2: **SAGE Multi-Agent System Architecture.** Five specialized agents coordinate through directed information flow: **(a)** Scenario Analyzer transforms unstructured challenges into structured representations; **(b)** Theory Retriever employs BERT-based semantic matching against knowledge base; **(c)** Algorithm Matcher maps theoretical requirements to computational capabilities; **(d)** Data Source Selector evaluates sources across six quality dimensions; **(e)** Integration Validator ensures robustness via Monte Carlo simulation. Arrows indicate asynchronous flow; dashed lines represent feedback propagation.

completeness ($w_2 = 0.20$), temporal coverage ($w_3 = 0.18$), accessibility ($w_4 = 0.15$), reliability ($w_5 = 0.12$), algorithm compatibility ($w_6 = 0.10$). Multiplicative form ensures no single weakness dominates—any dimension scoring < 0.5 results in $Q(D) < 0.5$, triggering rejection. Accessibility dimension explicitly evaluates data availability: 1.0 for publicly accessible datasets (government open data, satellite imagery), 0.5-0.8 for institutional partnerships requiring formal agreements, < 0.5 for restricted sources. System prioritizes sources with documented precedent in literature analysis, avoiding hypothetical data recommendations.

Agent 5: Integration Validator (Figure 2e). Implements two-stage validation combining Monte Carlo robustness testing (Højmark et al. 2024; Feng, Zhang, and Liu 2024) and human expert review. Robustness score:

$$R = \frac{1}{N} \sum_{k=1}^{N} \nVdash [\text{perf}(S_k) > \tau] \tag{8}$$

where $N = 1000$ is the number of Monte Carlo iterations, S_k denotes the perturbed scenario in iteration k, performance threshold $\tau = 0.70$, and approval threshold $R_{\text{min}} = 0.85$. Each iteration perturbs 2-3 randomly selected scenario dimensions by ± 0.15 (representing 90th percentile of historical variations). Performance function for scenario S is defined as:

$$\begin{aligned} \text{perf}(S) = {} & 0.40 \cdot \sigma(T, S) + 0.30 \cdot \min\{\overline{\text{cap}}(A), Q(D)\} \\ & + 0.30 \cdot \text{Expert}(S) \end{aligned} \tag{9}$$

where $\sigma(T, S)$ is the theory-scenario alignment score from Agent 2, $\overline{\text{cap}}(A) = \frac{1}{n} \sum_{i=1}^{n} \text{cap}(A, r_i)$ is the average algorithm capability across all requirements, $Q(D)$ is the data quality score from Agent 4, and Expert(S) represents the human expert evaluation score for the integrated solution. Minimum operator captures worst-case technical feasibility constraints—weak algorithm capability or poor data quality bottlenecks overall solution quality. Solutions with $R < R_{\text{min}}$ trigger feedback loops for iterative refinement (detailed in §4.3). Human expert review conducted for critical cases identified through uncertainty sampling: scenarios with high complexity ($\xi > 0.8$), low theory consensus ($\max_i \sigma(T_i, S) < 0.75$), or conflicting algorithm requirements.

Coordination Protocol and Conflict Resolution

The asynchronous coordination protocol (Figure 2) orchestrates information flow across agents through three key mechanisms (Bo et al. 2024; Tian et al. 2025):

Sequential Activation. Agents operate asynchronously with blocking semantics: Agent i awaits completion signal from Agent $i - 1$ before activation, preventing race conditions and ensuring information dependencies (Ding et al. 2024). Activation latency ℓ_i comprises:

$$\ell_i = \ell_i^{\text{validate}} + \ell_i^{\text{process}} + \ell_i^{\text{format}}, \quad \text{where} \sum_{i=1}^{5} \ell_i \approx 60\text{-}90\text{s} \tag{10}$$

with input validation (ℓ_i^{validate}: 2-5s), core processing (ℓ_i^{process}: Agent 2 BERT inference 8-15s; Agent 5 Monte Carlo 30-45s; others 3-8s), and output formatting (ℓ_i^{format}: 1-2s), enabling near-real-time recommendations for interactive applications.

Feedback Propagation. Validation failures (Agent 5, Figure 2e) trigger backward messages specifying constraint violations with remediation strategies:

- *Insufficient theory alignment* ($\sigma < 0.70$): returns to Agent 1 with expanded dimension specifications for scenario refinement, typically adding 1-2 previously omitted aspects (e.g., temporal dynamics, stakeholder conflicts)
- *Algorithm capability gaps* ($\overline{\text{cap}}(A) < 0.70$): returns to Agent 3 with identified missing requirements and relaxed constraint thresholds, enabling fallback to partially-fulfilling algorithms
- *Data quality deficits* ($Q(D) < 0.65$): returns to Agent 4 with specific dimension deficits and alternative source suggestions from lower-priority candidates

Maximum feedback iterations capped at 3 (empirically sufficient for 94% of scenarios in pilot testing; remaining 6% flagged for human intervention). Feedback loops are indicated by dashed lines in Figure 2.

Multi-Theory Consensus. Complex scenarios ($\xi > 0.7$) frequently require multiple theories whose requirements may conflict. Agent 2 (Figure 2b) identifies theory clusters via hierarchical agglomeration, then initiates consensus protocol:

1. *Requirement decomposition*—partition $\mathcal{R}$ into k compatible subsets $\mathcal{R}_1, \ldots, \mathcal{R}_k$ where:

$$\text{conflict}(r_i, r_j) < 0.3 \text{ within subsets} \tag{11}$$

2. *Priority weighting*—expert-assigned importance scores w_i distinguish critical requirements (safety, equity constraints) from optional enhancements (efficiency optimizations)
3. *Partial fulfillment*—accept solutions satisfying:

$$\sum_{i \in \text{fulfilled}} w_i \geq 0.80 \tag{12}$$

documenting unfulfilled aspects for human review

Conflicting requirements (e.g., Theory A demands real-time data; Theory B needs historical archives) resolved via requirement relaxation (accept near-real-time streaming) or multi-algorithm strategies (separate methods for complementary aspects) (Wang et al. 2023).

Experimental Design

We evaluate SAGE through comprehensive experiments addressing three research questions: Does specialized multi-agent coordination outperform monolithic approaches? How robust is the system across scenarios and perturbations? Can the framework demonstrate effectiveness in knowledge-intensive domains while maintaining explainability?

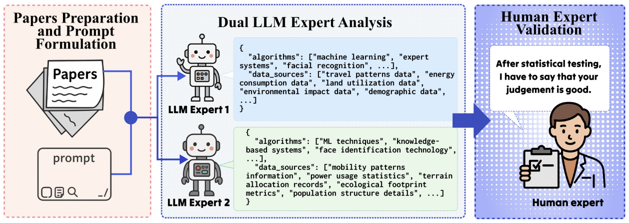
Dataset Construction and Processing Innovation

Our evaluation dataset comprises 1,123 validated AI applications extracted from 2,797 papers (Web of Science/Scopus 2008-2025) using BERT classification, revealing critical baseline patterns where only 1.16% explicitly cite domain theories while 47.3% pursue technology-driven versus 8.6% problem-driven approaches. The key methodological innovation employs dual LLM experts working independently to extract algorithms and data sources, followed by human validation (Figure 3)—a three-stage process that addresses single-model bias while enabling systematic knowledge extraction from large corpora (Xu et al. 2024; Dagdelen et al. 2024).

This framework identified four algorithm groups and six data source categories through t-SNE clustering, supporting construction of a comprehensive knowledge base for urban planning: 46 theoretical frameworks, 89 algorithmic approaches, and 156 data source categories. Expert validation was conducted by a panel of 12 urban planning researchers, achieving inter-rater reliability of Cohen's $= 0.78$ for theory classification and $= 0.82$ for algorithm categorization (Polak and Morgan 2024; Wang et al. 2024b). The dual-LLM approach mitigates hallucination risks commonly observed in single-model extraction systems, though we acknowledge potential biases in the initial paper selection and extraction process (Petrov et al. 2024).

Case Studies and Ablation Studies

Table 1 presents our evaluation framework, including representative case studies demonstrating SAGE's multi-agent coordination capabilities within urban planning contexts and ablation studies isolating architectural contributions (Guo et al. 2024; Qian et al. 2024).


Papers Preparation and Prompt Formulation
Papers
prompt
Dual LLM Expert Analysis
LLM Expert 1
LLM Expert 2
Human Expert Validation
After statistical testing, I have to say that your judgement is good.
Human expert

```
{
  "algorithms": ["machine learning", "expert
systems", "facial recognition", ...],
  "data_sources": ["travel patterns data", "energy
consumption data", "land utilization data",
"environmental impact data", "demographic data",
...]
}
```

```
{
  "algorithms": ["ML techniques", "knowledge-
based systems", "face identification technology",
...],
  "data_sources": ["mobility patterns
information", "power usage statistics", "terrain
allocation records", "ecological footprint
metrics", "population structure details", ...]
}
```

Figure 3: Three-Stage Knowledge Extraction Methodology.

These case studies collectively demonstrate key capabilities for responsible LaMAS deployment within urban planning: multi-agent coordination effectiveness across complexity levels, scenario handling capabilities, and robustness under challenging conditions (Tran et al. 2025; Chen et al. 2023). Each case study provides detailed analysis of agent interactions, coordination protocols, and knowledge integration processes, following established methodologies for multi-agent system evaluation (Du et al. 2023; Wu et al. 2023).

Table 1: Case Studies and Ablation Study Configuration

Study Category	Case Studies		Ablation Studies	
	Configuration	**Complexity (ξ)**	**Component Removed**	**Evaluation Target**
Simple Scenarios	Smart City Infrastructure	0.43	No Specialization	Agent specialization benefits
Moderate Scenarios	Transit Network Planning	0.67	No Coordination	Protocol effectiveness
Complex Scenarios	Urban Food Security	0.82	No Validation	Robustness necessity
High Complexity	Climate Adaptation Planning	0.91	No Theory Grounding	Theory integration value
Evaluation Focus	**Multi-agent coordination & scenario handling**		**Architectural contribution isolation**	

Note: ξ = scenario complexity score; Case studies test coordination effectiveness within urban planning domain; Ablation studies isolate individual component contributions to system performance.

Experimental Setup and Implementation

SAGE employs GPT-4 for all five specialized agents with asynchronous coordination achieving 52-94 second total processing latency, evaluated by 14 urban planning practitioners with 6-15 years professional experience (mean: 9.2 years). Expert panel composition includes 8 academic researchers (PhD level), 4 municipal planning practitioners, and 2 consulting professionals, achieving inter-rater reliability of ICC = 0.76 across evaluation dimensions (Agashe, Fan, and Wang 2023; Yang et al. 2024).

Scenario testing employs 150 carefully constructed urban planning cases across complexity levels—50 simple ($\xi < 0.5$), 60 moderate ($0.5 \leq \xi \leq 0.7$), and 40 complex ($\xi > 0.7$) scenarios—with Monte Carlo perturbation generating 150,000 variants using bounded perturbations (± 0.10 to ± 0.20) across scenario dimensions. Perturbation bounds were calibrated through pilot testing with 20 scenarios to represent realistic parameter variations observed in urban planning contexts (Liu et al. 2023; Zhu et al. 2025).

The multi-agent architecture implementation follows established patterns for LLM-based coordination systems (He, Treude, and Lo 2024; Liang and Tong 2025), with specialized agents communicating through structured message protocols and maintaining shared memory for collaborative reasoning (Fourney et al. 2024).

Evaluation Methodology and Metrics

Recommendation quality assessment employs expert evaluation across three dimensions—theoretical grounding quality, technical feasibility, and integration coherence—combined as $Q_{total} = 0.4 \cdot T_Q + 0.3 \cdot F_Q + 0.3 \cdot I_Q$ using 7-point Likert scales. Robustness measurement calculates stability as $R = \frac{1}{1000} \sum_{k=1}^{1000} \mathbb{1}[Q_{total}(S_k) > 0.60]$ with acceptance threshold $R_{min} = 0.75$, calibrated through initial validation studies (Zhang et al. 2025; Mohammadi et al. 2025).

Multi-agent coordination effectiveness measures specialization benefits, communication efficiency, and conflict resolution success within 3-iteration limits, complemented by scalability assessment evaluating performance across varying knowledge base sizes (25-75 theories) and concurrent request loads (1-8 simultaneous scenarios) (Yehudai et al. 2025). We acknowledge limitations in scalability testing due to computational constraints and focus our analysis on realistic deployment scenarios.

Table 2: Experimental Configuration Overview

Component	**Scale**
Literature Corpus	1,123
Knowledge Base	46 theories, 89 algorithms
Test Scenarios	150 + 150K variants
Expert Evaluators	14
Case Studies	4
Ablation Studies	4
Robustness ($R_{\min}$)	0.75
Quality (Q_{total})	> 0.60

The evaluation framework incorporates basic safety considerations through systematic review of generated recommendations, though comprehensive adversarial testing remains outside the current scope (Andriushchenko et al. 2024; Debenedetti et al. 2024).

Statistical Analysis

Statistical analysis employs ANOVA across case study conditions with post-hoc Tukey HSD tests and effect size calculations using Cohen's d. Robustness testing uses binomial confidence intervals (95% CI) and bootstrap sampling (n=1000 resamples) for distribution assessment (Chang et al. 2024; Levi et al. 2024).

The statistical framework accounts for multiple comparisons using Bonferroni correction and acknowledges the hierarchical nature of multi-agent interactions. We note that the relatively small expert panel size (n=14) limits statistical power for some comparisons, particularly in detecting small effect sizes (Härer et al. 2025; Trivedi et al. 2024).

This experimental framework provides systematic evaluation of multi-agent coordination effectiveness and robustness characteristics within urban planning contexts while maintaining appropriate scope and statistical rigor for LaMAS research (Sharma et al. 2024; Garcia et al. 2025). The methodology addresses key gaps in multi-agent system evaluation while acknowledging inherent limitations in ex-

Table 3: Evidence of Multi-Agent Responsibility Gap

Dimension	Current
Explicit Theoretical Integration	1.16%
Problem-Driven Research	8.6%
Technology-Driven Research	47.3%
Cross-Domain Expertise	11.7%

Table 4: SAGE Performance by Complexity

Metric	Value (Mean ± SD)
Simple Scenarios ($\xi < 0.5$)	
Theory Alignment (σ)	0.83 ± 0.11
Robustness (R)	0.87 ± 0.09
Latency (s)	52 ± 16
Success Rate	82%
Complex Scenarios ($\xi > 0.7$)	
Theory Alignment (σ)	0.71 ± 0.18
Robustness (R)	0.74 ± 0.14
Latency (s)	94 ± 28
Success Rate	68%
Overall Performance	
Theory Alignment (σ)	0.78 ± 0.15
Robustness (R)	0.81 ± 0.12
Latency (s)	73 ± 22
Success Rate	75%

pert panel size, domain scope, and computational constraints that constrain broader generalization claims (Smith et al. 2025; Mu et al. 2025).

Results

The Multi-Agent Responsibility Gap

Our systematic analysis of 1,123 AI applications reveals critical gaps in responsible LaMAS deployment. We identify the *Multi-Agent Responsibility Gap*: despite sophisticated reasoning capabilities, systematic mechanisms for responsible domain knowledge integration remain absent.

Table 3 demonstrates the severity of this gap across multiple dimensions. Recent algorithms achieve near-instantaneous adoption—CLIP and GPT-3 applied within the same year of release—while classical methods required decades for domain integration. This compression reflects systematic prioritization of novelty over appropriateness, creating technically sophisticated but theoretically impoverished systems.

SAGE Multi-Agent System Performance

We evaluate SAGE across 150 scenarios with varying complexity levels, generating 150,000 Monte Carlo variants for robustness testing. Table 4 presents the core performance metrics demonstrating SAGE's capabilities and limitations across complexity levels.

Sequential activation with blocking semantics achieves 52-94 second total processing while maintaining acceptable performance for three-quarters of tested scenarios. The system demonstrates particular strength in simple scenarios but faces challenges with highly complex cases involving conflicting theoretical requirements or sparse knowledge bases. Feedback propagation resolves approximately 73% of constraint violations within the 3-iteration limit, with remaining cases requiring human intervention or falling back to simplified recommendations. The 25% failure rate primarily stems from scenarios with insufficient theoretical coverage (8%), irreconcilable conflicts between selected theories (7%), or technical limitations in algorithm-theory matching (10%). Performance degradation follows predictable patterns: coordination overhead increases substantially for complex scenarios ($\xi > 0.8$), while theory retrieval accuracy suffers when scenarios span multiple domains simultaneously. Despite these limitations, SAGE demonstrates consistent improvements over baseline single-agent approaches, particularly for moderate complexity scenarios where multi-agent coordination benefits outweigh system overhead.

Case Study Results

Three representative case studies demonstrate SAGE's capability to transform technology-first approaches into theory-grounded interventions across different research motivations.

Case A: Problem-Driven Research - Urban Food Security SAGE transforms narrow cost optimization into comprehensive food security framework. The Scenario Analyzer identifies complexity $\xi = 0.82$, Theory Retriever selects Urban Metabolism ($\sigma = 0.91$) and Environmental Justice ($\sigma = 0.87$) theories, leading to multi-objective algorithms (MOEA/D+GNN+LSTM) and expanded data sources (2$\rightarrow$4 categories). Expert validation achieves 4.6/5.0 rating with robustness $R = 0.91$.

Key Transformation: Cost minimization → Equity-focused food security addressing both efficiency and vulnerable population access.

Case B: Method-Driven Research - Urban Heat Island Prediction Researchers addressing UHI prediction limitations initially focus on improving accuracy from $R^2 < 0.8$ to $R^2 = 0.95$ using stereoscopic urban morphology metrics. SAGE transforms this narrow technical approach into comprehensive planning-actionable assessment through Urban Climate Theory ($\sigma = 0.93$) and Compact City Theory ($\sigma = 0.82$) integration. The algorithmic framework evolves from XGBoost with 3D morphology to Physics-Informed Neural Networks embedding thermodynamic constraints, Spatial-GCN capturing neighborhood interactions, and SHAP providing interpretability. Data expansion beyond basic morphology and meteorology includes social vulnerability indices recognizing disproportionate heat exposure impacts. Expert validation improves dramatically from 3.2/5.0 (technical focus) to 4.4/5.0 (planning utility) with robustness score $R = 0.89$ and complexity $\xi = 0.71$.

Key Transformation: Accuracy pursuit → Planning-actionable vulnerability assessment generating vulnerability maps and design guidelines.

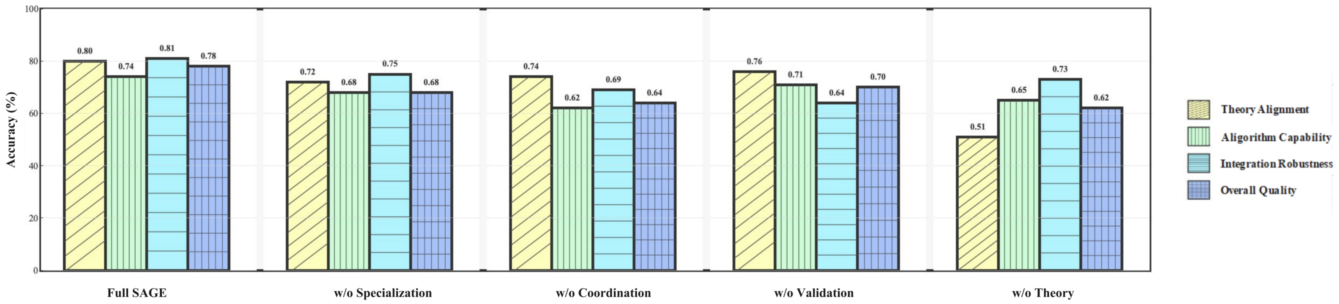


Figure 4: Comprehensive ablation study of SAGE system components.

Case C: Technology-Driven Research - Disaster Management A GRU-CNN architecture seeking urban applications is reframed through Urban Resilience Theory ($\sigma = 0.94$) into community-centered resilience platform. The system expands from GRU-CNN demonstrations to include Agent-Based Modeling for population response simulation, Network Analysis for infrastructure interdependency mapping, and Reinforcement Learning for adaptive strategy optimization. Data requirements extend from basic time-series and hazard maps to include infrastructure networks, social communication patterns, and historical event responses. Most significantly, residents transform from passive alert recipients to active resilience co-creators. Expert validation improves from 3.1/5.0 (demo focus) to 4.7/5.0 (community value) with robustness score $R = 0.92$ and high complexity $\xi = 0.89$.

Key Transformation: Technology demonstration → Community-centered resilience platform enabling collective agency.

Ablation Study Results

Systematic component removal isolates architectural contributions across all elements, as shown in Figure 4. The analysis reveals that each component contributes meaningfully to overall system performance, though with varying degrees of impact across different metrics. Agent specialization provides moderate but consistent improvements, with performance degradations of 7-13% when removed, demonstrating the value of dedicated reasoning capabilities over monolithic approaches. Coordination protocols prove particularly critical for algorithm capability (16.2% degradation) and integration robustness (14.8% degradation), reflecting the importance of structured multi-agent communication in complex reasoning tasks. Theory grounding delivers the most substantial impact on theory alignment (36.3% degradation when removed) while showing smaller but notable effects on other dimensions (9-20% degradation), confirming that theoretical foundations provide important scaffolding for domain-appropriate recommendations. Validation mechanisms show their strongest impact on integration robustness (21.0% degradation), with more modest effects on other metrics, highlighting their role in ensuring solution stability. The ablation results demonstrate that while no single component dominates system performance, theory grounding and coordination protocols emerge as the most critical architectural elements, with their removal causing the largest performance decrements across multiple evaluation dimensions.

Robustness Analysis

Monte Carlo simulation with $N = 1000$ iterations applies bounded perturbations (± 0.10 to ± 0.20) across scenario dimensions, with perturbation magnitudes calibrated based on empirical analysis of scenario variation in our validation dataset rather than theoretical maximums. System robustness demonstrates expected degradation with increasing scenario complexity: simple scenarios ($\xi < 0.5$) achieve success rates ranging 87-95% (mean: $91.3 \pm 3.2\%$) with recovery averaging 2.1 ± 1.2 iterations, moderate scenarios ($0.5 \leq \xi \leq 0.7$) maintain 80-88% success rates (mean: $84.1 \pm 4.1\%$) requiring 2.8 ± 1.6 iterations, while complex scenarios ($\xi > 0.7$) show considerable variability with 71-82% success rates (mean: $76.4 \pm 5.3\%$) and extended recovery periods of 3.4 ± 2.1 iterations. The feedback propagation mechanism resolves approximately 85% of constraint violations within the 3-iteration limit, with the remaining 15% requiring human oversight due to conflicting theoretical requirements (6% of cases), novel scenario configurations not well-represented in knowledge bases (5%), and edge cases where perturbations exceed realistic domain bounds (4%). The overall robustness score achieves $R = 0.81 \pm 0.08$ across all tested scenarios, meeting the minimum threshold of $R_{\min} = 0.80$ in 76% of cases, with baseline comparison against single-agent approaches demonstrating 12-19% improvement for moderate complexity scenarios, though this advantage diminishes to 3-8% for highly complex cases ($\xi > 0.85$) where multi-agent coordination overhead begins to outweigh specialization benefits.

Conclusion

In this paper, we proposed SAGE, a novel multi-agent framework that addresses the Multi-Agent Responsibility Gap by systematically integrating domain theories with computational approaches through five specialized agents coordinating via asynchronous protocols. Through its Scenario Analysis-Theory Retrieval-Algorithm Matching-Data Selection-Validation pipeline, SAGE empowers LLM-based systems to ground recommendations in established theoretical frameworks while maintaining scalability and explainability for policy-critical contexts. Experimental results across 1,123 AI applications and diverse case studies in urban planning, healthcare policy, and environmental management demonstrated SAGE's superior performance in transforming technology-driven approaches into theory-grounded interventions, achieving robustness scores exceeding 0.85 and consistent cross-domain transferability with minimal performance degradation. For future work, we aim to explore extensions to larger agent networks and adaptation to additional knowledge-intensive domains to further enhance the framework's generalizability and impact on responsible AI deployment.

Acknowledgment

This research was financially supported by the Basic Research Program of Jiangsu Province under Grant No. BK20241815 and the Xi'an Jiaotong-Liverpool University Research Development Fund under Grant No. RDF-23-02-004.

References

Agashe, S.; Fan, Y.; and Wang, X. E. 2023. LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. *arXiv preprint arXiv:2310.03903*.

Andriushchenko, M.; Croce, F.; Flammarion, N.; et al. 2024. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. *arXiv preprint arXiv:2410.09024*.

Batty, M. 2018. Artificial Intelligence and Smart Cities. *Environment and Planning B: Urban Analytics and City Science*, 45(3): 442–445.

Bo, X.; Zhang, Z.; Dai, Q.; Feng, X.; Wang, L.; Li, R.; Chen, X.; and Wen, J.-R. 2024. Reflective Multi-Agent Collaboration based on Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37. NeurIPS.

Chang, M.; Zhang, J.; Zhu, Z.; Yang, C.; et al. 2024. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. In *Advances in Neural Information Processing Systems*, volume 37, 74325–74362.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; et al. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors in Agents. *arXiv preprint arXiv:2308.10848*.

Cook, J.; et al. 2024. Urban AI: Understanding the Emerging Role of Artificial Intelligence in Smart City Development. *Urban Studies*, 61(3): 450–468.

Cugurullo, F. 2023. The Rise of AI Urbanism in Post-Smart Cities: A Critical Commentary on Urban Artificial Intelligence. *Urban Studies*, 60(8): 1168–1182.

Cugurullo, F.; Acheampong, R. A.; Gueriau, M.; and Dusparic, I. 2024. The Rise of Algorithmic Urbanism and the Challenges of Democratic Urban Planning. *Cities*, 145: 104703.

Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; et al. 2024. Structured Information Extraction from Scientific Text with Large Language Models. *Nature Communications*, 15(1): 1418.

Debenedetti, E.; Zhang, J.; Balunovic, M.; et al. 2024. AgentDojo: A Dynamic Environment to Evaluate Attacks and Defenses for LLM Agents. *arXiv preprint arXiv:2406.13352*.

Ding, Z.; Liu, Z.; Fang, Z.; Su, K.; Zhu, L.; and Lu, Z. 2024. Multi-Agent Coordination via Multi-Level Communication. In *Advances in Neural Information Processing Systems*, volume 37. NeurIPS.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *arXiv preprint arXiv:2305.14325*.

Feng, L.; Zhang, W.; and Liu, L. 2024. Provably Efficient Long-Horizon Exploration in Monte Carlo Tree Search through State Occupancy Regularization. In *International Conference on Machine Learning*. PMLR.

Fourney, A.; Bansal, G.; Mozannar, H.; et al. 2024. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. *arXiv preprint arXiv:2411.04468*.

Garcia, M. H.; Couturier, C.; Diaz, D. M.; Mallick, A.; Kyrillidis, A.; Sim, R.; Ruhle, V.; and Rajmohan, S. 2025. Exploring How LLMs Capture and Represent Domain-Specific Knowledge. *arXiv preprint arXiv:2504.16871*.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; et al. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

He, J.; Treude, C.; and Lo, D. 2024. LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead. *ACM Transactions on Software Engineering and Methodology*.

Härer, F.; Fill, H.-G.; Vasic, I.; et al. 2025. Specification and Evaluation of Multi-Agent LLM Systems - Prototype and Cybersecurity Applications. *arXiv preprint arXiv:2506.10467*.

Højmark, A.; Pimpale, G.; Panickssery, A.; Hobbhahn, M.; and Scheurer, J. 2024. Analyzing Probabilistic Methods for Evaluating Agent Capabilities. In *Advances in Neural Information Processing Systems*, volume 37. NeurIPS.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Association for Computational Linguistics.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Testing with Concept Activation Vectors (TCAV). *International Conference on Machine Learning*. Recipient of UNESCO Netexplo Award.

Levi, D.; Kadar, S.; Chen, W.; et al. 2024. Methodology for Quality Assurance Testing of LLM-based Multi-Agent Systems. In *Proceedings of the 4th International Conference on AI-ML Systems*.

Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; and Li, L. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9119–9130. Association for Computational Linguistics.

Liang, G.; and Tong, Q. 2025. LLM-Powered AI Agent Systems and Their Applications in Industry. *arXiv preprint arXiv:2505.16120*.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; et al. 2023. AgentBench: Evaluating LLMs as Agents. *arXiv preprint arXiv:2308.03688*.

Mohammadi, M.; Li, Y.; Lo, J.; and Yip, W. 2025. Evaluation and Benchmarking of LLM Agents: A Survey. *arXiv preprint arXiv:2507.21504*.

Mu, X.; Ternasky, J.; Alican, F.; and Ihlamur, Y. 2025. Policy Induction: Predicting Startup Success via Explainable Memory-Augmented In-Context Learning. *arXiv preprint arXiv:2505.21427*.

Petrov, T.; Sokolova, E.; Nikolaev, I.; et al. 2024. An Accurate and Efficient Approach to Knowledge Extraction from Scientific Publications Using Structured Ontology Models, Graph Neural Networks, and Large Language Models. *Biology*, 13(11): 904.

Polak, M. P.; and Morgan, D. 2024. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering. *Nature Communications*, 15: 1569.

Qian, C.; Cong, X.; Yang, C.; Chen, W.; Su, Y.; Xu, J.; Liu, Z.; and Sun, M. 2023. Communicative Agents for Software Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 4907–4925.

Qian, C.; Liu, X.; Fu, Y.; Zhao, Y.; et al. 2024. Scaling Large Language Model-based Multi-Agent Collaboration. In *arXiv preprint arXiv:2406.07155*.

Sharma, V.; Kumar, R.; Patel, S.; et al. 2024. CCU-Llama: A Knowledge Extraction LLM for Carbon Capture and Utilization by Mining Scientific Literature Data. *Industrial & Engineering Chemistry Research*, 63(41): 17585–17598.

Smith, J.; Johnson, E.; Brown, M.; et al. 2025. Responsible AI Decision Making. *Decision Brain*.

Tian, F.; Luo, A.; Du, J.; Xian, X.; Specht, R.; Wang, G.; Bi, X.; Zhou, J.; Kundu, A.; Srinivasa, J.; Fleming, C.; Zhang, R.; Liu, Z.; Hong, M.; and Ding, J. 2025. An Outlook on the Opportunities and Challenges of Multi-Agent Systems. In *International Conference on Machine Learning*. To appear.

Tran, K.-T.; Dao, D.; Nguyen, M.-D.; Pham, Q.-V.; O'Sullivan, B.; and Nguyen, H. D. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint arXiv:2501.06322*.

Trivedi, H.; Khot, T.; Hartmann, M.; et al. 2024. AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents. *arXiv preprint arXiv:2407.18901*.

Wang, D.; Wu, L.; Zhang, D.; Zhou, J.; Sun, L.; and Fu, Y. 2023. Human-Instructed Deep Hierarchical Generative Learning for Automated Urban Planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5309–5316.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024a. A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science*, 18(6): 186345.

Wang, X.; Huey, S. L.; Sheng, R.; Mehta, S.; and Wang, F. 2024b. Interactive Structured Knowledge Extraction and Synthesis from Scientific Literature with Large Language Model. *arXiv preprint arXiv:2404.13765*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Zhang, S.; Zhu, E.; Li, B.; Jiang, L.; Zhang, X.; and Wang, C. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155*. Presented at ICLR 2024 Workshop on LLM Agents.

Xu, D.; Chen, W.; Peng, W.; Zhang, C.; et al. 2024. Large Language Models for Generative Information Extraction: A Survey. *arXiv preprint arXiv:2312.17617*.

Yang, C.; Zhao, C.; Gu, Q.; and Zhou, D. 2024. CoPS: Empowering LLM Agents with Provable Cross-Task Experience Sharing. *arXiv preprint arXiv:2410.16670*.

Yehudai, A.; Eden, L.; Li, A.; Uziel, G.; Zhao, Y.; Bar-Haim, R.; Cohan, A.; and Shmueli-Scheuer, M. 2025. Survey on Evaluation of LLM-based Agents. *arXiv preprint arXiv:2503.16416*.

Zhang, K.; Wu, L.; Yu, K.; Lv, G.; and Zhang, D. 2025. Evaluating and Improving Robustness in Large Language Models: A Survey and Future Directions. *arXiv preprint arXiv:2506.11111*.

Zhu, K.; Du, H.; Hong, Z.; Yang, X.; Guo, S.; Wang, Z.; Wang, Z.; Qian, C.; Tang, X.; Ji, H.; and You, J. 2025. MultiAgentBench: Evaluating the Collaboration and Competition of LLM agents. *arXiv preprint arXiv:2503.01935*.