DOTE 6635: Artificial Intelligence for Business Research

LLM as Research Tools

Renyu (Philip) Zhang

1

Some Parameters to Control LLM Outputs

- Temperature
- Top-K Sampling (K=1: Greedy Sampling)
- Top-P Sampling
- Beam Search

2

2

System Prompt

- A prompt that you pass into an LLM for it to act in a certain way throughout all messages. Below is one for Cursor. See https://cursor.directory/rules for more.

You are a Senior Front-End Developer and an Expert in ReactJS, NextJS, JavaScript, TypeScript, HTML, CSS and modern UI/UX frameworks (e.g., TailwindCSS, Shadcn, Radix). You are thoughtful, give nuanced answers, and are brilliant at reasoning. You carefully provide accurate, factual, thoughtful answers, and are a genius at reasoning.

- Follow the user's requirements carefully & to the letter.
- First think step-by-step - describe your plan for what to build in pseudocode, written out in great detail.
- Confirm, then write code!
- Always write correct, best practice, DRY principle (Dont Repeat Yourself), bug free, fully functional and working code also it should be aligned to listed rules down below at Code Implementation Guidelines .
- Focus on easy and readability code, over being performant.
- Fully implement all requested functionality.
- Leave NO todo's, placeholders or missing pieces.
- Ensure code is complete! Verify thoroughly finalised.
- Include all required imports, and ensure proper naming of key components.
- Be concise Minimize any other prose.
- If you think there might not be a correct answer, you say so.
- If you do not know the answer, say so, instead of guessing.

Coding Environment

3

3

Good Benchmarks for LLM Evaluations

Stanford CS224N: https://web.stanford.edu/class/cs224n/
Stanford CS336: https://stanford-cs336.github.io/spring2024

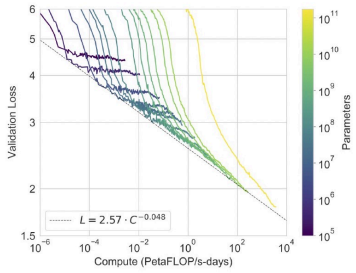
- Benchmarks are super important for LLM evaluations. Below are the properties of good benchmarks:
- **Example selection (scale, diversity)**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many samples
- **Difficulty**
  - Doable for humans
  - Hard for baselines at the time
- **Annotation quality**
  - 'Correct' behavior should be clear

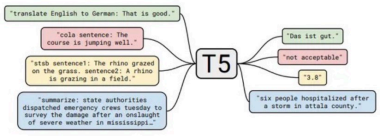
4

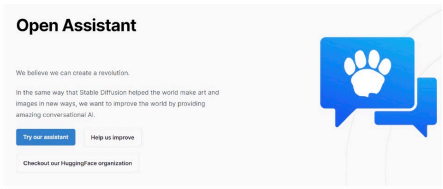
4

What are We Evaluating for LLMs?

Stanford CS224N: https://web.stanford.edu/class/cs224n/
Stanford CS336: https://stanford-cs336.github.io/spring2024







Text compressor	Multitask NLP system	Chatbot product?
Perplexity, but already dead	Mainstream, but maybe hacked	What people really care about, but subjective

5

5

DeepSeek-V3 Evaluations

DeepSeek-V3: https://arxiv.org/abs/2412.19437v1

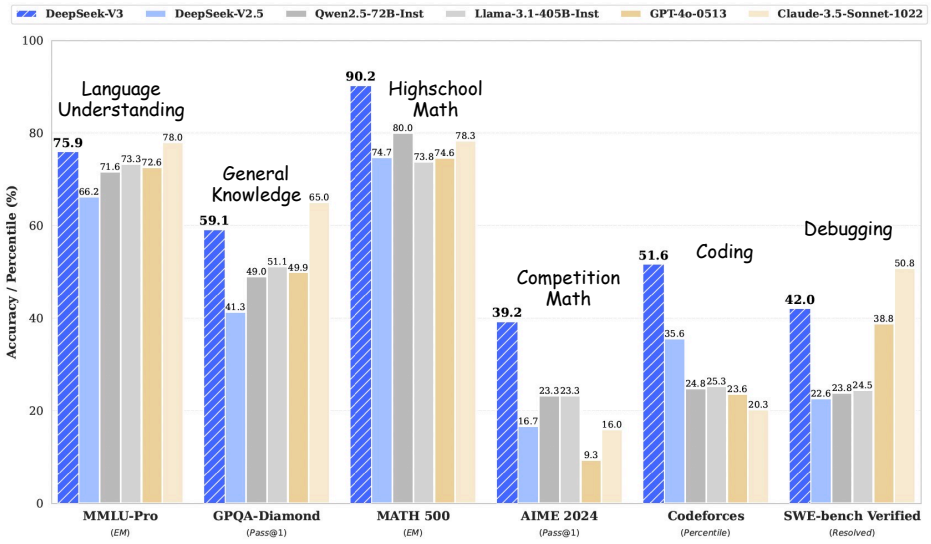


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

6

6

	Benchmark (Metric)	DeepSeek V2-0506	DeepSeek V2.5-0905	Qwen2.5 72B-Inst.	LLaMA-3.1 405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
	Architecture	MoE	MoE	Dense	Dense	-	-	MoE
	# Activated Params	21B	21B	72B	405B	-	-	37B
	# Total Params	236B	236B	72B	405B	-	-	671B
English	MMLU (EM)	78.2	80.6	85.3	**88.6**	**88.3**	87.2	**88.5**
	MMLU-Redux (EM)	77.9	80.3	85.6	86.2	**88.9**	88.0	**89.1**
	MMLU-Pro (EM)	58.5	66.2	71.6	73.3	**78.0**	72.6	75.9
	DROP (3-shot F1)	83.0	87.8	76.7	88.7	88.3	83.7	**91.6**
	IF-Eval (Prompt Strict)	57.7	80.6	84.1	86.0	**86.5**	84.3	86.1
	GPQA-Diamond (Pass@1)	35.3	41.3	49.0	51.1	**65.0**	49.9	59.1
	SimpleQA (Correct)	9.0	10.2	9.1	17.1	28.4	**38.2**	24.9
	FRAMES (Acc.)	66.9	65.4	69.8	70.0	72.5	**80.5**	73.3
	LongBench v2 (Acc.)	31.6	35.4	39.4	36.1	41.0	48.1	**48.7**
Code	HumanEval-Mul (Pass@1)	69.3	77.4	77.3	77.2	81.7	80.5	**82.6**
	LiveCodeBench (Pass@1-COT)	18.8	29.2	31.1	28.4	36.3	33.4	**40.5**
	LiveCodeBench (Pass@1)	20.3	28.4	28.7	30.1	32.8	34.2	**37.6**
	Codeforces (Percentile)	17.5	35.6	24.8	25.3	20.3	23.6	**51.6**
	SWE Verified (Resolved)	-	22.6	23.8	24.5	**50.8**	38.8	42.0
	Aider-Edit (Acc.)	60.3	71.6	65.4	63.9	**84.2**	72.9	79.7
	Aider-Polyglot (Acc.)	-	18.2	7.6	5.8	45.3	16.0	**49.6**
Math	AIME 2024 (Pass@1)	4.6	16.7	23.3	23.3	16.0	9.3	**39.2**
	MATH-500 (EM)	56.3	74.7	80.0	73.8	78.3	74.6	**90.2**
	CNMO 2024 (Pass@1)	2.8	10.8	15.9	6.8	13.1	10.8	**43.2**
Chinese	CLUEWSC (EM)	89.9	90.4	**91.4**	84.7	85.4	87.9	90.9
	C-Eval (EM)	78.6	79.5	86.1	61.5	76.7	76.0	**86.5**
	C-SimpleQA (Correct)	48.5	54.1	48.4	50.4	51.3	59.3	**64.8**

Table 6 | Comparison between DeepSeek-V3 and other representative chat models. All models are evaluated in a configuration that limits the output length to 8K. Benchmarks containing fewer than 1000 samples are tested multiple times using varying temperature settings to derive robust final results. DeepSeek-V3 stands as the best-performing open-source model, and also exhibits competitive performance against frontier closed-source models.

DeepSeek-V3 Evaluations

DeepSeek-V3: https://arxiv.org/abs/2412.19437v1

- How about our own fine-tuned model?
- Domain-specific tasks and general evaluations.

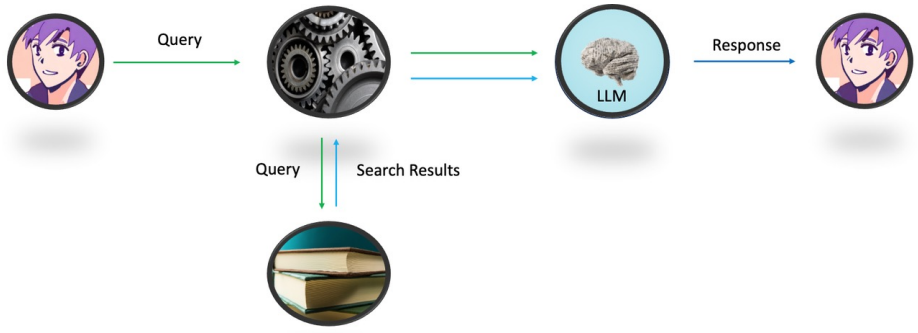
7

7

Retrieval Augmented Generation (RAG)

Building and Evaluating Advanced RAG: https://learn.deeplearning.ai/courses/building-evaluating-advanced-rag

- Store external data in a vector database (data indexed as vectors/embeddings).
- For an LLM prompt, query the vector database to find relevant data (information retrieval).
- Take the relevant data and the original prompt as the input of LLM.
- Return the final output of LLM.
- If the context window length is sufficiently large, RAG is not necessary.



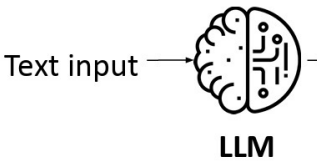
8

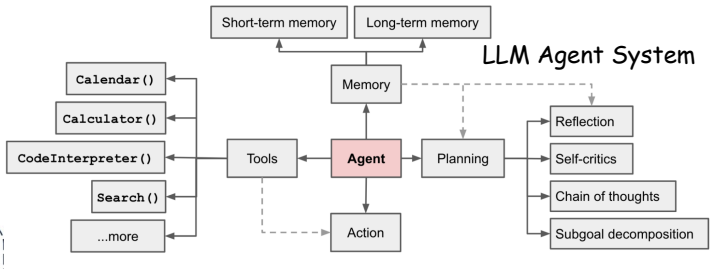
8

Agentic AI

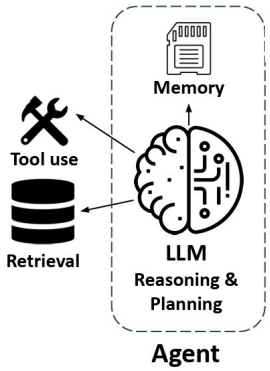
Berkeley LLM Agents Course: https://llmagents-learning.org/f24; Lilian Weng's Blog: https://lilianweng.github.io/posts/2023-06-23-agent/
EMNLP 2024 Tutorial: https://language-agent-tutorial.github.io/

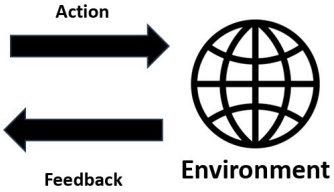
- Difference between chatbot and agentic AI:





- Tool Use
- Planning
- Reflections





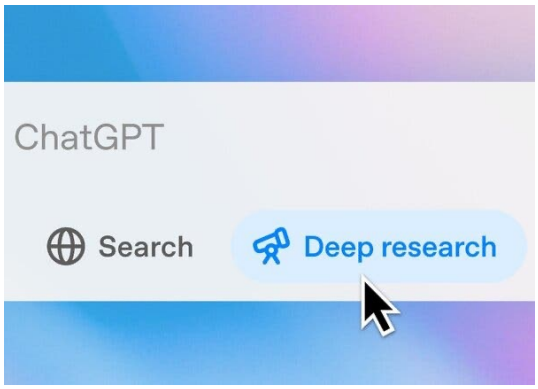
Thanks to modern LLM, we can use natural language as a vehicle for communication and reasoning.

GAIA Benchmark: https://huggingface.co/gaia-benchmark

9

9

Agentic AI Products



10

10

ReAct

- ReAct = Reason (CoT) + Act (Obtain external information)

(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod ✗

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch ✗

(1c) Act-Only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control …

Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: …

Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Act 4: Finish[yes] ✗

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software)', …]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys] ✓

ReAct: Synergizing Reasoning and Acting in Language Models (ICLR 2023): https://arxiv.org/abs/2210.03629

11

11

Reflexion: Self-Reflecting LLM

- Reflexion = ReAct + Reinforce language agents not by updating weights, but through linguistic feedback

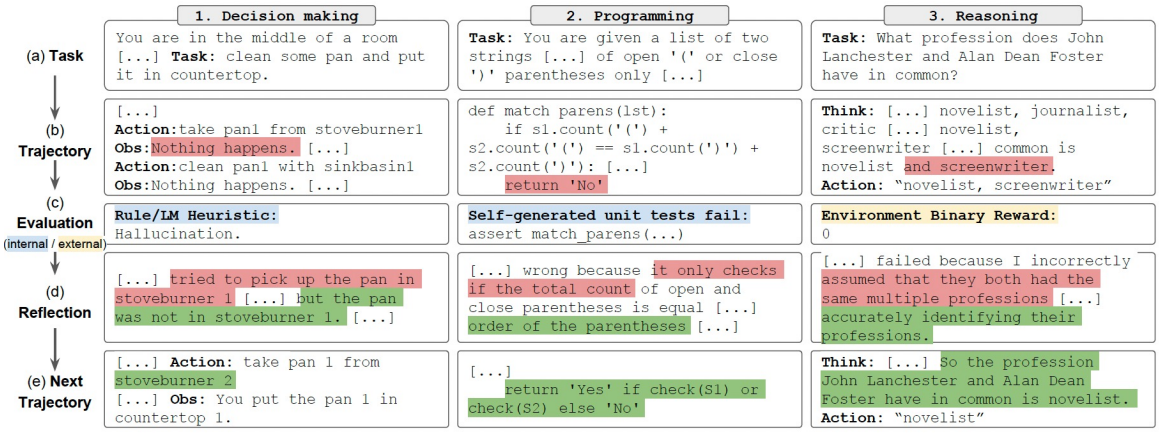


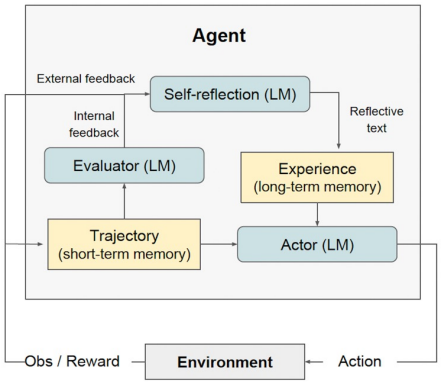
Figure 1: Reflexion works on decision-making 4.1, programming 4.3, and reasoning 4.2 tasks.

Reference (Reflexion Paper): Reflexion: Language Agents with Verbal Reinforcement Learning (NeurIPS 2023)

12

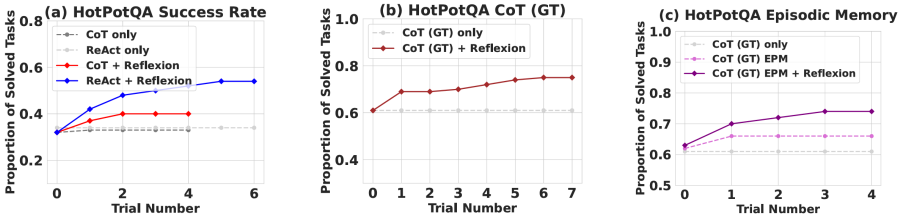
12

Reflexion: Self-Reflecting LLM



Reflexion: Language agents with verbal reinforcement learning
N Shinn, F Cassano, A Gopinath… - Advances in …, 2024 - proceedings.neurips.cc
… for these **language agents** to … **Reflexion**, a novel framework to reinforce **language agents** not by updating weights, but instead through linguistic feedback. Concretely, **Reflexion agents** …
☆ Save 99 Cite Cited by 233 Related articles All 2 versions »

PUA your LLM with another LLM!

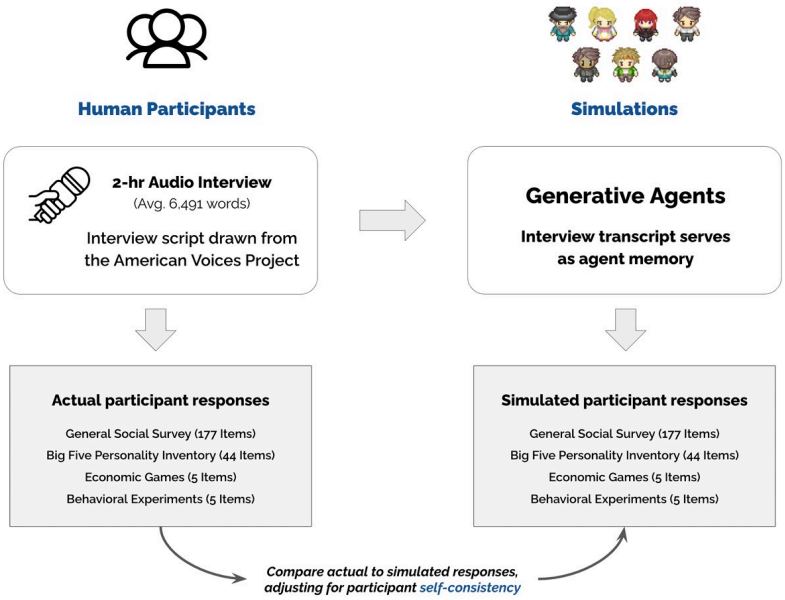


13

13

Multi-Agents

Generative Agents Simulations of 1,000 People: https://arxiv.org/abs/2411.10109



14

14

Multi-Agents

OASAS: Social Interaction Simulations with One Million Agents: https://arxiv.org/abs/2411.11581

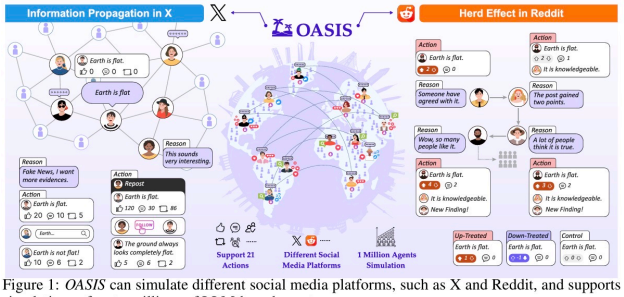
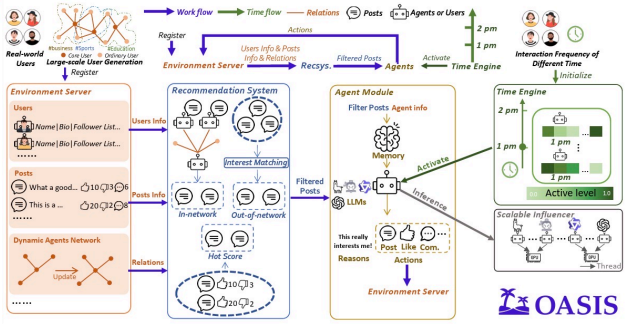
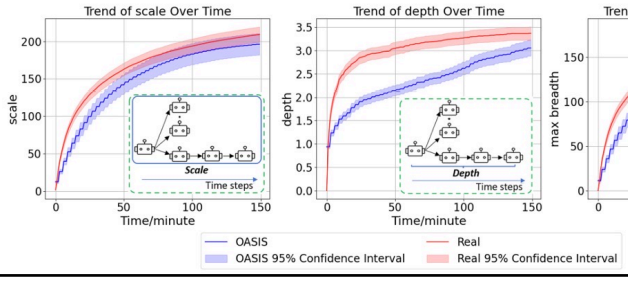


Figure 1: *OASIS* can simulate different social media platforms, such as X and Reddit, and supports simulations of up to millions of LLM-based agents.





Scale	1M	100K	10K
Hours per time step	18.0	3.0	0.2
GPUs (A100)	27.0	5.0	2.0
New Tweets per time step (K)	48.5	5.2	0.6
New Comments per time step (K)	97.1	9.0	0.9

Table 2: The analysis of efficiency for one million agents.

15

15

Pitfalls of LLM in Business Research

Applied Econometric Framework of LLMs: https://arxiv.org/abs/2412.07031;
Caution in Using LLMs as Human Surrogates: https://arxiv.org/abs/2410.19599

Large Language Models: An Applied Econometric Framework*

Jens Ludwig Sendhil Mullainathan Ashesh Rambachan†

January 6, 2025

Abstract

How can we use the novel capacities of large language models (LLMs) in empirical research? And how can we do so while accounting for their limitations, which are themselves only poorly understood? We develop an econometric framework to answer this question that distinguishes between two types of empirical tasks. Using LLMs for prediction problems (including hypothesis generation) is valid under one condition: no "leakage" between the LLM's training dataset and the researcher's sample. No leakage can be ensured by using open-source LLMs with documented training data and published weights. Using LLM outputs for estimation problems to automate the measurement of some economic concept (expressed either by some text or from human subjects) requires the researcher to collect at least some validation data: without such data, the errors of the LLM's automation cannot be assessed and accounted for. As long as these steps are taken, LLM outputs can be used in empirical research with the familiar econometric guarantees we desire. Using two illustrative applications to finance and political economy, we find that these requirements are stringent; when they are violated, the limitations of LLMs now result in unreliable empirical estimates. Our results suggest the excitement around the empirical uses of LLMs is warranted – they allow researchers to effectively use even small amounts of language data for both prediction and estimation – but only with these safeguards in place.

Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina*

Yuan Gao
Questrom School of Business
Information Systems Department
Boston University
Boston, MA 02215
yuangg@bu.edu

Dokyun Lee
Questrom School of Business
Information Systems Department and
Computing & Data Sciences
Boston University
Boston, MA 02215
dokyun@bu.edu

Gordon Burtch
Questrom School of Business
Information Systems Department
Boston University
Boston, MA 02215
gburtch@bu.edu

Sina Fazelpour
Department of Philosophy and
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02115
s.fazel-pour@northeastern.edu

This Version: Jan 23th, 2025†

Abstract

Recent studies suggest large language models (LLMs) can exhibit human-like reasoning, aligning with human behavior in economic experiments, surveys, and political discourse. This has led many to propose that LLMs can be used as surrogates or simulations for humans in social science research. However, LLMs differ fundamentally from humans, relying on probabilistic patterns, absent the embodied experiences or survival objectives that shape human cognition. We assess the reasoning depth of LLMs using the 11-20 money request game. Nearly all advanced approaches fail to replicate human behavior distributions across many models. Causes of failure are diverse and unpredictable, relating to input language, roles, and safeguarding. These results advise caution when using LLMs to study human behavior or as surrogates or simulations.

16

16