Graph-based Agent Memory

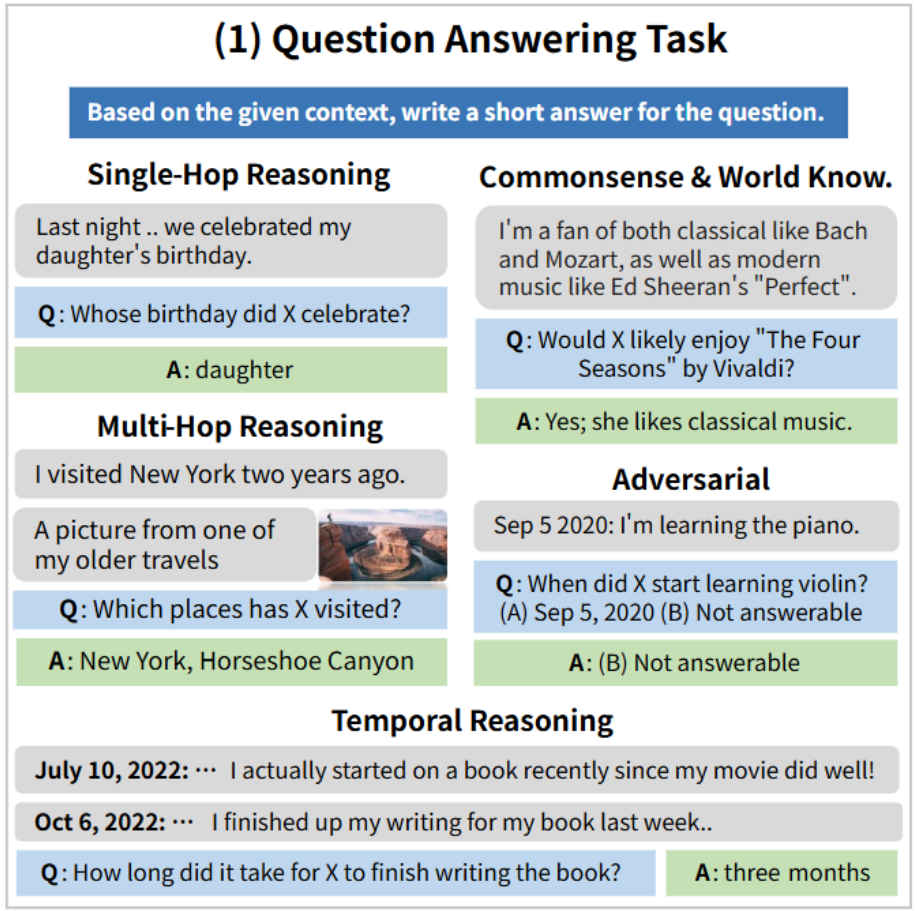
2026.5.29

Long-term Conversation QA Task

- QA Based on a very long conversation

- LoCoMo QA dataset:
- 1. Single-Hop Question (42.3%)
- 2. Multi-Hop Question (14.2%)
- 3. Temporal Reasoning Question (16.1%)
- 4. Open-domain Knowledge Question (4.8%)
- 5. Adversarial Question (22.4%)

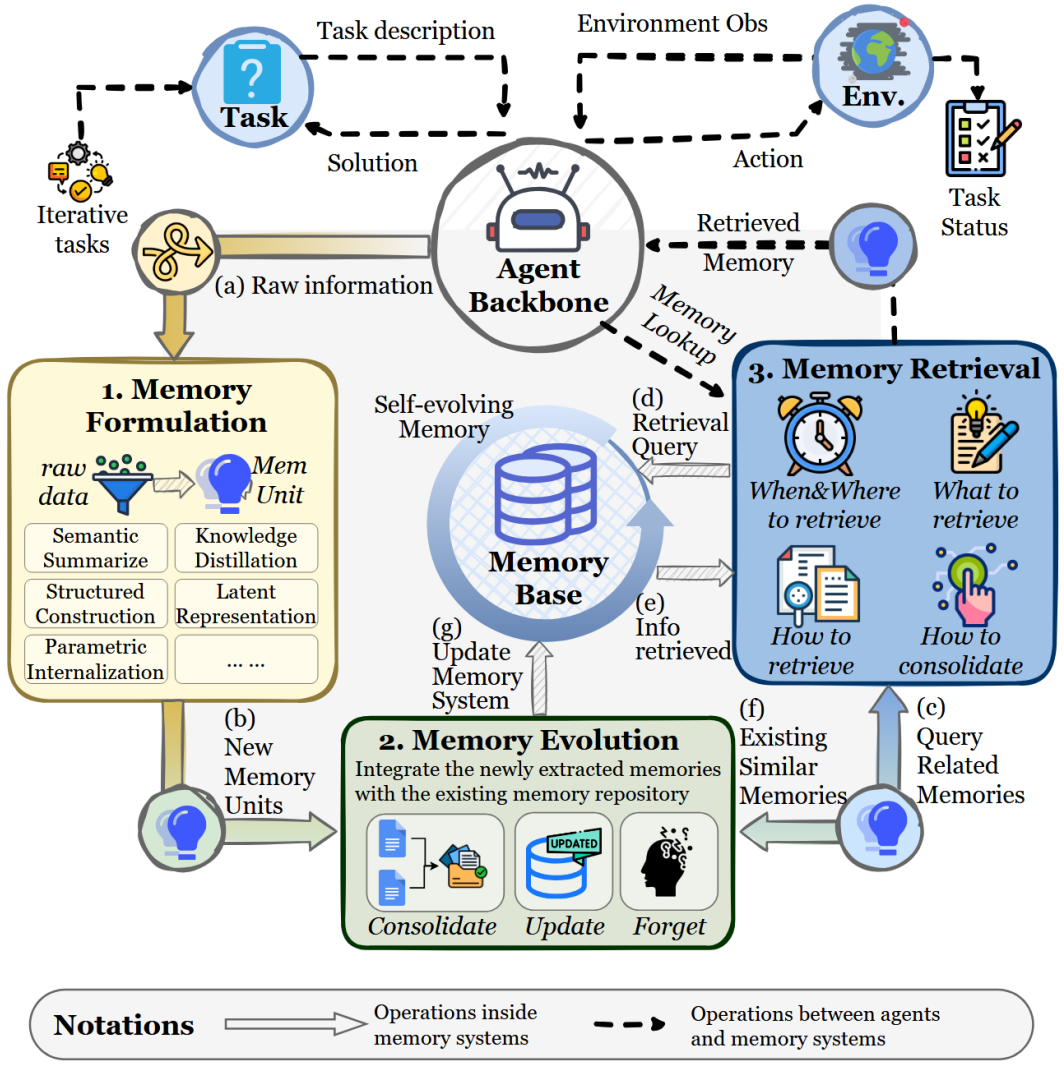
- 10 very long conversations, 1986 questions
- Avg. 588.2 turns, 27.2 sessions, 16618.1 tokens per conv.



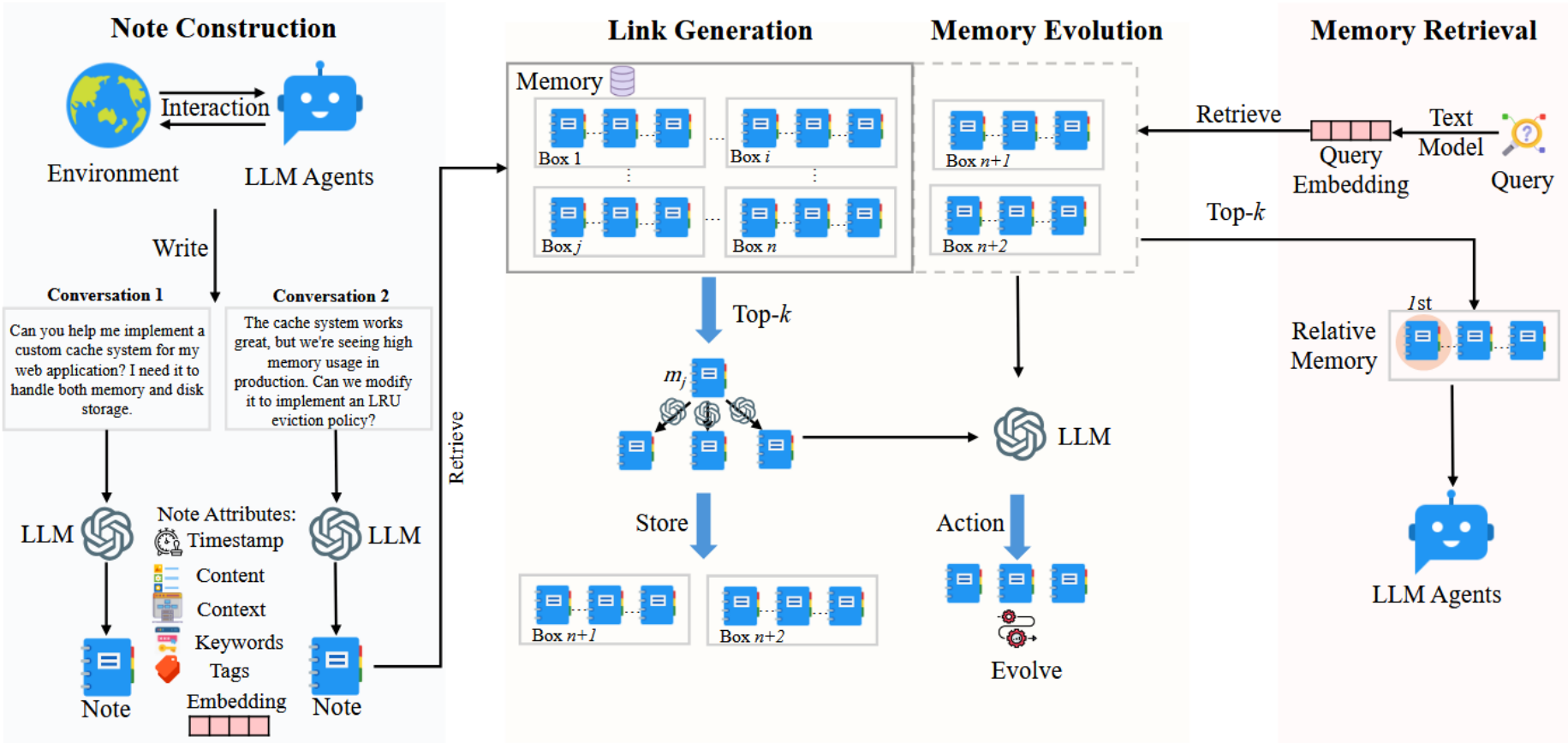
Agent Memory

- Use agent to process and retrieve memory

- 3 Steps:
- Formulation
- Evolution
- Retrieval



A-Mem: Agentic Memory for LLM Agents



A-Mem

- Traditional memory systems require developers to predefine memory storage structures.

- Need a flexible and universal memory system that supports LLM agents' long-term interactions

(a) Traditional memory system. **(b)** Our proposed agentic memory.

A-Mem

- Note Construction

$$m_i = \{c_i, t_i, K_i, G_i, X_i, e_i, L_i\}$$

$$K_i, G_i, X_i \leftarrow LLM(c_i, t_i, P_{s1})$$

$$e_i = f_{enc}[concat(c_i, K_i, G_i, X_i)]$$

c_i: original content
t_i: timestamp
K_i: keywords
G_i: tag for categorization
X_i: semantic understanding
e_i: embedding
L_i: linked memories
P: Prompt

A-Mem

- Link Generation

$$s_{n,j} = \frac{e_n \cdot e_j}{|e_n||e_j|}$$

$$M_{near}^{n} = \{m_j | rank(s_{n,j}) \leq k, m_j \in M\}$$

$$L_i \leftarrow LLM(m_n || M_{near}^{n} || P_{s2})$$

c_i: original content
t_i: timestamp
K_i: keywords
G_i: tag for categorization
X_i: semantic understanding
e_i: embedding
L_i: linked memories
P: Prompt

A-Mem

- Memory Evolution

$$m_j^* \leftarrow LLM\left(m_n || M_{near}^n \backslash \mathrm{m}_j || m_j || P_{s3}\right)$$

c_i: original content
t_i: timestamp
K_i: keywords
G_i: tag for categorization
X_i: semantic understanding
e_i: embedding
L_i: linked memories
P: Prompt

A-Mem

- Memory Retrieval

$$e_q = f_{enc}(q)$$

$$s_{q,i} = \frac{e_q \cdot e_i}{|e_q||e_i|}$$

$$M_{retrieved} = \{m_i | rank(s_{q,i}) \leq k, m_i \in M\}$$

c_i: original content
t_i: timestamp
K_i: keywords
G_i: tag for categorization
X_i: semantic understanding
e_i: embedding
L_i: linked memories
P: Prompt

A-Mem

- Experiment

Model		Method	Category										Average		
			Multi Hop		Temporal		Open Domain		Single Hop		Adversial		Ranking		Token
			F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	Length
GPT	4o-mini	LoCoMo	25.02	19.75	18.41	14.77	12.04	11.16	40.36	29.05	**69.23**	**68.75**	2.4	2.4	16,910
		ReadAgent	9.15	6.48	12.60	8.87	5.31	5.12	9.67	7.66	9.81	9.02	4.2	4.2	643
		MemoryBank	5.00	4.77	9.68	6.99	5.56	5.94	6.61	5.16	7.36	6.48	4.8	4.8	432
		MemGPT	26.65	17.72	25.52	19.44	9.15	7.44	41.04	34.34	43.29	42.73	2.4	2.4	16,977
		A-Mem	**27.02**	**20.09**	**45.85**	**36.67**	**12.14**	**12.00**	**44.65**	**37.06**	50.03	49.47	**1.2**	**1.2**	2,520
	4o	LoCoMo	28.00	18.47	9.09	5.78	16.47	14.80	**61.56**	**54.19**	**52.61**	**51.13**	2.0	2.0	16,910
		ReadAgent	14.61	9.95	4.16	3.19	8.84	8.37	12.46	10.29	6.81	6.13	4.0	4.0	805
		MemoryBank	6.49	4.69	2.47	2.43	6.43	5.30	8.28	7.10	4.42	3.67	5.0	5.0	569
		MemGPT	30.36	22.83	17.29	13.18	12.24	11.87	60.16	53.35	34.96	34.25	2.4	2.4	16,987
		A-Mem	**32.86**	**23.76**	**39.41**	**31.23**	**17.10**	**15.84**	48.43	42.97	36.35	35.53	**1.6**	**1.6**	1,216
Qwen2.5	1.5b	LoCoMo	9.05	6.55	4.25	4.04	9.91	8.50	11.15	8.67	40.38	40.23	3.4	3.4	16,910
		ReadAgent	6.61	4.93	2.55	2.51	5.31	12.24	10.13	7.54	5.42	27.32	4.6	4.6	752
		MemoryBank	11.14	8.25	4.46	2.87	8.05	6.21	13.42	11.01	36.76	34.00	2.6	2.6	284
		MemGPT	10.44	7.61	4.21	3.89	13.42	11.64	9.56	7.34	31.51	28.90	3.4	3.4	16,953
		A-Mem	**18.23**	**11.94**	**24.32**	**19.74**	**16.48**	**14.31**	**23.63**	**19.23**	**46.00**	**43.26**	**1.0**	**1.0**	1,300
	3b	LoCoMo	4.61	4.29	3.11	2.71	4.55	5.97	7.03	5.69	16.95	14.81	3.2	3.2	16,910
		ReadAgent	2.47	1.78	3.01	3.01	5.57	5.22	3.25	2.51	15.78	14.01	4.2	4.2	776
		MemoryBank	3.60	3.39	1.72	1.97	6.63	6.58	4.11	3.32	13.07	10.30	4.2	4.2	298
		MemGPT	5.07	4.31	2.94	2.95	7.04	7.10	7.26	5.52	14.47	12.39	2.4	2.4	16,961
		A-Mem	**12.57**	**9.01**	**27.59**	**25.07**	**7.12**	**7.28**	**17.23**	**13.12**	**27.91**	**25.15**	**1.0**	**1.0**	1,137
Llama 3.2	1b	LoCoMo	11.25	9.18	7.38	6.82	11.90	10.38	12.86	10.50	51.89	48.27	3.4	3.4	16,910
		ReadAgent	5.96	5.12	1.93	2.30	12.46	11.17	7.75	6.03	44.64	40.15	4.6	4.6	665
		MemoryBank	13.18	10.03	7.61	6.27	15.78	12.94	17.30	14.03	52.61	47.53	2.0	2.0	274
		MemGPT	9.19	6.96	4.02	4.79	11.14	8.24	10.16	7.68	49.75	45.11	4.0	4.0	16,950
		A-Mem	**19.06**	**11.71**	**17.80**	**10.28**	**17.55**	**14.67**	**28.51**	**24.13**	**58.81**	**54.28**	**1.0**	**1.0**	1,376
	3b	LoCoMo	6.88	5.77	4.37	4.40	10.65	9.29	8.37	6.93	30.25	28.46	2.8	2.8	16,910
		ReadAgent	2.47	1.78	3.01	3.01	5.57	5.22	3.25	2.51	15.78	14.01	4.2	4.2	461
		MemoryBank	6.19	4.47	3.49	3.13	4.07	4.57	7.61	6.03	18.65	17.05	3.2	3.2	263
		MemGPT	5.32	3.99	2.68	2.72	5.64	5.54	4.32	3.51	21.45	19.37	3.8	3.8	16,956
		A-Mem	**17.44**	**11.74**	**26.38**	**19.50**	**12.53**	**11.83**	**28.14**	**23.87**	**42.04**	**40.60**	**1.0**	**1.0**	1,126

GAAMA: Graph Augmented Associative Memory for Agents

- Existing approaches face distinct limitations:
- HippoRAG: entity-centric design causes a mega-hub problem
- A-Mem: retrieves primarily via embedding similarity without graph-based relevance propagation

- Highlight:
- Concept-mediated knowledge graph
- Hybrid retrieval with Personalized PageRank

GAAMA

- Construction
- Step 1: Episode preservation

- Save raw context
- Each conversation message maps to one episode node
- Episodes are temporally chained via NEXT edges

GAAMA

- Construction
- Step 2: Fact and concept extraction

- Facts:
- Atomic factual assertions about
- Facts are linked to their source episodes via DERIVED_FROM edges, preserving provenance.
- top k similar historical episodes and facts are passed into the context of the LLM to derive the new fact set.

GAAMA

- Construction
- Step 2: Fact and concept extraction

- Concepts:
- Topic-level labels that capture the thematic content of episodes and facts.
- Episodes are linked to their concepts via HAS_CONCEPT edges; facts are linked via ABOUT_CONCEPT edges.
- top k similar historical episodes and facts are passed into the context of the LLM to derive the new fact set.

GAAMA

- Construction
- Step 3: Reflection synthesis

- Reflections:
- higher-order insights from multiple facts
- Reflections capture generalized patterns, preferences, or lessons that span multiple conversations
- Each reflection is linked to its supporting facts via DERIVED_FROM_FACT edges.

GAAMA

- Construction
- Step 3: Reflection synthesis

- Reflections:
- higher-order insights from multiple facts
- Reflections capture generalized patterns, preferences, or lessons that span multiple conversations
- Each reflection is linked to its supporting facts via DERIVED_FROM_FACT edges.

GAAMA

- Retrieval
- Step 1: KNN candidate retrieval and seed selection

- retrieve a broad candidate pool of 2B nodes by cosine similarity (B = max_facts + max_reflections + max_episodes)
- From this pool, the top-k nodes (default k = 40) are selected as PPR seeds.
- The remaining 2B – k KNN candidates are retained in the candidate pool for final scoring.

GAAMA

- Retrieval
- Step 2: Graph expansion

- expand outward along graph edges to depth d (default d = 2) from k seed nodes
- Any graph-discovered nodes not already in the KNN candidate pool are fetched and added to it
- Enable PPR to discover nodes that are structurally connected to the seeds but may not appear in the KNN results

GAAMA

- Retrieval
- Step 3: Edge-type-aware transition weights

- Transition probabilities are computed by per-source normalization

$$P_{ij} = \frac{\tilde{w}_{ij}}{\sum_k \tilde{w}_{ik}}$$

- Hub dampening: $\theta = 50$

$$\tilde{w}_{ij}^{\text{damped}} = \tilde{w}_{ij} \cdot \min\left(1, \frac{\theta}{\deg(i)}\right)$$

GAAMA

- Retrieval
- Step 4: Personalized PageRank

- run iterative PPR on the local subgraph

$$r_j^{(t+1)} = (1-\alpha) \cdot v_j + \alpha \sum_i r_i^{(t)} \cdot P_{ij} + \alpha \cdot S^{(t)} \cdot v_j$$

where $S^{(t)} = \sum_{i:\, \deg(i)=0} r_i^{(t)}$

GAAMA

- Retrieval
- Step 5: Additive scoring

- The final relevance score is computed over all candidate nodes
- Both PPR scores and similarity scores are max-normalized to [0, 1]

$$\text{score}(n) = w_{\text{ppr}} \cdot \text{ppr}(n) + w_{\text{sim}} \cdot \text{sim}(n, q)$$

where $w_{\text{ppr}} = 0.1$ and $w_{\text{sim}} = 1.0$

GAAMA

- Memory packing and context assembly

- Retrieved nodes are bucketed by type and subject to per-type budget caps: max_facts=60, max_reflections=20, max_episodes=80.
- the lowest-scored items are removed first until a total word budget (max_memory_words=1000) is satisfied.
- Episodes are additionally sorted by their temporal sequence number to preserve chronological order in the assembled context.

GAAMA

- Experiment
- QA dataset without adversarial questions, total 1540 questions

$$r = \frac{\text{number of reference key facts found in hypothesis}}{\text{total key facts in reference answer}}$$

System	Multi-hop (Cat1)	Temporal (Cat2)	Open Domain (Cat3)	Single Hop (Cat4)	Overall
A-Mem [Xu et al., 2025]	44.7	37.4	50.0	51.5	47.2
Nemori [Nan et al., 2025]	49.4	45.0	36.8	57.4	52.1
HippoRAG [Gutierrez et al., 2024]	61.7	67.0	67.7	74.1	69.9
RAG Baseline	67.5	59.0	44.6	87.1	75.0
GAAMA (ours)	**72.2**	**71.9**	**49.3**	**87.2**	**78.9**

GAAMA

- Experiment

Table 5: Ablation: Semantic vs. PPR=0.1 retrieval, per category. Δ = PPR=0.1 − Semantic (pp).

Category	n	Semantic (%)	PPR=0.1 (%)	Δ (pp)
Cat1 (Multi-hop)	282	72.2	72.2	0.0
Cat2 (Temporal)	321	71.4	71.9	+0.5
Cat3 (Open Domain)	96	49.1	49.3	+0.2
Cat4 (Single Hop)	841	85.7	87.2	+1.6
Overall	**1540**	**78.0**	**78.9**	**+1.0**

Table 7: Heatmap of reward delta (pp): PPR=0.1 − Semantic, per conversation and category.

$> +5$ | $+2$ to $+5$ | 0 to $+2$ | -2 to 0 | -5 to -2 | < -5

Conv.	Cat1	Cat2	Cat3	Cat4	Overall
conv-26	+1.8	+4.0	−3.8	+2.4	+2.1
conv-30	−2.3	+3.8	—	+5.3	+3.8
conv-41	+0.8	+3.7	0.0	−2.0	−0.3
conv-42	−4.6	+5.0	−9.1	+2.3	+0.9
conv-43	−4.0	+5.8	−2.4	+2.8	+1.6
conv-44	+6.0	−4.2	+11.9	−0.3	+1.2
conv-47	+0.8	−4.9	−3.8	−0.6	−1.7
conv-48	−1.2	−4.4	0.0	+3.6	+1.1
conv-49	+1.6	+1.5	+5.2	−1.4	+0.5
conv-50	0.0	−4.7	+14.3	+3.3	+1.5
Avg.	**0.0**	**+0.5**	**+0.2**	**+1.6**	**+1.0**