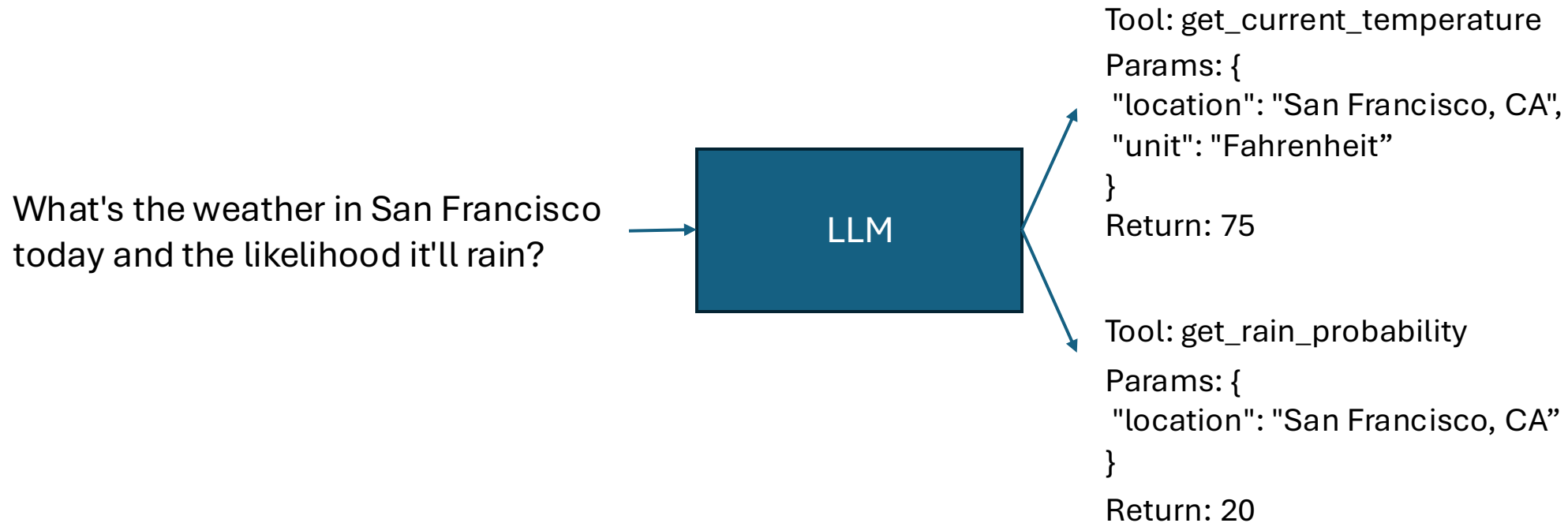


ToolGen: Unified Tool Retrieval and Calling via Generation

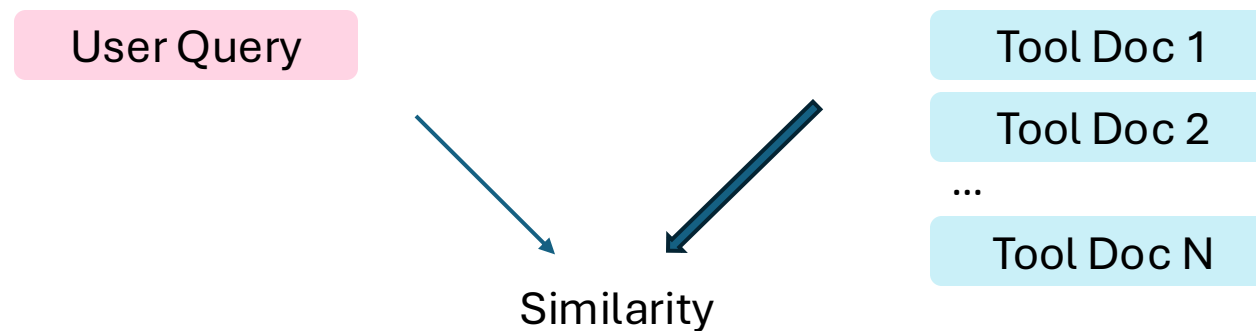
LLM Agent for Tool Learning

- Equip LLMs with tools to complete tasks automatically



Tool Retrieval

- A very large set of tools: up to 50k
- Retrieval based methods:
 - Discrete Retrieval: TF-IDF, BM25
 - Dense Retrieval: OpenAI Ada Embedding, Contrastive Learning BERT



Tool Learning: four-stage paradigm

- Planning: Decompose a user query into multiple sub-objectives
- **Tool Selection**: Select the best tool based on planning

Sub-objective 1: What is the current price of gold per ounce in USD?

Candidate Tools:

1. Metals Prices Rates API: The latest API endpoint will return real-time exchange rate data updated every 60 seconds.
2. Medium: Get official news from Medium.
3. Cryptocurrency Markets: Recently published cryptocurrencies videos.

Output: 1. Metals Prices Rates API: The latest API endpoint will return real-time exchange rate data updated every 60 seconds.

- Tool Calling: Fill in all parameters of the tool and execute the tool
- Response Generation: synthesize information and integrate knowledge to construct comprehensive responses

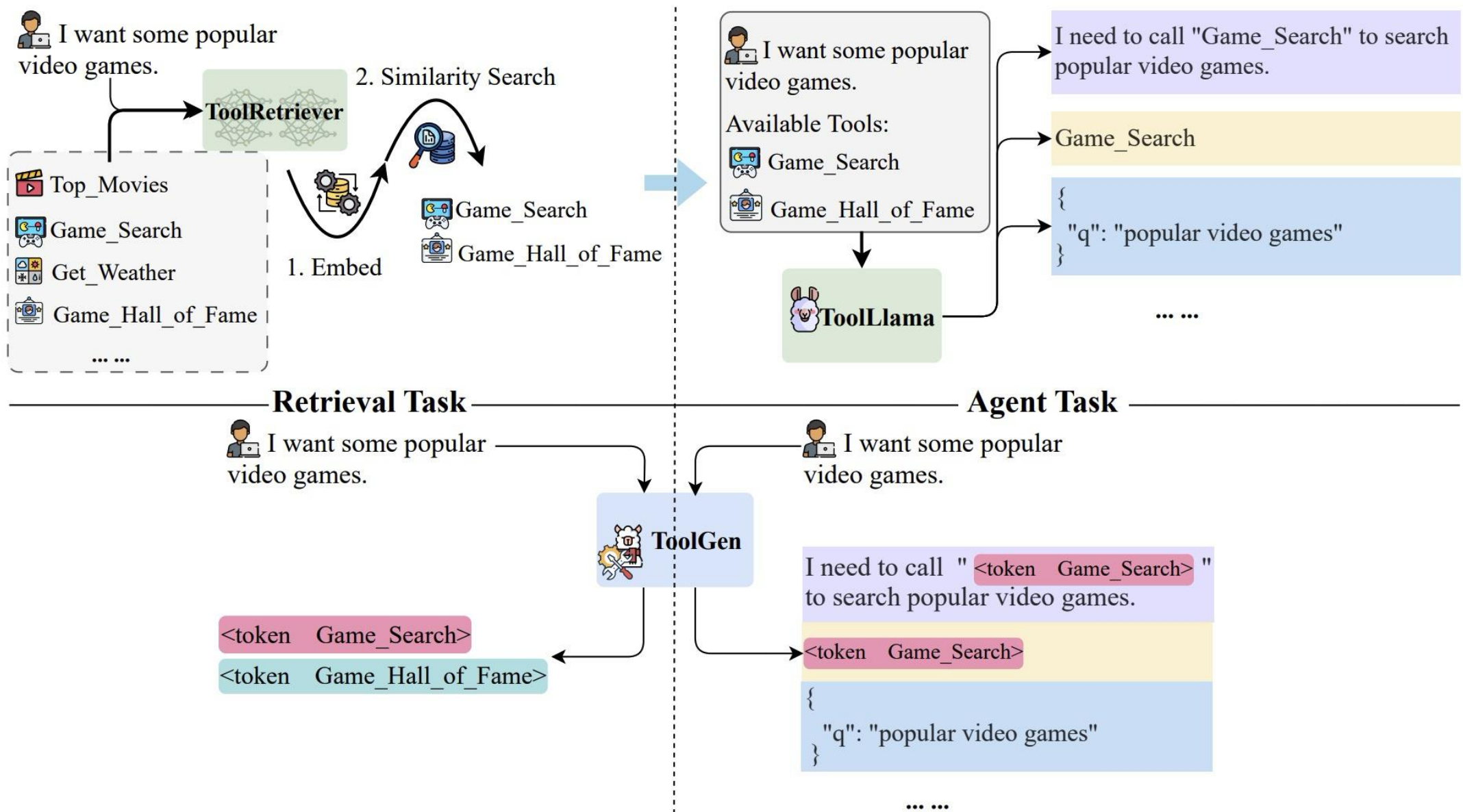
Tool Selection

- Some problems:
 - Rely on unsupervised or small models: (e.g. BM25 and BERT)
 - Can not capture full semantics/functionality in user queries and tool documents
 - Tool selection and calling are separated

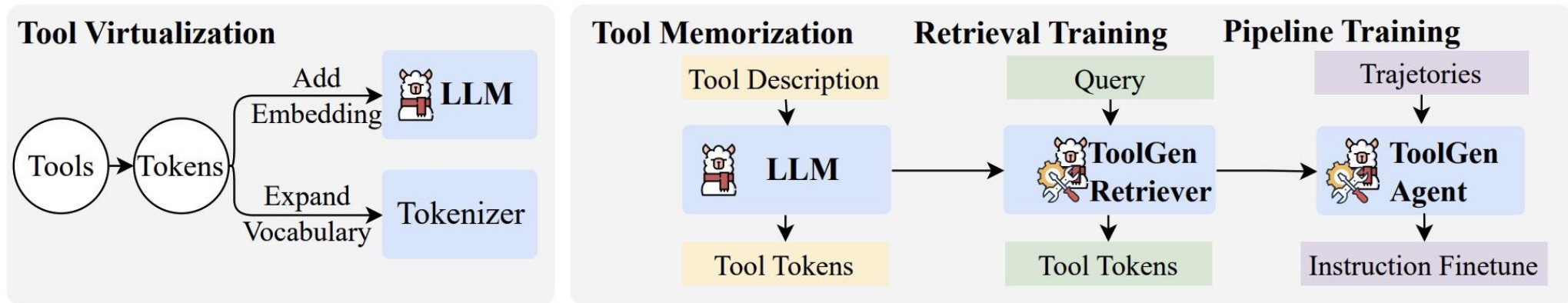
ToolGen: Unified Tool Retrieval and Calling

- Transform tool retrieval and execution into unified generation
 - Represent each tool as a unique token

Comparison: Traditional vs ToolGen

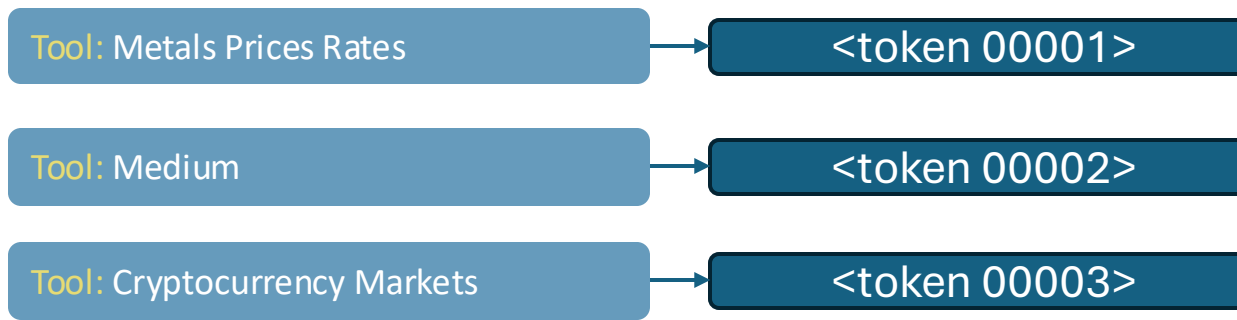


ToolGen: Unified Tool Retrieval and Calling



ToolGen: Tool Virtualization

- Tool virtualization: Map each tool into token for generation
 - Represent each tool with a single token



- Add tool tokens to vocabulary, expand embeddings

ToolGen: Tool Memorization

- Fine-tuning LLMs to memorize all tools
 - Input: Tool description
 - Output: Tool tokens

Name: QRCheck, Description: Check the quality of any QRCode

LLM

0 0 1 2 3

Numeric

QRCheck

Semantic

0 3 2 5 4

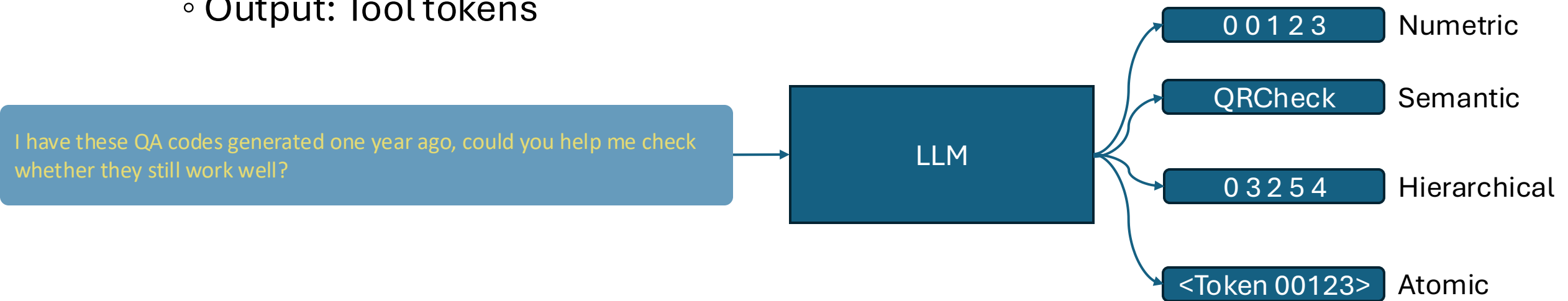
Hierarchical

<Token 00123>

Atomic

ToolGen: Retrieval Training

- Fine-tuning LLMs to generate tools tokens based on queries
 - Input: Queries
 - Output: Tool tokens



ToolGen: End-to-End Agent-Tuning

- Fine-tuning LLMs with full trajectories

User: I have these QR codes generated one year ago, could you help me check whether they still work well? The file is data/qr_code.png

ToolGen: I need to call <token 00123> to check quality of a QR code.

User: Generation the action.

ToolGen: <token 00123>

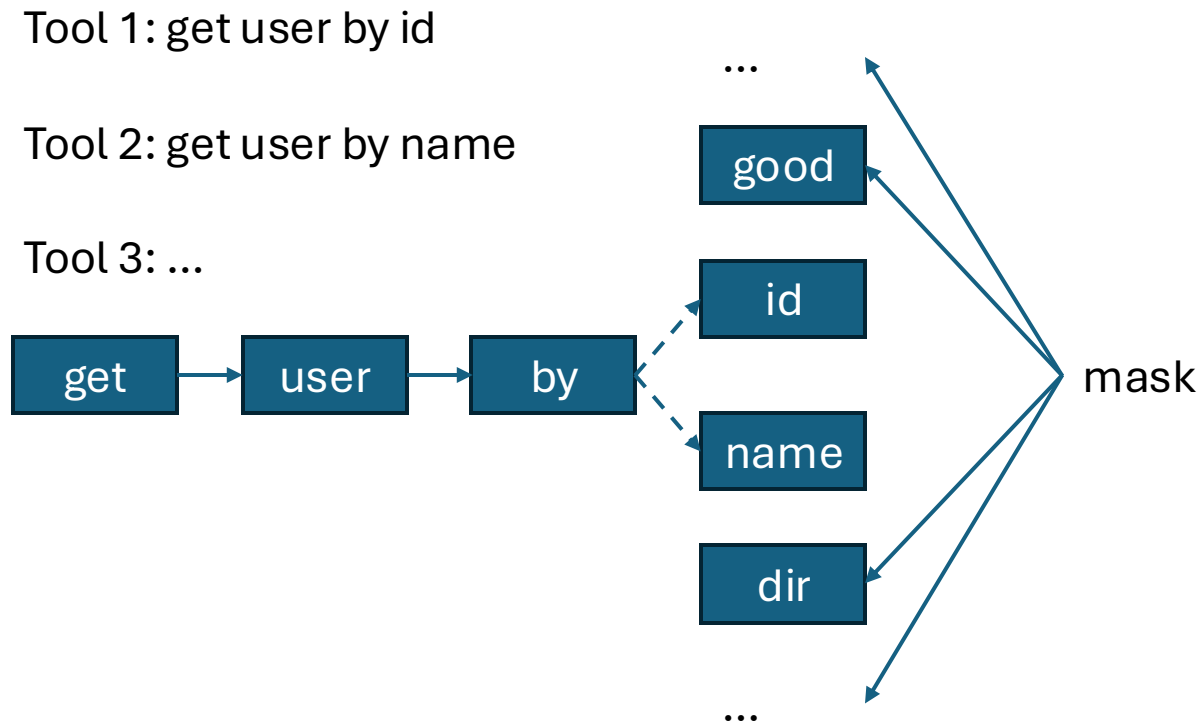
User: Documentation: {"name": <token 00123>, "description": "check the quality of a QR code..."}

ToolGen: {"file": "data/qr_code.png"}

User: {"response": "The qr code is ...

ToolGen: Inference

- Constrained Beam Search
 - Constrain beam search path to feasible tokens



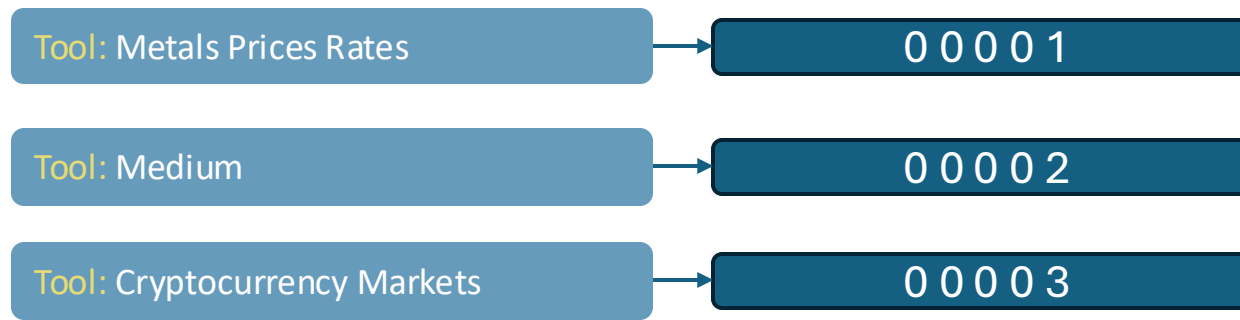
Tool Retrieval

- Model: Llama-3-8B

Model	I1			I2			I3		
	NDCG1	NDCG3	NDCG5	NDCG1	NDCG3	NDCG5	NDCG1	NDCG3	NDCG5
In-Domain									
BM25	29.46	31.12	33.27	24.13	25.29	27.65	32.00	25.88	29.78
EmbSim	63.67	61.03	65.37	49.11	42.27	46.56	53.00	46.40	52.73
Re-Invoke*	69.47	—	61.10	54.56	—	53.79	59.65	—	59.55
IterFeedback*	90.70	90.95	<u>92.47</u>	<u>89.01</u>	<u>85.46</u>	<u>87.10</u>	91.74	87.94	90.20
ToolRetriever	80.50	79.55	84.39	71.18	64.81	70.35	70.00	60.44	64.70
ToolGen	<u>89.17</u>	<u>90.85</u>	92.67	91.45	88.79	91.13	<u>87.00</u>	<u>85.59</u>	<u>90.16</u>
Multi-Domain									
BM25	22.77	22.64	25.61	18.29	20.74	22.18	10.00	10.08	12.33
EmbSim	54.00	50.82	55.86	40.84	36.67	39.55	18.00	17.77	20.70
ToolRetriever	72.31	70.30	74.99	64.54	57.91	63.61	52.00	39.89	42.92
ToolGen	87.67	88.84	91.54	83.46	86.24	88.84	79.00	79.80	84.79

ToolGen: Indexing

- Numeric Index: Represent each tool with a number



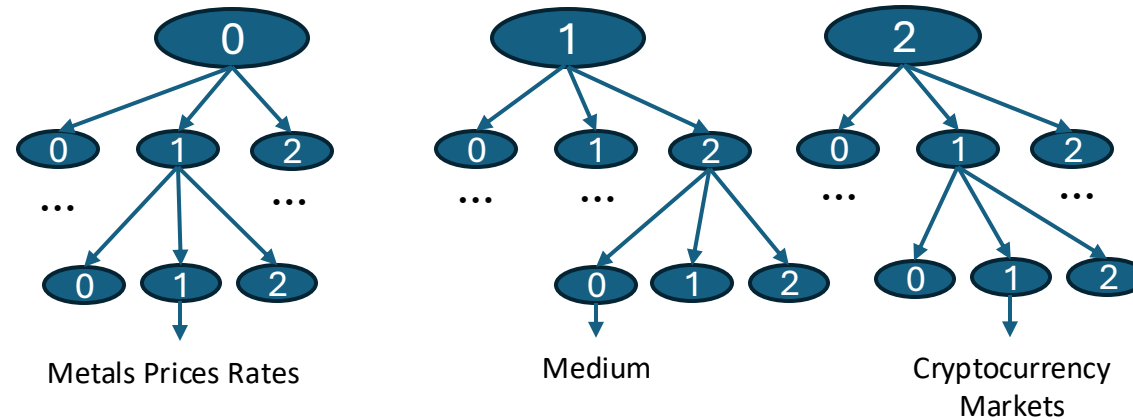
ToolGen: Indexing

- Hierarchical Index: Clustering tools into a hierarchical structure and map each tool with their path

Clustering tools into:

Further clustering into:

Further clustering into:



Tool: Metals Prices Rates

0 1 1

Tool: Medium

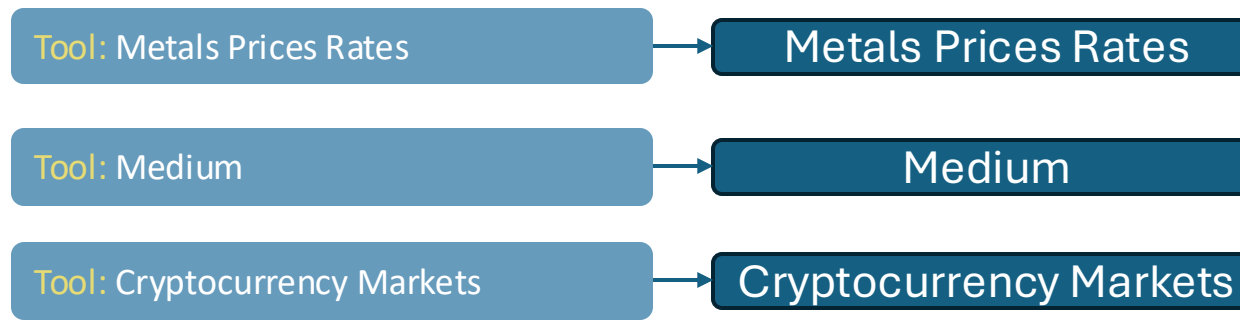
1 2 0

Tool: Cryptocurrency Markets

2 1 1

ToolGen: Indexing

- Semantic Index: Represent each tool with its name

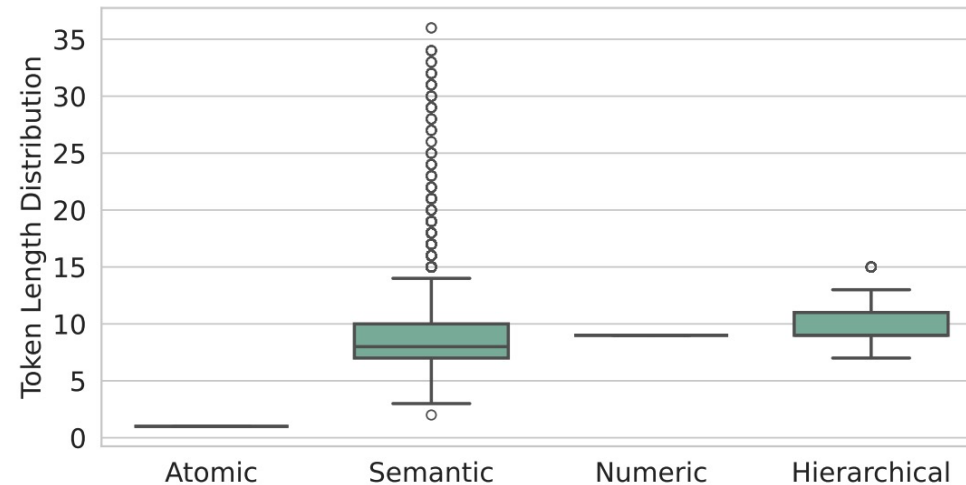


ToolGen: Indexing

- Results

Model	I1			I2			I3		
	NDCG1	NDCG3	NDCG5	NDCG1	NDCG3	NDCG5	NDCG1	NDCG3	NDCG5
Numeric	83.17	84.99	88.73	79.20	79.23	83.88	71.00	74.81	82.95
Hierarchical	85.67	87.38	90.26	82.22	82.70	86.63	78.50	<u>79.47</u>	84.15
Semantic	89.17	91.29	93.29	83.71	<u>84.51</u>	<u>88.22</u>	82.00	78.86	85.43
Atomic	<u>87.67</u>	<u>88.84</u>	<u>91.54</u>	<u>83.46</u>	86.24	88.84	<u>79.00</u>	79.80	<u>84.79</u>

- Token Length



ToolGen: End-to-End

- Results

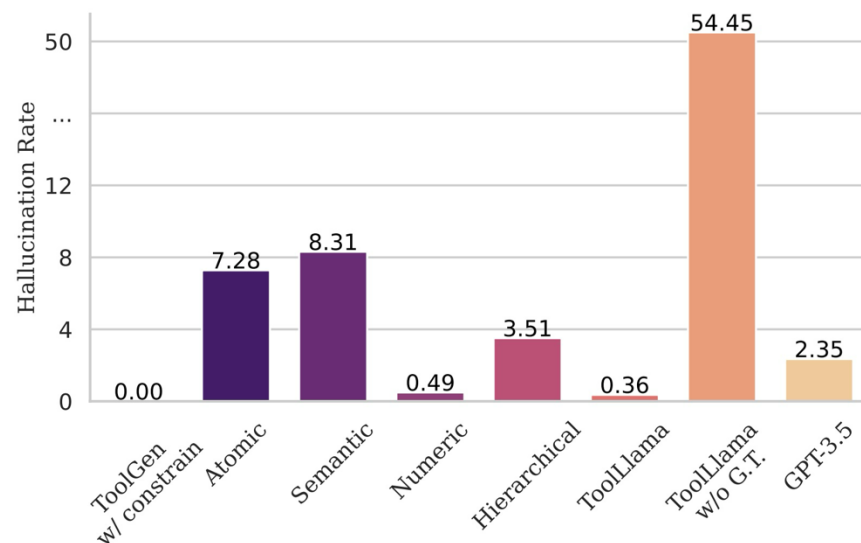
Model	SoPR				SoWR			
	I1	I2	I3	Avg.	I1	I2	I3	Avg
w/ Ground Truth Tools (G.T.)								
GPT-3.5	56.60	47.80	54.64	50.91	-	-	-	-
ToolLlama-2	53.37	41.98	46.45	48.43	47.27	59.43	27.87	47.58
ToolLlama	55.93	48.27	52.19	52.78	50.31	53.77	31.15	47.88
ToolGen	61.35	49.53	43.17	54.19	51.53	57.55	31.15	49.70
w/ Retriever (R.)								
GPT-3.5	51.43	41.19	34.43	45.00	53.37	53.77	37.70	50.60
ToolLlama-2	56.13	49.21	34.70	49.95	50.92	53.77	21.31	46.36
ToolLlama	54.60	49.96	51.37	51.55	49.08	61.32	31.15	49.70
ToolGen	56.13	52.20	47.54	53.28	50.92	62.26	34.42	51.51

ToolGen: End-to-End

- Indexing

Indexing	SoPR				SoWR			
	I1	I2	I3	Avg.	I1	I2	I3	Avg
Numeric	34.76	29.87	46.99	35.45	25.77	33.02	29.51	28.79
Hierarchical	50.20	45.60	32.79	45.50	38.04	43.40	29.51	38.18
Semantic	58.79	45.28	44.81	51.87	49.69	57.55	26.23	47.88
Atomic	58.08	56.13	44.81	55.00	47.85	57.55	29.51	47.58

- Hallucination



Conclusions

- Conclusions

- ToolGen outperforms pervious unsupervised discrete or supervised dense retrieval
- ToolGen Agent outperforms previous retriever-based paradigm