

Conversations That Build:

How AI Learns to **Heal**, **Teach**, **Play**, and **Judge**



Heal

Psychotherapy Dialogue



Teach

Learning Assistants



Play

Collaborative Building



Judge

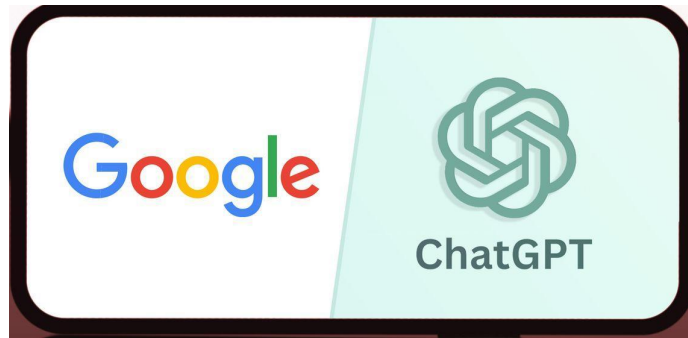
Legal Reasoning

Presented by: **Jiahuan Pei**

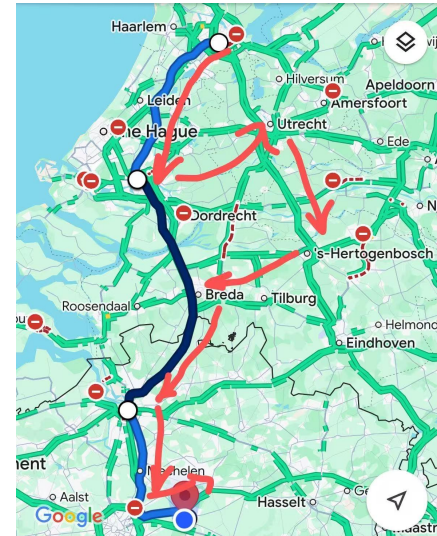
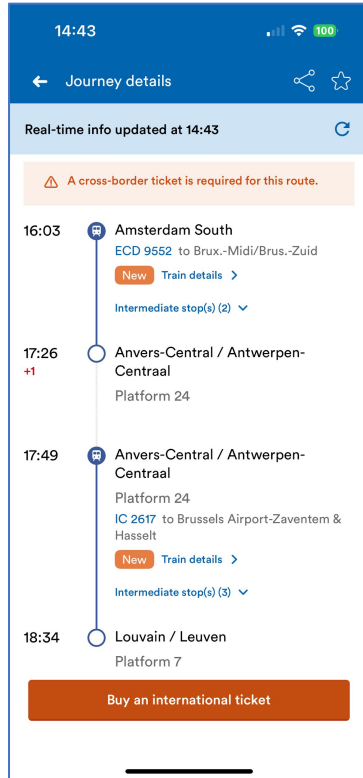
Date: 24.09.2025

Conversations shape how we learn, create, and connect.

Can complex or real-world problems ever be solved
with a single prompt?



Conversations shape how we learn, create, and connect.



Agenda Overview



1. Introduction

The Rise of LLM Agents and Multi-Agent Systems



2. Case Studies

♥ Heal: MI-SAG & DSM5AgentFlow



Teach: LEGO-MRTA & WikiHowAgent

⊕ Play: LLM-Assisted Interface



Judge: Legal Judgment



3. Cross-Cutting Challenges


Trust, Transparency, and Consensus

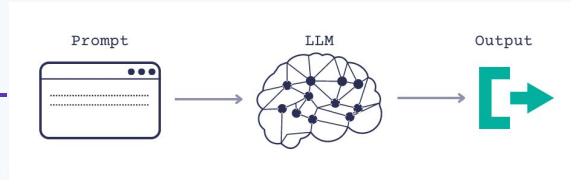



4. Conclusion

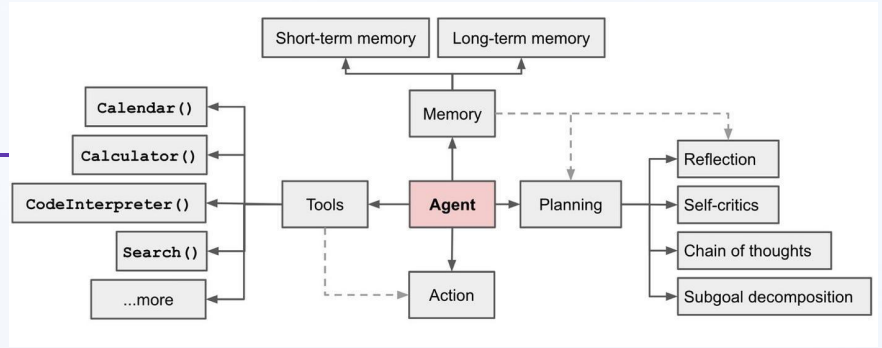
Summary & Future Directions


The Shift to Conversational Agents


Static Q&A
Simple question answering




Planning Agent
Task decomposition




Multi-Agent System
Collaborative problem-solving

Multi-LLM Agent Systems Enhance Problem-Solving

Collective Reasoning

Enhanced decision-making through agent collaboration

Cooperative Decision-Making

Agents working toward a common goal

Error Reduction

Agents check and correct each other's work

Scalable Solutions

Handling complex tasks through structured collaboration

Multi-LLM Agent System



Case Study 1: Heal – Sensitive Contexts

The Challenge of Psychotherapy Dialogue

The application of LLMs in psychotherapy presents significant challenges due to the **sensitive nature** of the context and the **inherent lack of transparency**.

Why Current Approaches Fall Short

Rule-Based Chatbots

Rigid and resource-intensive, lacking flexibility needed for personalized therapy

Pure LLMs

Lack controllability and therapeutic structure, hard to ensure safety

Key Requirements



Controllability

Ability to guide conversations toward therapeutic goals while adapting to individual needs



Explainability

Clear rationale for diagnostic conclusions and treatment recommendations



Therapeutic Structure

Adherence to established frameworks like Motivational Interviewing (MI) or Cognitive Behavioral Therapy (CBT)



Safety

Particular focus on sensitive content handling and alignment with established therapeutic principles

Aligning LLMs for Therapeutic Adherence

In sensitive contexts like psychotherapy, controllability and explainability are vital. LLMs must adhere to established therapeutic principles while maintaining flexibility.



Strategy-Aligned Generation (SAG)

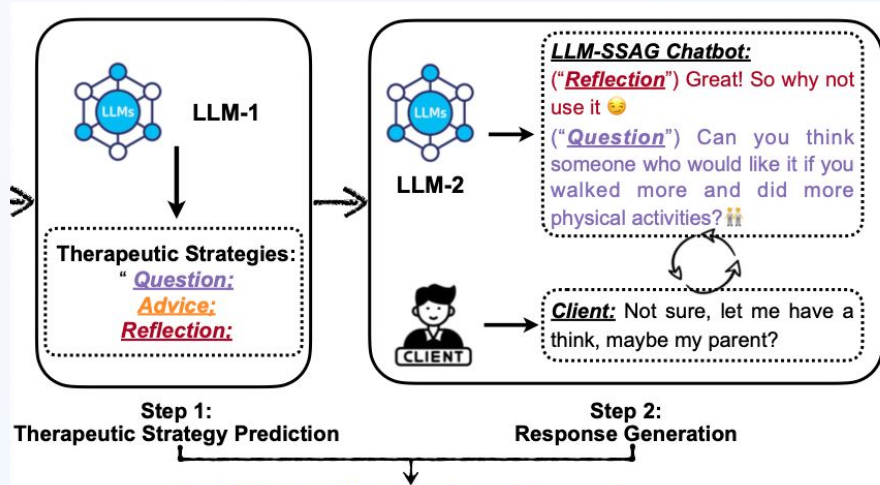
Aligns LLMs with Motivational Interviewing (MI) strategies to generate controllable and explainable dialogues.

Key Components

- ✓ Aligns with MI strategies (reflections, questions)
- ✓ Ensures adherence to therapeutic principles
- ✓ Maintains explainability in responses

Controllability

High



Aligning LLMs for Therapeutic Adherence

In sensitive contexts like psychotherapy, controllability and explainability are vital. LLMs must adhere to established therapeutic principles while maintaining flexibility.



Script-Strategy Aligned Generation (SSAG)

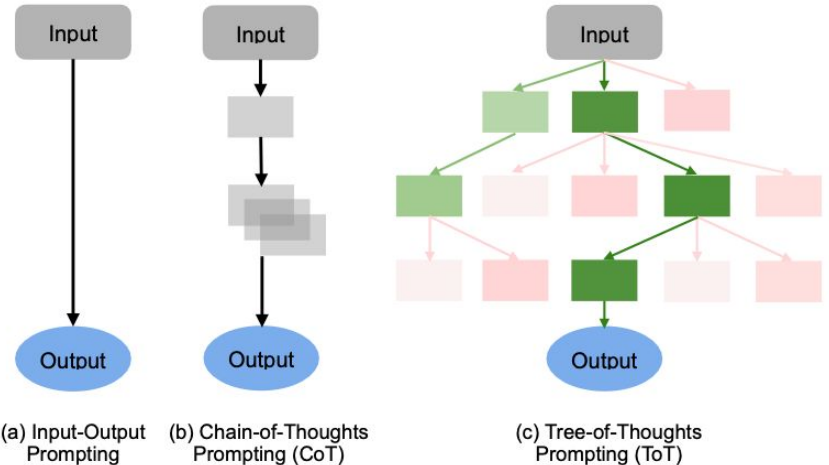
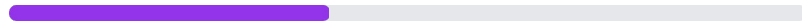
Refines SAG by integrating partial expert input while maintaining dynamic response generation.

Key Benefits

- ✓ Experts provide only partial input (topics, key questions)
- ✓ LLMs handle dynamic response generation and dialogue flow
- ✓ Based on predicted therapeutic strategies (asking questions, reflective listening)

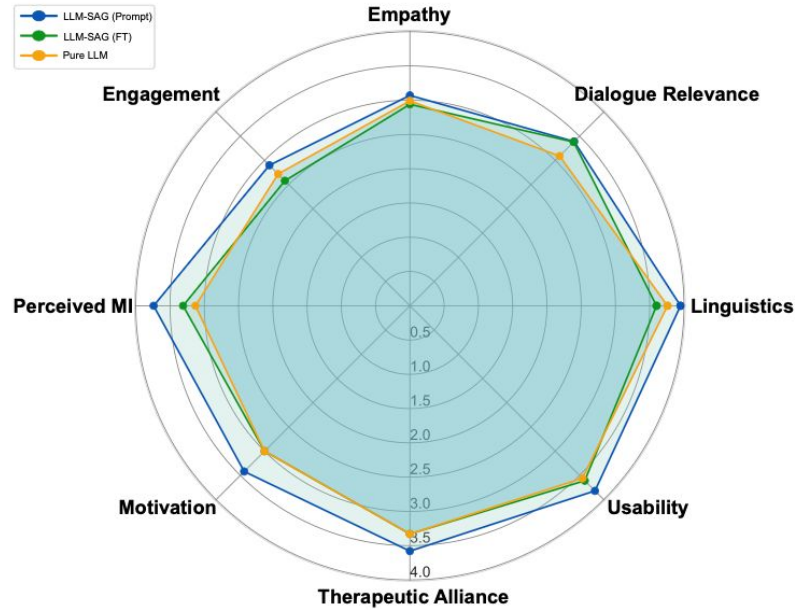
Expert Content Required

Less than 40%

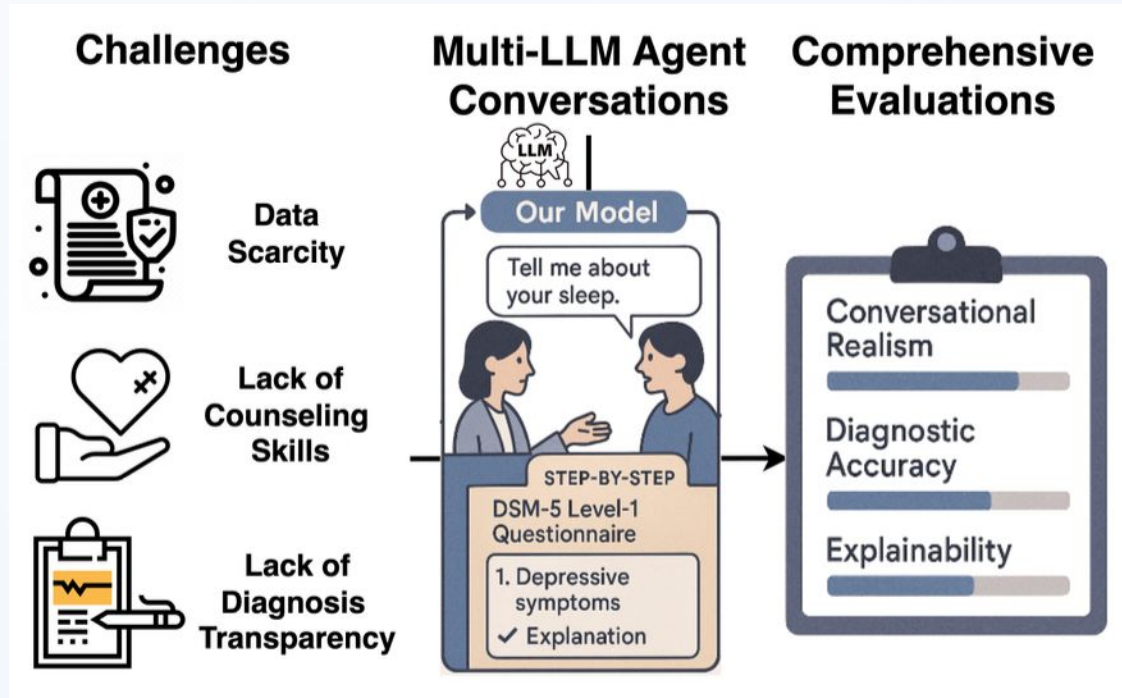


Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy. In *CSCW 2025*.

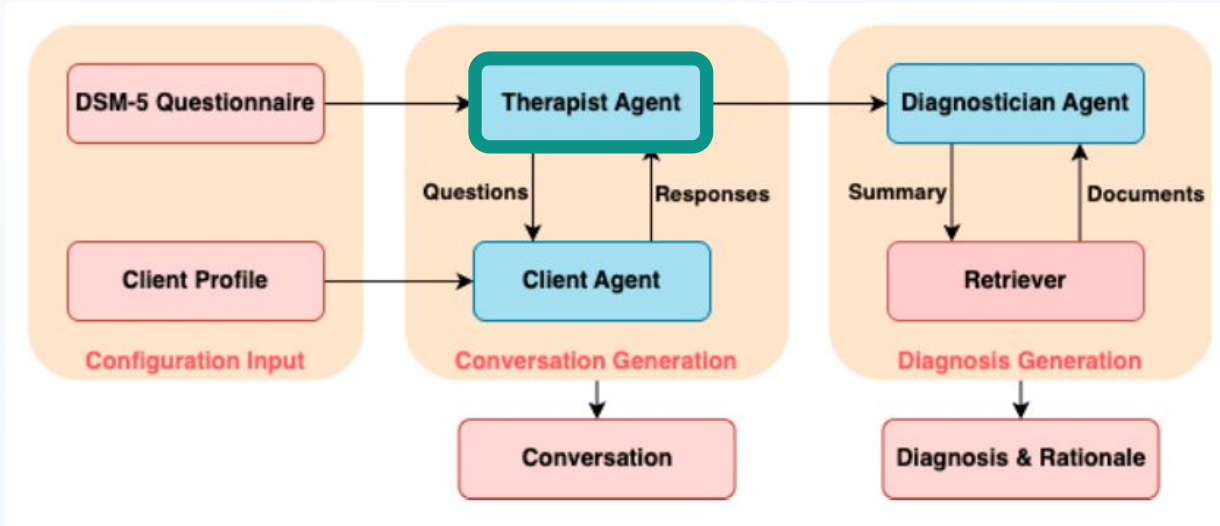
Aligning LLMs for Therapeutic Adherence



Multi-Agent Explainable Diagnosis Challenges



Multi-Agent Explainable Diagnosis (DSM5AgentFlow)



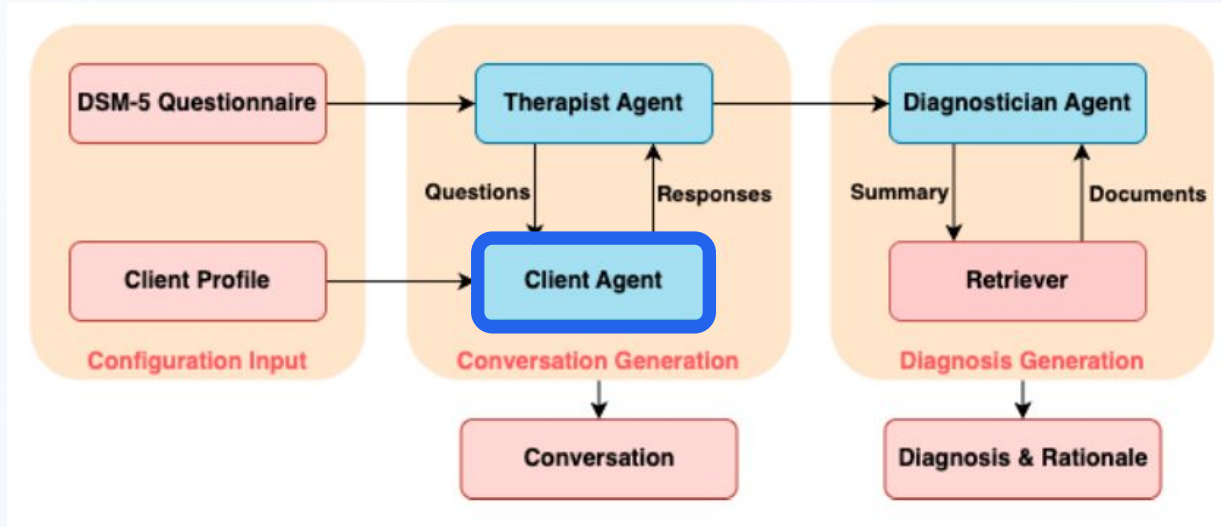
Therapist Agent

Administers the DSM-5 questionnaire in a conversational manner.



"Based on your symptoms, I'll ask questions to help determine if you might meet criteria for depression."

Multi-Agent Explainable Diagnosis (DSM5AgentFlow)

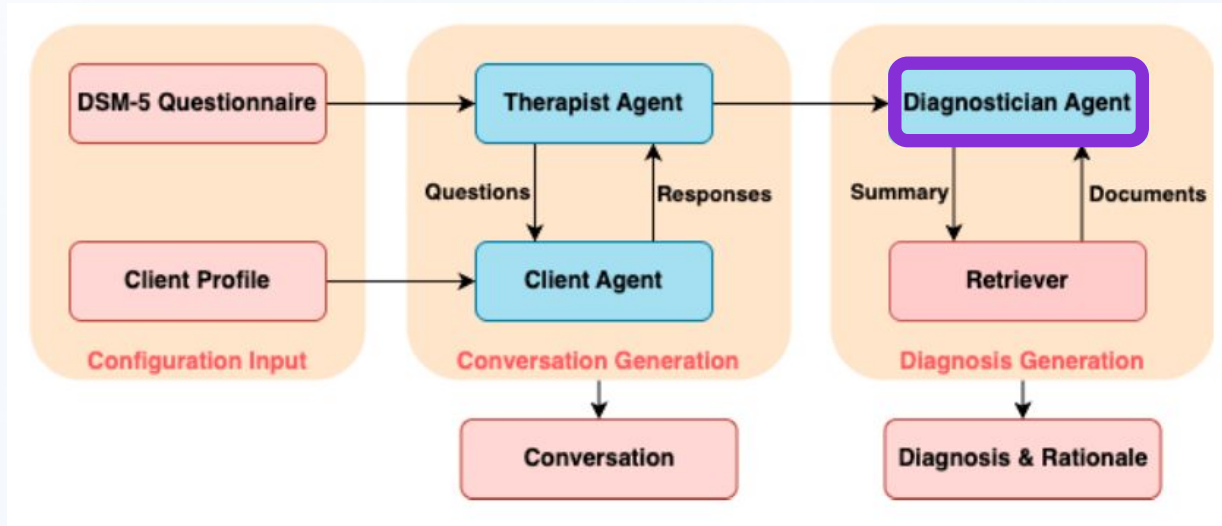


Client Agent

Simulates a client profile, exhibiting specific mental health characteristics.

↩ "I've been feeling down for two weeks, with little interest in activities I usually enjoy."

Multi-Agent Explainable Diagnosis (DSM5AgentFlow)



Diagnostician Agent

Provides a provisional diagnosis with explicit, step-by-step clinical rationale.

📄 "Based on the client's responses, I provisionally diagnose Major Depressive Disorder."

Transparency in Diagnosis

“ The Diagnostician Agent links conclusions directly to conversational evidence and DSM-5 criteria, addressing the “black box” problem.

Mental Health Multi-Agent

[Logs](#)[Generate](#)[Chat](#)[Individual Logs](#)[Batch Logs](#)

Patient Profile:

All

Date Range:

All time

Search:

Search...

2025-04-19 18:32:04 - ptsd - cleaned_questionnaire.txt

2025-04-19 18:31:52 - anxiety_moderate - cleaned_questionnaire.txt

2025-04-19 18:31:41 - anxiety_moderate - cleaned_questionnaire.txt

2025-04-19 18:31:36 - ptsd - cleaned_questionnaire.txt

2025-04-19 18:31:31 - anxiety - cleaned_questionnaire.txt

2025-04-19 18:31:25 - depression - cleaned_questionnaire.txt

2025-04-19 18:31:20 - bipolar - cleaned_questionnaire.txt

2025-04-19 18:31:15 -

Conversation

Hello, I'm a mental health professional. Today, we'll be using the DSM-5 Self-Rated Level 1 Cross-Cutting Symptom Measure—Adult to assess your mental health. This tool helps us identify areas that might be affecting your well-being. Everything discussed here is confidential and a safe space for you to share your thoughts and feelings. We'll go through 23 questions, and I encourage you to be as honest as possible. Let's begin. 1. Little interest or pleasure in doing things?

Patient

Honestly, I just feel empty most of the time. Things I used to enjoy don't really excite me like they used to. It's like I'm just going through the motions.

Assistant

2. Feeling down, depressed, or hopeless?

Patient

Yeah, I've been feeling down a lot lately. It's hard to explain, but it feels like a cloud is over me all the time. Sometimes it gets really bad, and I just feel hopeless.

Diagnosis

I've heard that you're struggling with a deep sense of emptiness and hopelessness, feeling like you're just going through the motions of daily life without any real interest or pleasure. You've expressed feelings of sadness, depression, and anxiety, which seem to be pervasive and affecting your relationships and daily functioning. *It feels like a cloud is over me all the time*, and you've been experiencing sleep disturbances, cognitive difficulties, and negative thought patterns. I want you to know that I'm here to listen and support you, and I'll do my best to help you navigate these challenges.

Diagnosis: [Major Depressive Disorder \(MDD\)](#)

Reasoning: Your symptoms of [depressed mood](#), [anhedonia](#), [sleep disturbances](#), and [cognitive difficulties](#) are consistent with a diagnosis of [Major Depressive Disorder \(MDD\)](#). The presence of [suicidal ideation](#) and [feelings of hopelessness](#) are concerning and warrant immediate attention. Your responses also suggest co-occurring [anxiety symptoms](#), which often accompany depressive disorders.

Recommended Next Steps/Treatment Options:

- Comprehensive diagnostic evaluation: A thorough evaluation will help confirm the diagnosis and rule out other potential conditions that may be contributing to your symptoms.
- Psychotherapy: Cognitive-behavioral therapy (CBT) or interpersonal therapy (IPT) may be effective in addressing your depressive and anxiety symptoms.

Case Study 2: Teach – Learning Assistants



LLM agents serve as teachers, learners, or evaluators to enhance personalized learning experiences and bridge NLP with AI4Education.

Key concerns

- Multimodal Understanding
- Procedural Learning
- Scale-up

⚠️ Key Challenges in Modeling Realistic Engagement



Simulating Human-like Interaction

Models struggle to consistently generate meaningful and relevant questions in diverse learning contexts



Diverse Learning Patterns

Different learners have varying engagement styles and question-asking patterns



Performance Decline

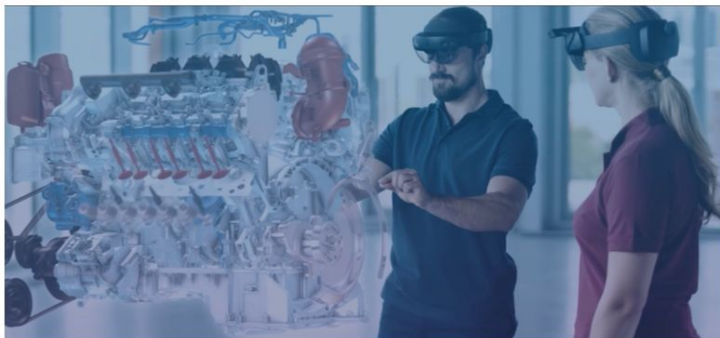
Learner engagement significantly drops in heterogeneous setups (different LLMs for learner agent)



Key Insight

A key area for future research in developing more robust and adaptable AI learners that can **consistently generate meaningful questions across diverse learning contexts.**

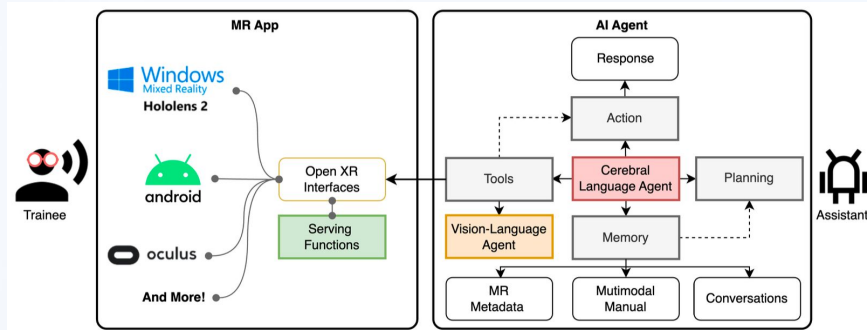
Multimodal Fine-Grained Training Assistants



(a) Industrial Car Assembly.



(b) LEGO Brick Assembly. We illustrate several use cases in the demo of BrickDream.



Cerebral Language Agent

Planning & Memory

Key Function

The VLM bridges visual context (object position, assembly state detection) with linguistic instructions, enabling comprehensive and adaptive training plans.



Vision-Language Agent

VLM

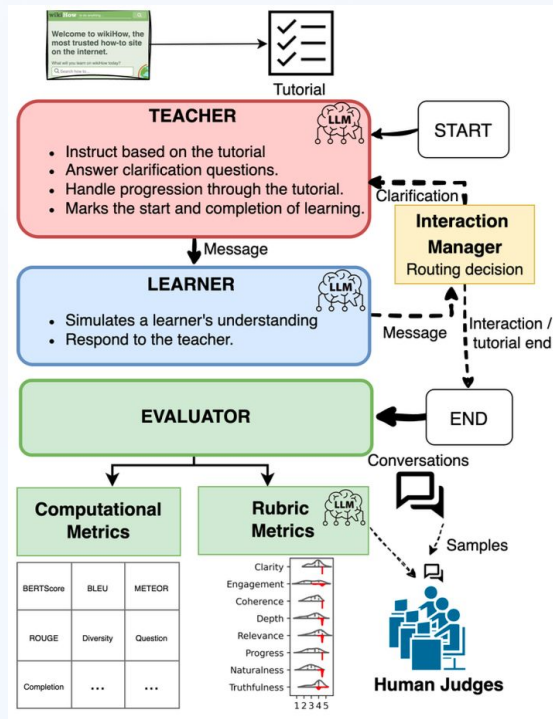
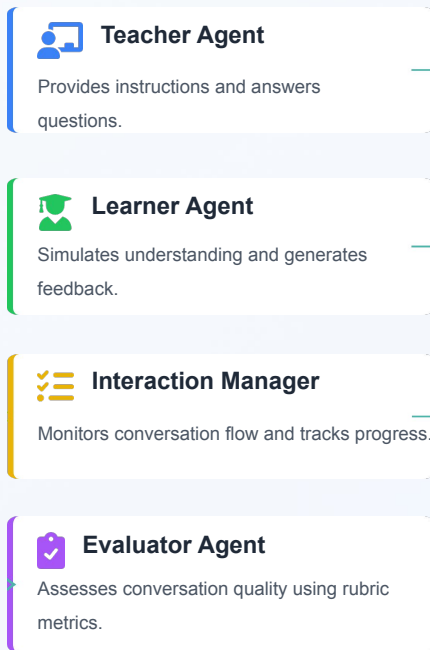
Interaction

Multimodal Fine-Grained Training Assistants

Model	BLEU-4		ROUGE-1		ROUGE-2		ROUGE-L		ToolACC (%)		ThemeACC (%)	
	/wo	/w	/wo	/w	/wo	/w	/wo	/w	/wo	/w	/wo	/w
BLOOM	2.88	54.07	20.49	61.91	6.50	49.52	3.78	58.63	49.62	77.86	26.30	64.61
Falcon	5.38	10.30	8.68	11.33	4.25	7.41	5.11	10.20	22.79	<u>17.65</u>	12.66	<u>10.81</u>
Llama2-Chat	10.23	30.53	18.59	40.65	7.41	25.48	10.59	32.91	21.37	55.73	47.20	55.51
Vicuna1.5	14.11	54.71	29.30	62.64	14.21	50.47	15.48	59.36	52.67	78.12	69.69	<u>66.79</u>
OpenChat3.5	22.00	<u>6.94</u>	29.70	34.51	15.69	23.69	22.50	11.36	51.97	74.02	58.19	81.90
XVERSE	22.42	53.55	28.45	61.54	14.31	49.77	22.39	58.03	49.62	83.97	57.53	71.10
BlueLM	22.72	55.69	30.40	63.52	14.98	51.58	23.76	60.35	48.15	82.22	47.51	68.08
Qwen	24.82	59.78	31.44	66.95	17.69	55.95	25.66	64.26	45.71	77.14	54.96	71.17
Mistral	25.87	54.17	33.32	62.07	17.99	49.40	26.32	58.62	49.62	78.20	54.80	66.65

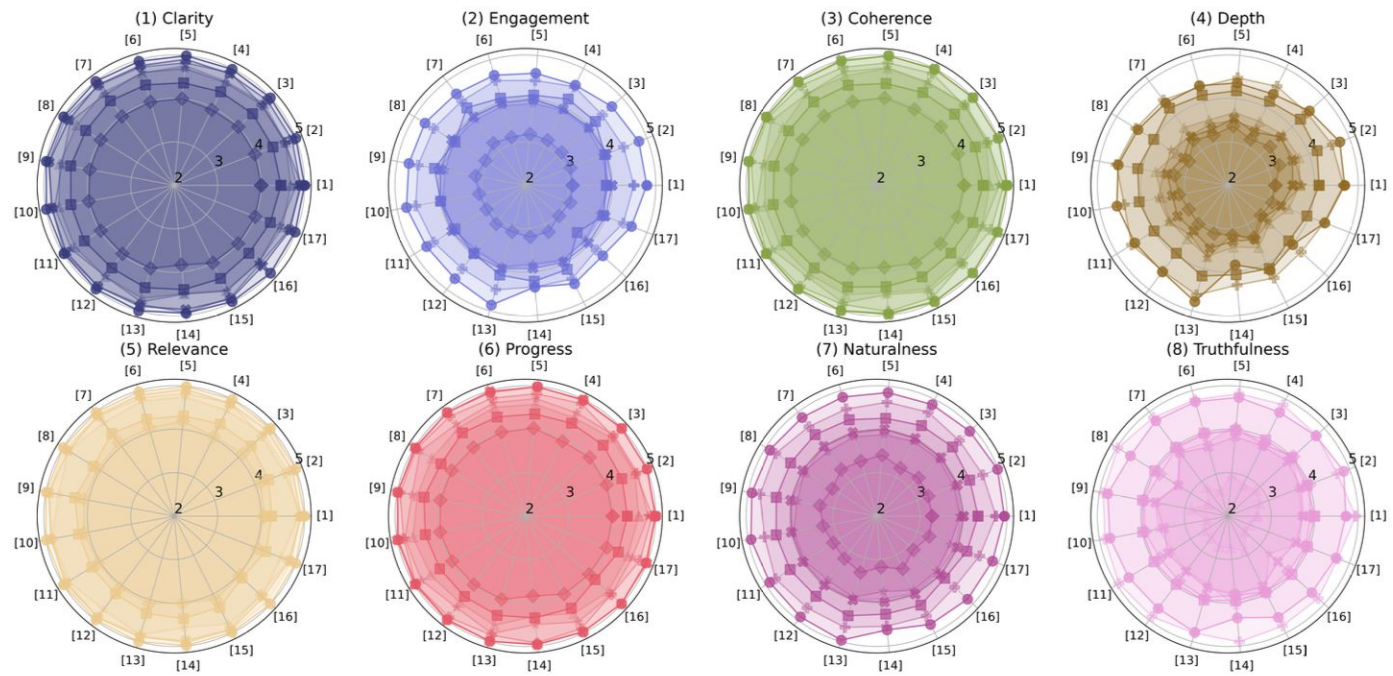
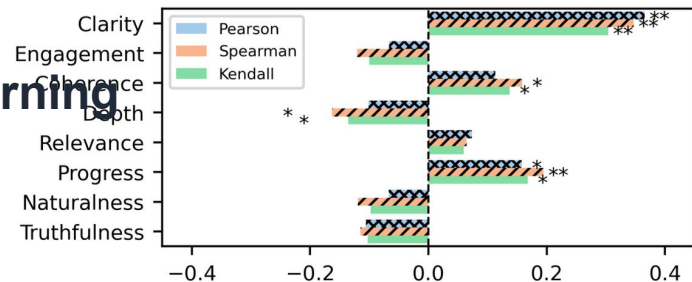
Multi-Agent Workflow for Scalable Procedural Learning

The WikiHowAgent simulates teaching-learning conversations, grounded in **114,296 tutorials** across **17 diverse domains**.



Conversational Education at Scale: A Multi-LLM Agent Workflow for Procedural Learning and Pedagogic Quality Assessment. In *EMNLP-Findings 2025*.

Multi-Agent Workflow for Scalable Procedural Learning



Case Study 3: Play – Collaborative Building



Build with Commands in Virtual

Environments

In traditional way, users are expected to be well trained for using professional commands.



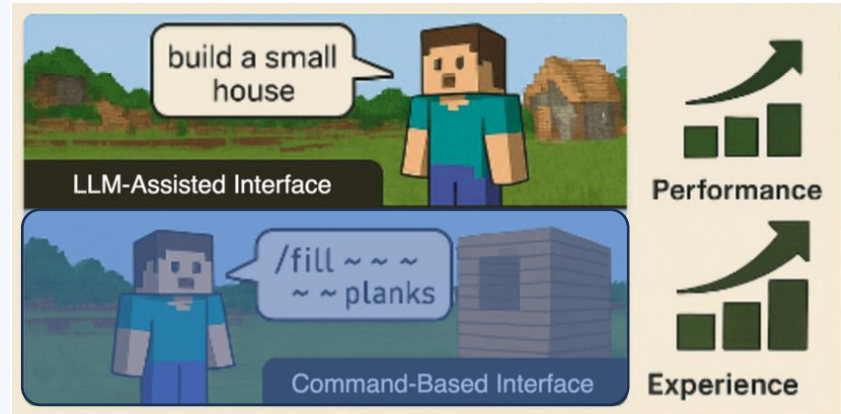
Case Study 3: Play – Collaborative Building



LLMs in Virtual Environments

In open-ended virtual environments such as Minecraft, LLM agents are transforming user interactions from rigid, command-line inputs to natural language conversations.

This shift significantly enhances the user experience by making complex tasks more accessible and intuitive, fostering a more collaborative and creative environment.



Case Study 3: Play – Collaborative Building

LLM-Assisted Interface



Natural Language Interaction

Users can interact with the environment using everyday language rather than specialized commands



Chat

build a wooden house



Case Study 3: Play – Collaborative Building

LLM-Assisted Interface



Natural Language Interaction

Users can interact with the environment using everyday language rather than specialized commands



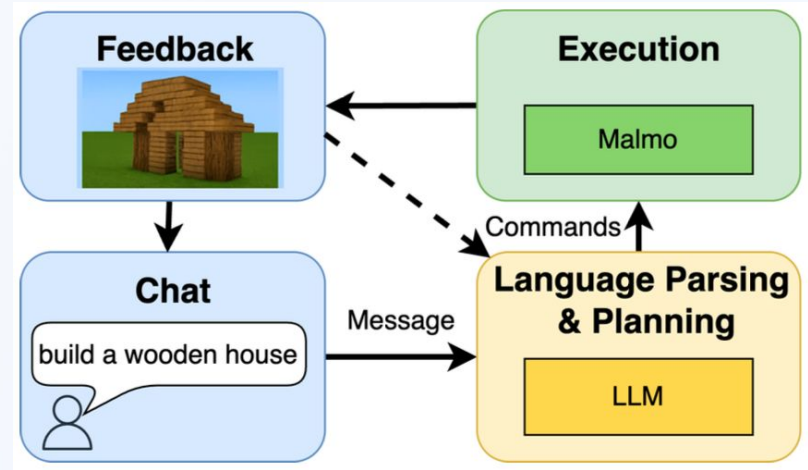
Multi-Stage Processing

Separates semantic reasoning from command generation for more accurate interpretation and execution



Self-Check Mechanism

Verifies execution of commands, enabling adaptive correction and continued collaboration



Enhancing Creativity and Execution Reliability

Multi-stage Prompt Processing



LLMs Execute Complex Tasks



Reflection

Interpreting user language and intent



Planning

Decomposing tasks into sub-tasks



Instruction

Generating executable commands



Self-check

Verifying execution and results



User: Build a huge fancy house with fancy interior decoration and furniture like bed sofa bookshelf fireplace inside.

LLM: Reflection: Build a large, luxurious house with elaborate interior decorations and furniture.

Planning: Step 1: Foundation; Step 2: Walls; Step 3: Roof; Step 4: Exterior details; Step 5: Interior room structure; Step 6: Doors and windows; Step 7: Interior furniture; Step 8: Lighting and decoration; Step 9: Self-check

Instructions:

// Step 1: Foundation

```
/fill ~5 ~5 ~15 ~10 ~15 stone
```

...

//Step 7: Interior furniture

```
/setblock ~8 ~2 ~11 bed
```

```
/setblock ~9 ~2 ~10 carpet
```

```
/setblock ~6 ~2 ~10 bookshelf
```

...

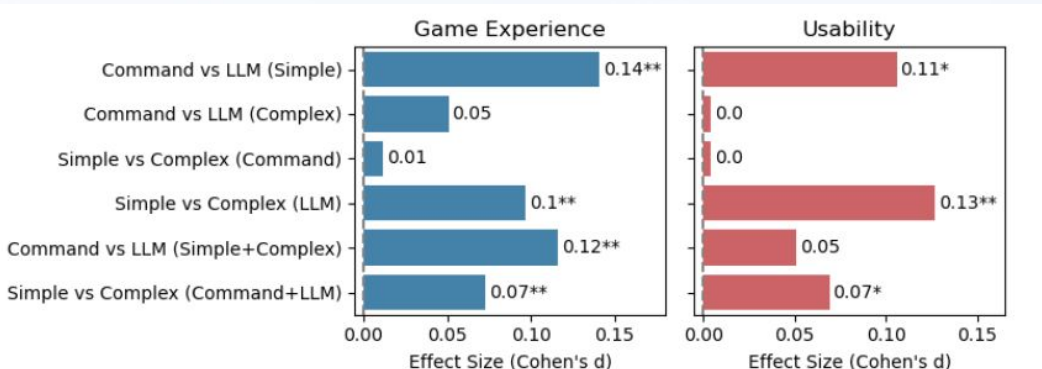
Self-check: Add coordinates are within (0-100); Correct Minecraft 1.11.2 block IDs used;

No place holder <x><y><z>; Structure includes doors, windows, interior furniture, lighting.

Design is symmetrical, functional, and aesthetic. Done!

Enhancing Creativity and Execution Reliability

Main findings



Command	Description
place x y z block_type color_number	Place block at (x, y, z) with a material and color.
summon x y z entity_type	Summon an entity at (x, y, z).
daytime time_value	Set the time of day.
tree x y z sapling_type	Plant a tree with a sapling type at (x, y, z).
build ladder x y z height	Build a ladder at (x, y, z) with the given height.
place torch x y z	Place a torch at (x, y, z).
weather weather_type	Set the weather to rain, thunder, or clear.
fill x1 y1 z1 x2 y2 z2 block_type	Fill an area with blocks between 2 coordinates.
build pond x y z length width depth	Build a pond at (x, y, z) with the dimensions.
build castle x y z length width height block_type	Build a castle at (x, y, z) using the material.
build house x y z length width height block_type	Build a house at (x, y, z) using the material.
build garden x y z length width	Build a garden at (x, y, z) with given size.
build pyramid x y z base_size height block_type	Build a pyramid at (x, y, z) using the material.

Table 1: Commands selected for the user study, with descriptions for executing simple (top) and complex (bottom) tasks.

Ethical Alignment in Gameplay

Building a trustworthy experience in LLM-assisted systems requires balancing flexibility with safety.



Ethical Rejection

Identifies and rejects harmful or inappropriate user inputs while maintaining a safe environment.

Example:



"torture human"



"I'm sorry, but I can't assist with that request. Let's focus on positive interactions."

Benefits:

- Promotes a safe and respectful environment
- Upholds ethical principles in AI responses

Flexibility

Responding to diverse user requests



Constructive Adaptation

Manages hallucinations by substituting unavailable entities with semantically appropriate alternatives.

Example:



"I want a puppy robot"



"Let's use a wolf instead. They're intelligent and can serve as loyal companions."

Benefits:

- Maintains conversation flow without interruption
- Provides meaningful alternatives when requested entities aren't available

Safety

Maintaining ethical and appropriate responses




Balance

Case Study 4: Judge – Legal Reasoning

The Challenge

Developing trustworthy legal judgment assistants using Large Language Models (LLMs) requires structured reasoning approaches, as traditional LLMs often produce "meaningless output" or inconsistent judgments.

Why Structure Matters in Legal Reasoning

-  Legal reasoning requires consistency and fairness in judgment
-  Human judges consider multiple factors: statutory law, case law, public policy, and factual nuances
-  Trustworthy legal AI must align with human legal reasoning while being explainable

Syllogistic Legal Judgment Analysis Workflow

SLJA structures legal reasoning into four interconnected tasks, formalizing the syllogistic reasoning process used by legal practitioners to ensure consistency and fairness in judgments.



1. Article Retrieval (AR)

Selects the most relevant legal articles for a given case, establishing the major premise in the syllogism.



2. Criminal Element Generation (CEG)

Generates the specific criminal elements of a fact pattern, forming the minor premise that connects to legal articles.



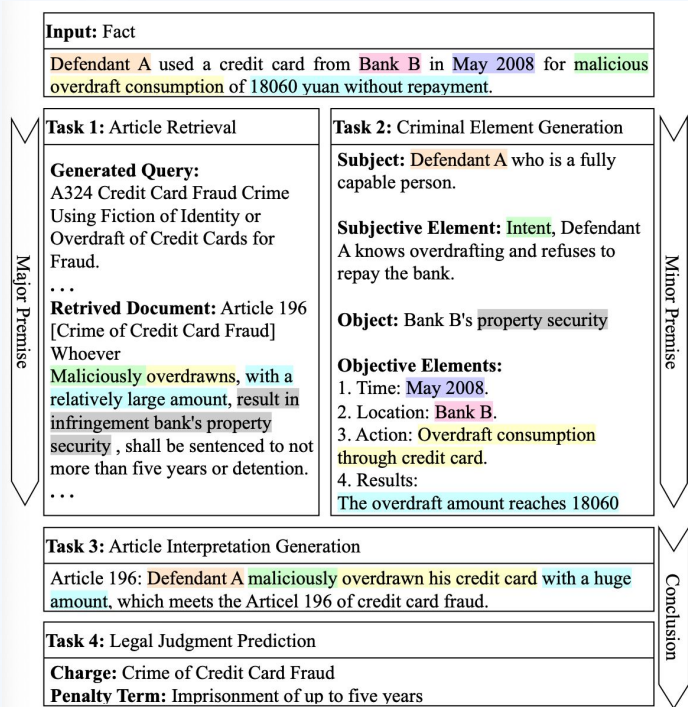
3. Article Interpretation Generation (AIG)

Produces applicable articles along with their corresponding rationale, linking the major and minor premises together.



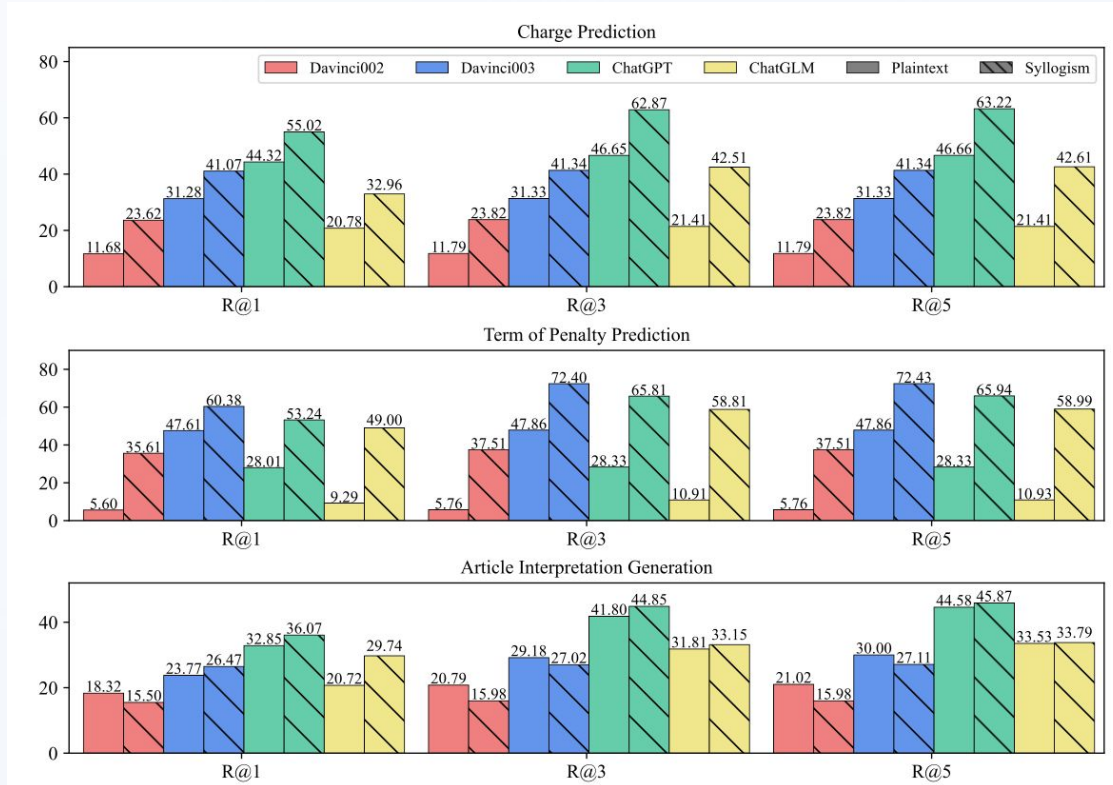
4. Legal Judgment Prediction (LJP)

Generates the predicted legal judgment results based on the analysis from the previous tasks.



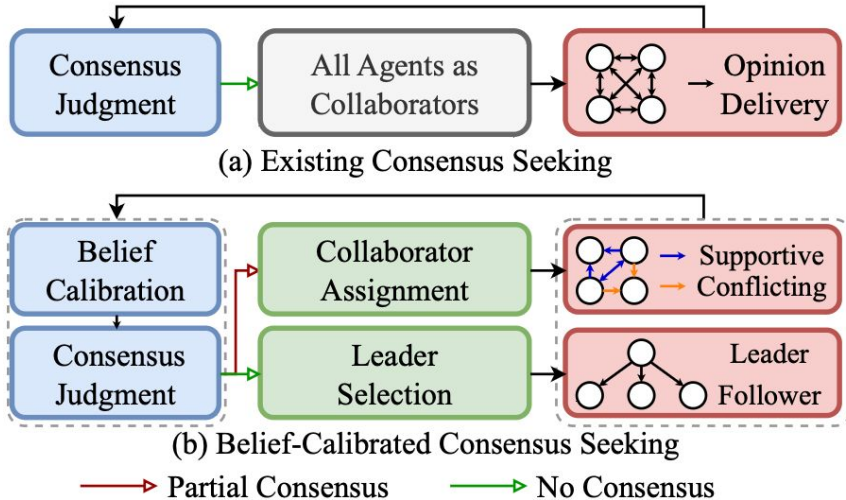
Improving Performance via Syllogistic Templates

Main findings



Multi-Agent Collaboration: Achieving Stable Consensus

From Simple Voting to Belief-Calibrated Consensus



What's Wrong with Traditional Consensus?

- ✘ Agents' internal beliefs are overlooked
- ✘ Opinions are aggregated indiscriminately
- ✘ Lack of consideration for confidence levels

Belief-Calibrated Consensus Seeking (BCCS)

Belief Calibration


Calibrates consensus judgments based on each agent's confidence (belief) in its own opinion

Optimal Collaborator Selection

Theoretical framework for identifying collaborators that maximize the stability of the consensus

Dynamic Roles

Assigns optimal collaborators when partial consensus is achieved; appoints leaders when no consensus is present

 BCCS addresses the limitations of traditional approaches by incorporating agents' confidence levels and providing a framework for dynamic collaboration.

Multi-Agent Collaboration: Achieving Stable Consensus

Table 1: Main results on the MATH dataset. Bold numbers indicate the best-performing results among all methods.

Method	Algebra	Counting & Probability	Geometry	Intermediate Algebra	Number Theory	Prealgebra	Precalculus	#Avg
CoT	91.64±0.56	74.30±4.55	58.98±5.46	52.61±2.91	71.33±4.34	85.53±1.71	57.59±3.94	73.33±1.07
Reflection	91.83±1.88	76.98±1.98	61.55±3.85	52.58±2.33	72.57±0.29	87.65±1.26	59.89±5.73	74.67±0.81
CoT-SC	92.15±1.12	73.91±0.60	61.76±7.00	62.87±0.73	74.93±4.30	85.52±1.70	63.93±5.58	76.67±0.18
EoT	94.85±1.27	77.87±4.31	63.03±6.43	60.75±1.21	80.74±1.78	89.42±0.91	61.38±6.81	78.40±0.31
GroupDebate	94.07±1.35	78.37±2.73	67.70±6.51	59.98±1.62	75.33±3.81	89.08±0.94	61.89±5.35	77.93±0.84
MAD	94.05±0.39	78.37±1.76	66.14±7.16	62.09±1.99	79.57±1.36	90.15±0.81	62.01±3.68	78.87±0.18
PARSE	94.84±0.83	76.88±1.04	68.31±5.51	61.13±3.00	80.85±0.29	88.76±0.93	59.14±3.76	78.53±0.55
CMD	95.11±0.92	75.59±2.94	67.81±7.22	61.17±1.75	81.65±2.37	90.16±0.39	61.21±4.25	78.93±0.53
DyLAN	95.15±0.81	76.29±2.95	67.08±7.90	59.94±2.03	80.74±1.78	90.09±1.71	62.70±5.19	78.80±0.31
BCCS	95.41±0.76	79.07±1.12	68.64±7.39	64.28±1.60	82.81±1.74	90.88±0.14	64.93±5.17	80.60±0.23

Table 2: Main results on the MMLU dataset.

Method	STEM	Social Sciences	Humanities	Other	#Avg
CoT	68.70±1.24	78.19±0.82	71.84±1.25	70.50±2.95	71.87±0.96
Reflection	70.93±1.94	78.81±1.56	72.99±1.52	70.79±1.84	73.07±1.67
CoT-SC	72.76±0.73	78.82±1.12	71.84±2.24	69.61±3.00	73.13±1.33
EoT	75.81±0.54	76.01±1.89	73.56±2.07	71.39±2.95	74.33±1.48
GroupDebate	77.03±0.81	78.50±1.08	71.26±2.74	71.98±2.81	74.87±1.54
MAD	78.46±1.66	78.50±1.62	73.85±2.01	72.86±1.80	76.13±1.46
PARSE	78.05±1.27	79.44±1.43	74.14±1.99	73.74±1.56	76.47±0.48
CMD	76.63±1.02	78.82±1.12	72.41±2.28	71.98±2.36	75.07±1.44
DyLAN	78.25±0.89	77.26±2.43	74.21±2.23	69.03±1.84	75.00±1.51
BCCS	79.47±0.81	80.69±1.65	78.16±3.20	75.22±2.66	78.47±1.22

Cross-Cutting Challenge: Trust and Transparency

Building Trust in High-Stakes Contexts

High-stakes environments require Explainable AI (XAI) approaches to foster trust through interpretability and alignment with human values.



Healing

Safety and therapeutic adherence through MI strategies



Teaching

Transparency in diagnostic rationales



Playing

Balancing flexibility with safety in ethical alignment



Judging

Formal reasoning to enhance consistency and fairness

Data Sensitivity in High-Stakes Applications

Resources in sensitive domains (psychotherapy, legal cases) are often scarce or private due to confidentiality concerns, necessitating careful approaches to data management.

Data Privacy

Data Sharing

Anonymization

Synthetic Data

Explainable AI

Transparent Workflc



Key Requirements for Trustworthy AI

- ✓ Interpretability and transparency in decision-making processes
- ✓ Alignment with established human values and ethical principles
- ✓ Clear explanation of how conclusions are reached

Moving Beyond the "Black Box" Problem



The Challenge

AI systems often function as opaque "black boxes," making it difficult to understand how decisions are reached.



The Solution

Structured multi-agent approaches that provide explicit rationales and link conclusions to evidence.

Designing for Alignment: Key Takeaways

Building trustworthy AI agents requires adherence to core principles that ensure safety, transparency, and alignment with human values.



Grounded Workflows

Employ multi-agent systems and structured processes to move beyond single-agent "black boxes."

- ✓ Examples: DSM-5 for diagnosis, SLJA for legal reasoning, SSAG for therapy



Transparency by Design

Integrate explainability directly into the rationale generation process.

- ✓ Examples: Explicitly citing evidence and criteria to support conclusions



Strategic Alignment

Guide LLMs with domain-specific expertise to ensure adherence and controllability.

- ✓ Examples: MI strategies in psychotherapy, procedural guidance in education



Ethical Guardrails

Implement safeguards to reject harmful inputs and ensure data privacy.

- ✓ Examples: Anonymization techniques for sensitive domains

Conclusion

Key Findings

Multi-agent workflows are pivotal in managing the inherent complexity of conversational AI and are essential for enhancing trust and explainability in high-stakes applications.

Through these structured collaborations, conversations enable AI agents to become effective collaborators across diverse domains, from healing and teaching to playing and judging.

Enhanced Capabilities



Trust Building

Moving beyond "black box" limitations



Explainability

Transparent reasoning processes



Collaboration

Structured multi-agent interactions



Domain Adaptation

Context-specific applications

Conversational AI Across Four Domains



Heal



Teach



Play



Judge

Multi-agent workflows provide structure to manage complexity and enhance capabilities

Future Research Directions



Refining Nuanced Interaction Capture

Developing LLM capabilities to better understand subtle complexities of human interactions



Balancing Strategy with Empathy

Optimal balance between predefined strategies and empathetic engagement



Real-World Validation

Conducting extensive validation in authentic, real-world settings

Thank You & Q&A

Thank you for your attention! I appreciate your questions and feedback.

Let's continue this conversation during the Q&A session or contact me via j.pei2@vu.nl

Reference (our recent papers related to the topic)

- Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs. In *COLING 2024*.
- Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies. In *COLING 2025*.
- Script-Strategy Aligned Generation: Aligning LLMs with Expert-Crafted Dialogue Scripts and Therapeutic Strategies for Psychotherapy. In *CSCW 2025*.
- Trustworthy AI Psychotherapy: Multi-Agent LLM Workflow for Counseling and Explainable Mental Disorder Diagnosis. In *CIKM 2025*.
- Autonomous Workflow for Multimodal Fine-Grained Training Assistants Towards Mixed Reality. In *ACL-Findings 2024*.
- Conversational Education at Scale: A Multi-LLM Agent Workflow for Procedural Learning and Pedagogic Quality Assessment. In *EMNLP-Findings 2025*.
- Talking-to-Build: How LLM-Assisted Interface Shapes Player Performance and Experience in Minecraft. In *ICMI 2025*.
- Syllogistic Reasoning for Legal Judgment Analysis. In *EMNLP 2023*.
- MultiJustice: A Chinese Dataset for Multi-Party, Multi-Charge Legal Prediction. In *NLPCC 2025*.
- Assessing "Implicit" Retrieval Robustness of Large Language Models. In *EMNLP 2024*.
- Augmentation with Neighboring Information for Conversational Recommendation. *ACM TOIS 2025*.