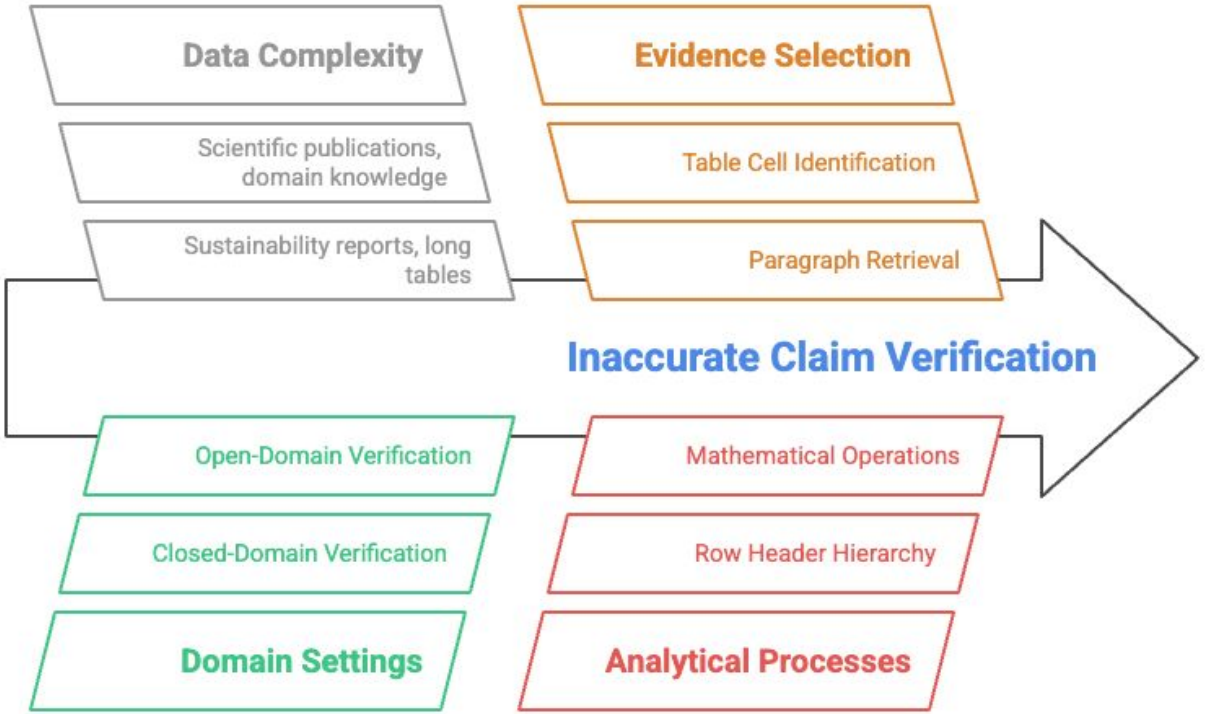
A Multi-Agent Approach for Claim Verification from Tabular Data Documents

Claim Verification Approaches

	Existing Research	Proposed Multi-Agentic Framework
Data Source	primarily on Wikipedia tables	Scientific publications, financial reports, sustainability reports
Methodology	Pretraining and fine-tuning	Prompt-based setup
Explanation	Table cell identification, limited generalizability	Step by step detailed explanation

Challenges in Claim Verification from Tabular Data

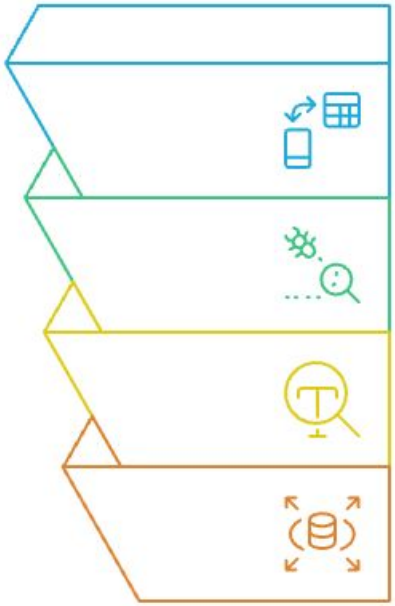


Multi-Agentic Claim Verification Process

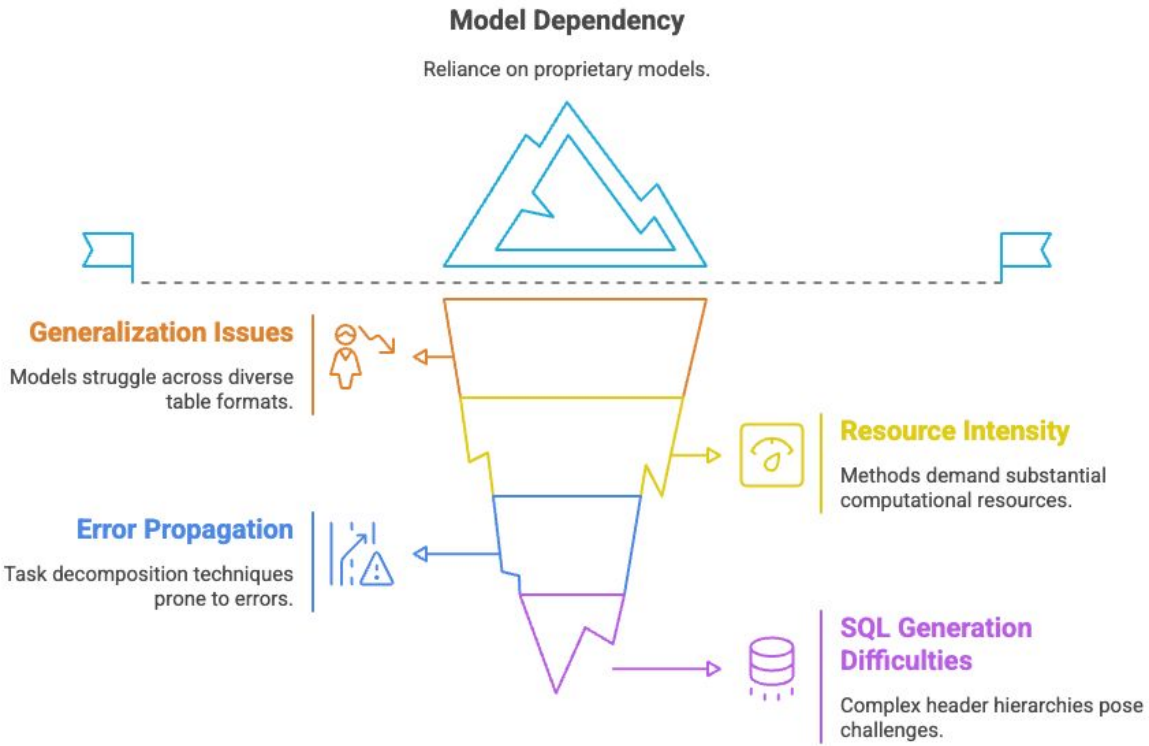
1. **Data Sufficiency Check**
   Determining if the table is sufficient for claimverification
2. **Reasoning Strategy Planning**
   Devising a strategy to verify the claim
3. **Verification Execution**
   Executing the strategy

Existing Works on Tabular Data Reasoning

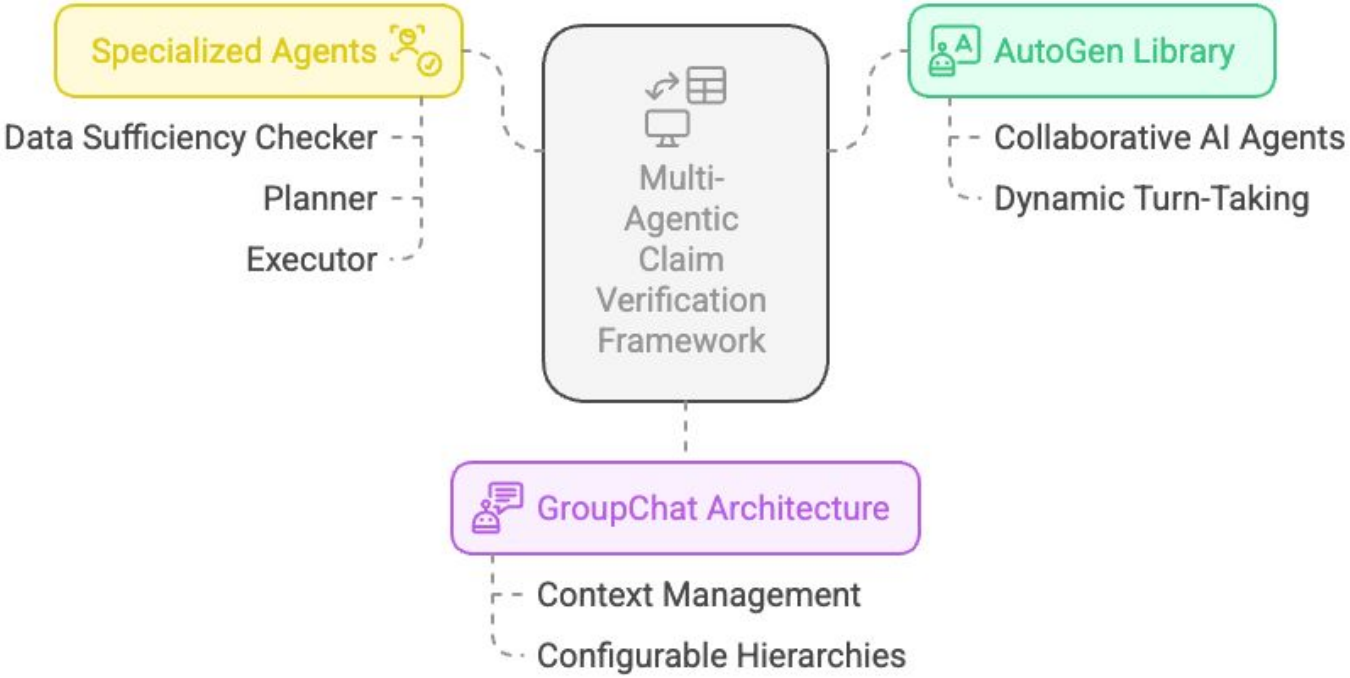
1. **Pre-training Models**
   Developing models to understand tabular data
2. **Fine-tuning Approaches**
   Refining models for specific tasks
3. **Tool-Augmented Reasoning**
   Enhancing models with specialized tools
4. **Open-Domain Reasoning**
   Applying models to diverse data sources



Challenges in Tabular Claim Verification

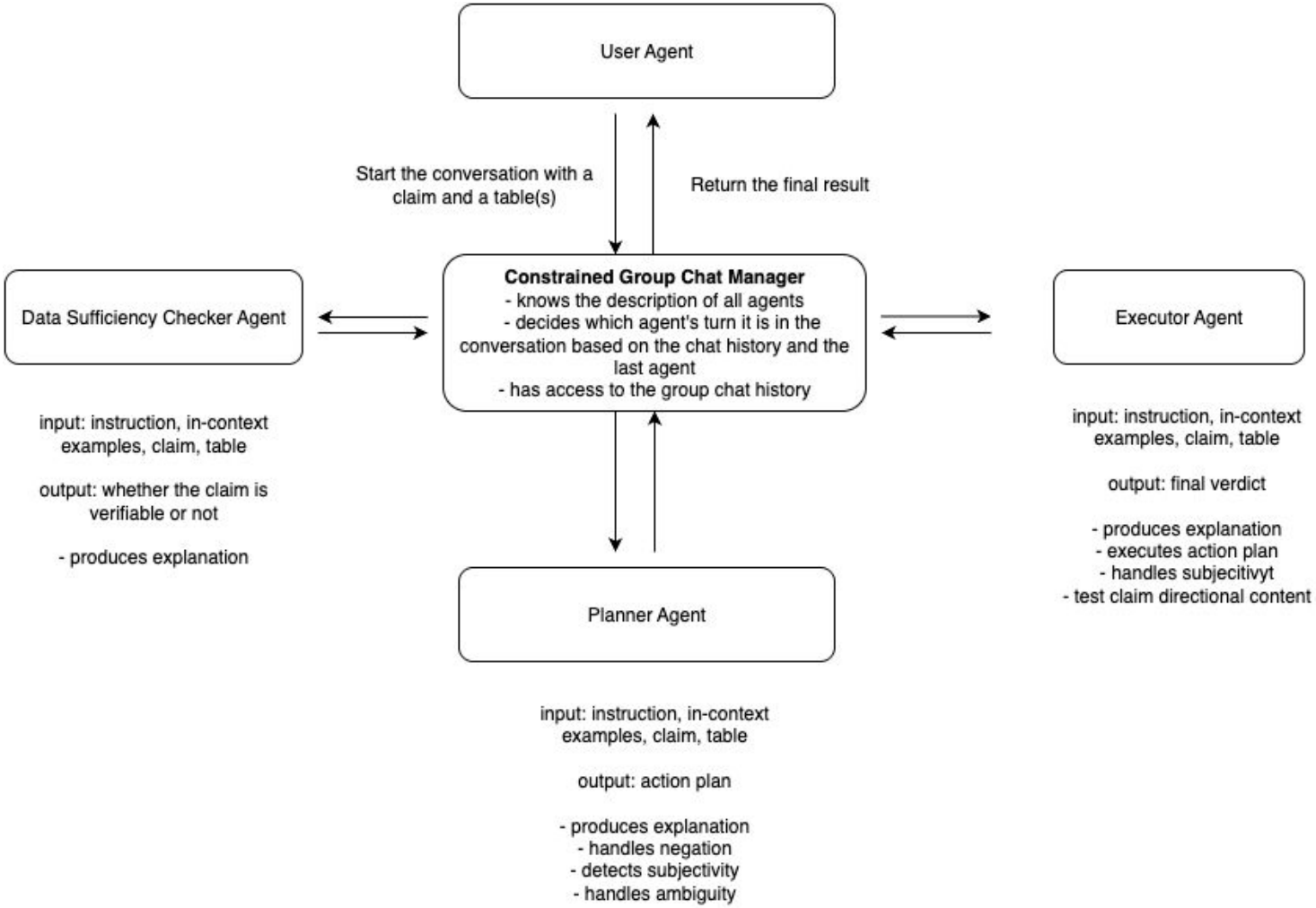


Multi-Agentic Claim Verification Framework (MACE)



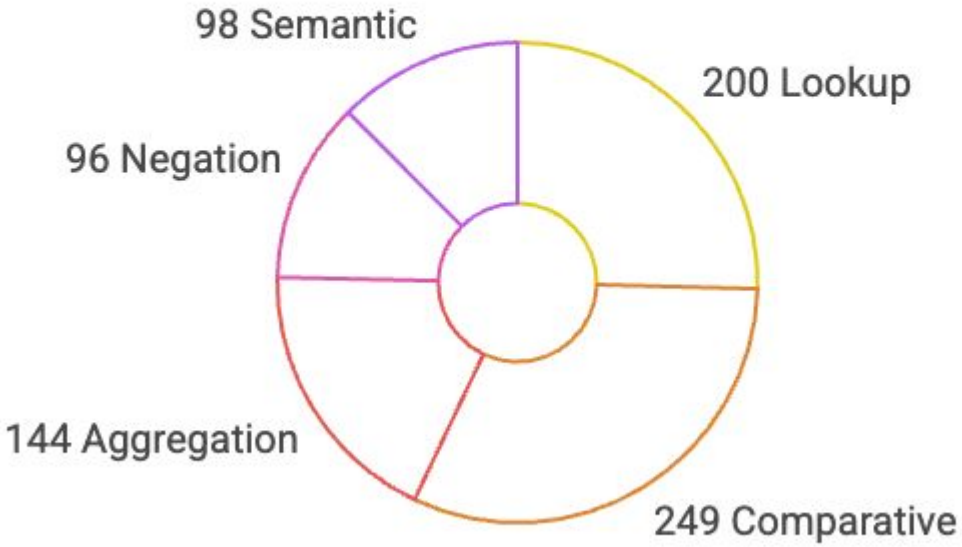
Role of Agents for Claim Verification

	Data Sufficiency Checker	Planner	Executor
Purpose	Evaluates data sufficiency	Devises verification plan	Implements action plan
Input	Table, caption, claim, column headers	Table, caption, claim, column headers	Action plan, Table, caption, claim, column headers
Process	Checks data relevance	determines the relevant evidence, interprets terms, proposes thresholds	Executes steps, cross-checks values
Output	Verdict (sufficient/insufficient)	Step-by-step action plan	Claim veracity and explanation
Key Feature	Early error mitigation	Structured reasoning process	Accurate and explainable outcomes

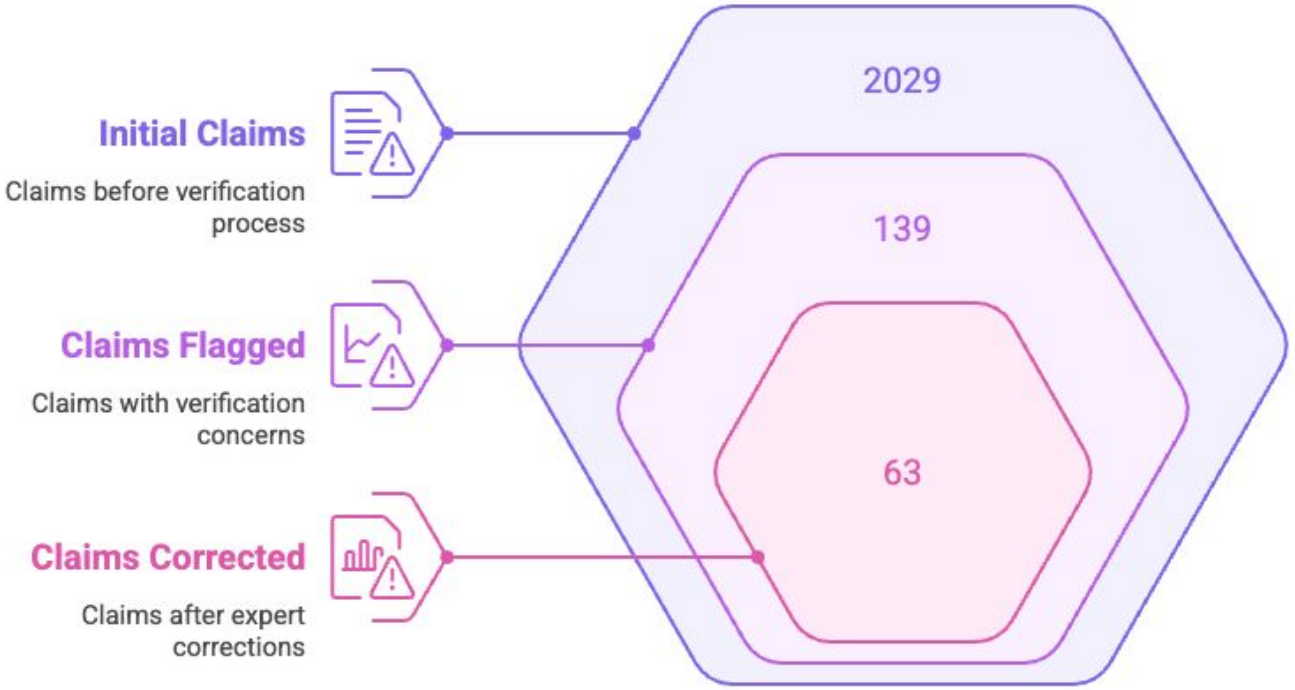

User Agent
Start the conversation with a claim and a table(s)
Return the final result
Constrained Group Chat Manager
- knows the description of all agents
- decides which agent's turn it is in the conversation based on the chat history and the last agent
- has access to the group chat history
Data Sufficiency Checker Agent
input: instruction, in-context examples, claim, table
output: whether the claim is verifiable or not
- produces explanation
Executor Agent
input: instruction, in-context examples, claim, table
output: final verdict
- produces explanation
- executes action plan
- handles subjecitivyt
- test claim directional content
Planner Agent
input: instruction, in-context examples, claim, table
output: action plan
- produces explanation
- handles negation
- detects subjectivity
- handles ambiguity

MineTabFact Dataset Creation and Verification

Distribution of "Supporting"Claim Types in MineTabFact Dataset (claims)



MineTabFact Claims Through Verification Stages



Dataset Comparison

	SCITAB	SEM-TAB-FACTS	FinDVer	MineTabFact
Description	Curated, scientific publications	Scientific tables	Open-domain, financial documents	Sustainability Reports
Claims	1,224	653	2,400	2029
Verification	3-way	3-way	2-way	3-way
Table Count	213	2,961	600 documents, 41.4 tables average	97
Difficulty	Challenges SOTA models	Requires complex reasoning	Challenges even GPT-4o	-

Method	Macro F1
PASTA $^{+}$	32.62
ProTrix $^{++}$	42.76
TART (GPT-4)*	63.6
Alpaca-7B†	28.95
Vicuna (13B)†	35.16
LLaMA-13B†	32.66
InstructGPT†	41.58
InstructGPT + COT†	42.6
GPT-4†	64.80
GPT-4 + COT†	63.21
MACE (Llama-3.3-70B-Instruct)	47.1
MACE (Typhoon2-70B-Instruct)	54.2
MACE (Qwen2.5-72B-Instruct)	58
MACE (Deepseek-Chat)	**70.8**

Table 3: Performance on Scitab (Macro F1). Models marked with †, *, and $^{++}$ are sourced from Lu et al. (2023), Lu et al. (2024) and Wu and Feng (2024) respectively; Model marked with $^{+}$ is the result from executing Gu et al. (2022a) on the Scitab dataset.

Method	F1 Micro
sattiy†	77.32
RyanStark†	81.51
THiFly-Queen†	83.76
King001†	84.48
Volta*	73.87
Tapas*	75.33
Tapex*	75.47
LKA*	78.54
DeBERTaV3*	78.92
PASTA*	84.1
MACE	**90.7**

Table 4: Performance on SEM-TAB-FACTS (F1 Micro score). Models marked with † are the best performing according to the leader-board (Wang et al., 2021); Those with * are sourced from Gu et al. (2022a).

Method	Testmini (Acc)	Test (Acc)
DeepSeek-V2-Lite	60.1	58.3
Qwen2.5	72.4	70.2
Llama-3.1 70B	75	74.5
Qwen2.5 72B	75.7	74.9
Mistral-Large 123B	74.8	75.8
Claude-3.5-Sonnet	73.1	70.4
Gemini-1.5-Pro	71.4	73.2
GPT-4o	75.3	76.2
MACE	**77.1**	**77.7**

Table 5: Accuracy on Findver (Testmini and Test). All baseline models are sourced from Zhao et al. (2024b).

Method	COT (F1 M)	w/o COT (F1 M)
glm-4-9b-chat	75.3	24.1
Llama-3.2-3B	54.9	17.5
mathstral-7B	64.4	24.1
Llama-3.1-8B	64.2	15.1
Mistral-7B	66	25.7
Phi-3.5-mini	68.9	25.9
Qwen2.5-7B	77.7	22.5
Gemini-1.5-pro	87.9	39
GPT-4o	92	39.9
Claude-3-5-sonnet	**93**	41.9
MACE	**93**	

Table 6: Performance comparison of various models on the MineTabFact dataset, showing F1 Macro scores with (COT) and without (w/o COT) Chain of Thought reasoning. All baseline models are adopted using the implementation of Zhao et al. (2024b).

Thanks