Initial Steps in Integrating Large Reasoning and Action Models for Service Composition

Ilche Georgievski and Marco Aiello, IAAS-SC, University of Stuttgart, DE

SummerSOC 2025

AI is eating the world

⑥ OpenAl	OpenAl	~	≅ cohere	Cohere	~	м	Mistral Al	~
G	Google	~	F Microsoft	Microsoft	~	ANTHROPIC	Anthropic	~
W deepseek	DeepSeek	~	Discover new ways to reate and connect with	Meta	~	01.AI	01.AI	~
Induction AT	Inflection AI	~	NVIDIA.	Nvidia	~	0	Contextual AI, Inc.	~
databricks	Databricks	~	Ś	GPT-4	~	∭, mosaic ^{w,}	Mosaic ML, Inc.	~
aws	AWS	~	la	Claude	~	IBM.	IBM	~
0	Meta Al	~	WHY LABS	Why Labs	~	EQ. Alibaba Group	Alibaba	~
amazon	Amazon	~	la	Claude 3 Haiku	~	≡	Falcon	~

AI is eating the wo

	R	ank 🗍	Nam	ne	🕴 Market Cap 🍦 Pric	ce 🕴	Today	Price (30 days)	Country
ing the world	\$	1 1		Microsoft MSFT	\$3.476 T	\$467.68	▲ 0.82%		🗾 USA
	☆ ∨	12		NVIDIA NVDA	\$3.414 T	\$139.99	- 1.36%	~~~~~	ISA USA
	☆	3	ć	Apple AAPL	\$2.996 T	\$200.63	- 1.08%	\sim	🗾 USA
	☆	4	G	Alphabet (Google)	\$2.049 T	\$169.81	• 0.25%	~~~~~	🗾 USA
	☆	5	∞	Meta Platforms (Facebook)	\$1.721 T	\$684.62	- 0.48%	~~~~	🗾 USA
	☆	6	T	Tesla TSLA	\$917.00 B	\$284.70	• 14.26%	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	🗾 USA
စြာ OpenAl OpenAl	☆	7	0	Oracle ORCL	\$479.91 B	\$171.14	- 1.81%	~~~~	🗾 USA
estimated at 300B	\$	8	Q	Palantir PLTR	\$282.97 B	\$119.91	~ 7.77%	m	🗾 USA
estimated at 500b	\$	9	IBM	IBM IBM	\$248.01 B	\$266.86	- 0.50%	~~~~	🗾 USA
		10	Δ	Adobe Adbe	\$176.95 B	\$415.20	• 0.31%		🗾 USA
	\$	11	C //	CoreWeave CRWV	\$64.82 B	\$135.05	~ 17.20%		🗾 USA
		12	Г	Cambricon Technologies 688256.SS	\$35.43 B	\$84.88	- 1.59%	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	📁 China
	\$	13	٢	Dynatrace	\$16.46 B	\$54.94	- 1.55%		🗾 USA
		14	'n	Mobileye MBLY	\$13.39 B	\$16.50	~ 1.70%	\sim	📼 Israel
	\$	15	"Г	Tempus Al	\$10.15 B	\$58.66	- 6.56%	m	🗾 USA
		16	А	Aurora Innovation	\$9.99 B	\$5.65	• 2.75%	m	🗾 USA
	\$	17	Ui	UiPath PATH	\$7.09 B	\$13.26	~ 1.61%	~~~~	III USA

"Company with no business plan buys company with no product" For 6.5 billion dollars







Builder.ai AI for software engineering or 700 SE humans



promised to make software creation "as easy as ordering pizza"

raised \$450 million and achieved a valuation of \$1.5 billion

reportedly owes \$85 million to Amazon and \$30 million to Microsoft in unpaid cloud services



on the other hand much more sustainable than real AI...

To estimate the **daily electricity consumption** for 700 people:

- Annual per capita consumption: 1,395 kWh
- Daily per capita consumption:
 1,395 kWh ÷ 365 days ≈ 3.82 kWh/day
- Total for 700 people:

3.82 kWh/day \times 700 people \approx 2,674 kWh/day



based on arxiv publications



COVER FEATURE VENTURE SCIENTISTS AND TECHNOLOGY ENTREPRENEURSHIP



Marco Aiello . University of Stuttgar

Al is transforming the global labor market, signaling a shift where jobs are both displaced and newly created. But is it really creating enough new jobs?

brings new entre- income per person and gross domestic product (GDP). Up preneurial opportunities and has the power until recently, economic growth has guaranteed that to revolutionize markets. In doing so, jobs can labor markets have many new opportunities and that swapped away, considerably downsized, or unemployment rates are low during periods of growth the fields to the factories, the end of the First Industrial machinery in a fig the neuron of the factories, the end of the errors mustations in machines in a target Recolution marked the shift of work from being avail-sable in factories to being in offices. Usually, the number experiencing the officie jobs created was greater than those lost. These fueled by the rapid revolutions brought economic growth in terms of average robotics, and the l robotics, and the I Digital Object Identifier 10.1209/MC.2024.3374279 Date of current version: 6 Mov 2074 also, what are the ket? In the presen

Authorized Journal use limited to: UNIVERSITAETSBIBL STUTTGART. Downloaded on May 03,2024 at 10-

A Challenge for the Next 50 Years of Automated Service Composition

Marco Aiello⁽⁵³⁾©

Department of Service Computing, IAAS, University of Stuttgart, Stuttgart, marco.aiello@iaas.uni-stuttgart.de

Abstract, Automated Service Composition emerged as a promising Abstract. Automated Service Composition emerged as a promising area of research at the biggining of the century. After treaty years, it appears to have reached a stagnating state where only little progress is made. In the present vision paper, 1 propose a challenge for automated service composition to be achieved in the sext 50 years. I set a scene in 2020 that service composition should be abs to handle by hom. Finally, I draws a parallel with autonomous driving to identify the major milestones in the quest to hilly autonomous service composition sputters.

Keywords: Automated Service Composition · Service-oriented Computing · Maturity Levels · Artificial Intelligence Planning

1 The Promises of Automated Service Composition

Automated service composition refers to systems that utilize distributed, discrete units of software by orderly invoking their execution with the goal of satisfying a set of user-defined specifications. The core idea is as old as the field of software engineering. In fact, as soon as software was complex enough to require artisan talent and engineering techniques, the intuition of using modular designs came about. Instead of writing code for every subtask, one could reuse parts of existing code, possibly resident remotely on a network. To make things simpler for the developer, the input/output syntax of these parts must be precisely specified in order to enable composition. These were the first steps in the direction of manual software composition and, with the subsequent advent of software services as units of invokable functionalities, of service composition.

1.1 A Parallel with the Automotive Industry

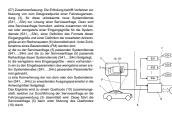
try and the process of driving a car will help mposition evolution. Since the first 'ride' of have been manually steering their vehicles . With the passing of time, more automation driver, such as synchronised gear shifting, Springer Nature Switzerland AG 2022 740, pp. 635–643, 2022.

	society and the job market in particular [5, 4]. There is con-				
	troversy on whether or not such systems manifest forms of				
	artificial intelligence. Researchers at Microsoft, for instance,				
	attribute signs of intelligence to the current fourth version				
	of Generative Pre-trained Transformer (GPT-4), which is in				
	development at the time of writing [5]. Some authors have				
	successfully solved Theory of Mind tasks using such tools.				
	Kosinski reports a success rate of 95% using GPT-4 in solv-				
	ing false-belief tasks. Other authors are more careful with				
(19) Deutsches Patent- und Markenamt	the excessive anthropomorphization of ChatGPT-like sys-				
	tems [6, 7]. What is sure is that the embedding of an LLM into E 10 2024 108 126 A1 2024 05.23				
(¹⁰⁾ D	E 10 2024 108 126 A1 2024.05.23 a system makes it a very powerful tool. Of interest to us in				
(12) Offenler	this editorial is the LLM capability to generate programs [8]				
(12) Offenleg	ungsachifftpotential impact on Service-Oriented Computing and				
	Applications.				
(21) Aktenzeichen: 10 2024 108 126.0 (22) Anmeldetag: 21.03.2024	(51) Int CL: The problem of automated service composition is central				
(43) Offenlegungstag: 23.05.2024	to the field of Service-Oriented Computing and Applica-				
(71) Anmelder:	(72) Effeder:				
Mercedes-Benz Group AG, 70372 Stuttgart, DE; Universität Stuttgart (Körperschaft des					
öffentlichen Rechts), 70174 Stuttgart, DE	FOR A RES available on a network is a potent way to build adap-				
(74) Vertreter: Dr. Weitzel & Partner Patent- und Rechtsamwälte	tive information systems. It is the idea that one can execute				
mbB, 89522 Heidenheim, DE	any task relying on multiple, loosely coupled services, pos-				
Mit Einverständnis des Anmelders offengelegte Anmeldung	gemäß § 31 A5s. 2 Ziffer 1 PallG				
Prüfungsantrag gemäß § 44 PatG ist gestellt.	Marco Aiello				
Die folgenden Angaben sind den vom Anm	elder eingereichen Unterformernoternoternass.uni-stuttgart.de				
(54) Bezeichnung: Verfahren zur Nutzung von unbekannt	en neuen Systemdiensten in einer Fahrzeuganwendung				

Ilche Georgievski

ilche.georgievski@iaas.uni-stuttgart.de

Service Computing Department, IAAS, University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, BW, Germany



Service Oriented Computing and Applications (2023) 17:233-238 https://doi.org/10.1007/s11761-023-00367-7

EDITORIAL

Service composition in the ChatGPT era

Marco Aiello¹ · Ilche Georgievski¹

Published online: 9 June 2023 © The Author(s) 2023

ChatGPT recently attracted vast attention in and outside the research community for its conversational abilities that mimic human ones exceptionally well. At the heart of systems like ChatGPT are Large Language Models (LLM). These models, rooted in deep neural networks, have the ability to predict the next textual token in a series of tokens based on statistical occurrences in extremely large data sets [1]. When the models are sufficiently big and well-tuned, one observes the "unreasonable effectiveness of data" [2] in how the system generates perfectly intelligible and believable sentoward havir tences. Such ability to have human-like conversations with a a web server software system is both stunning for the quality of the converenternrise inf sation and mind-blowing in terms of the potential impact on ation was als aty and the job market in particular [3, 4] There being strippe turned into a XML. Specif wiring of me describing se trations (BPI also propose such technole be discovered (publish-find The technol inspired man ing automate survey of the emerged from stand what a the signature the exchange mentation ar

sibly without prior knowledge of such services. This would guarantee that virtually any task can be executed by relying on third-party implementations and resources. Such a vision of automation is rooted in the fields of

software engineering and component-based software engineering. It emerged in a fervent moment of technological evolution. At the beginning of the century, the Internet was becoming pervasive, and the Web emerged as a central technology also for businesses. Every company was moving

composition

tems were in-

could not rea

would do, or:

the services h

and the sema

that made th

the road of

Compositio Prompto: An Architecture to Employ Large Language Models in Automated Service Computing

Robin D. Pesl¹(⊠), Carolin Mombrey^{1,3}, Kevin Klein^{1,2}, Denesa Zyberaj^{1,2}⁽ⁱ⁾, Ilche Georgievski¹⁽ⁱ⁾, Steffen Becker¹⁽ⁱ⁾, Georg Herzwurm¹⁽ⁱ⁾, and Marco Aiello¹⁽ⁱ⁾

¹ University of Stuttgart, Stuttgart, Germany {robin.pesl,marco.aiello}@iaas.uni-stuttgart.de ² Mercedes-Benz Group AG, Stuttgart, Germany ³ cellcentric GmbH & Co KG, Kirchheim/Teck-Nabern, Germany

Abstract. A classic, central Service-Oriented Computing (SOC) challenge is the service composition problem. It concerns solving a userdefined task by selecting a suitable set of services, possibly found at runtime, determining an invocation order, and handling request and response parameters. The solutions proposed in the past two decades mostly resort to additional formal modeling of the services, leading to extra effort, scalability issues, and overall brittleness. With the rise of Large Language Models (LLMs), it has become feasible to process semi-structured information like state-of-the-practice OpenAPI documentation containing formal parts like endpoints and free-form elements like descriptions. We propose Compositio Prompto to generate service compositions based on those semi-structured documents. Composition Prompto acts as an encapsulation of the prompt creation and the model invocation such that the user only has to provide the service specifications, the task, and which input and output format they expect, eliminating any manual and laborious annotation or modeling task by relying on already existing documentation. To validate our approach, we implement a fully operational prototype, which operates on a set of OpenAPIs a plain text task, and an input and output JSON schema as input and returns the generated service composition as executable Python code We measure the effectiveness of our approach on a parking spot booking case study. Our experiments show that models can solve several tasks. especially those above 70B parameters, but none can fulfill all tasks. Furthermore, compared with manually created sample solutions, the ones generated by LLMs appear to be close approximations.

Keywords: Automated service composition · Service discovery · Large language models · Code generation · Automotive services

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2025 W. Gaaloul et al. (Eds.): ICSOC 2024, LNCS 15405, pp. 276–286, 2025. https://doi.org/10.1007/978-981-96-0808-9_20

In the second se

Index Terms-Service Composition, LLM, Philosophy of Sci-

HERE TRANSACTIONS ON SERVICE COMPUTING

A Paradigm Shift in Service Research:

The Case of Service Composition

Marco Aiello, Senior Member, IEEE

Advance-Recent advancements in artificial intelligence, par-ticularly infrared version durantin, isolating arrive orange. In an advancements in Al that LLMs utter statements without having a model for truth, uttimately "buildating" [5]. Either way, the current advancements in Al thate takes our field and had of scientific this transformation, introducing new paradigms for automation and decidion-making. This paper cannues the revelving impose of LLMs on service composition, a fundamental problem in methodologie, and points michighest method complete and points and the set is transformation. The discussion paradise in adjust into emerging apportantials, relations are into exact the service in paradise in the set of LLMs of GPT in the third the discussion paradise in adjust into emerging apportantials, relation are interesting to a service composition in the rest of Ld-driven animation. The function for parating for a service composition in the rest of Ld-driven animation. The function for parating for a service composition in the rest of Ld-driven animation. The function for parating for a service composition in the rest of Ld-driven animation. The function for parating for a service composition in the rest of Ld-driven animation. The function for parating for a service composition of the rest of Ld-driven animation.

chitectures are changing between traditional approaches and

computing and the central problem of service composition, a field I have been active in for more than two decades.

Service Composition

The Case of Service Composition

Service Composition is the process of integrating independent looslely coupled services starting from a user request based on the ones available in the execution context. The services communicate over a network and are modular, allowing for flexible and dynamic composition. The *orchestrator* is responsible for coordinating the service composition.

Service Composition as AI Planning

- Artificial Intelligence Planning and Scheduling is a branch of Artificial Intelligence devoted to the study of algorithms and systems to empower intelligent agents with the ability to pursue their goals.
- Goal: a description of the state of the world to realise user request
- *Plan:* an algorithm that describes how to reach a goal state *a composition to orchestrate*
- Environment: a system the state of which can be sensed and changed by the planning actor APIs, service states, domain knowledge

Large Reasoning Models

- A neural network-based model optimized for multi-step logical and symbolic reasoning
- Trained on heterogeneous datasets: natural language, formal logic, math, code, and multimodal inputs
- Excels at structured problem-solving via in-context learning, chain-of-thought prompting, and tool augmentation
- Designed to perform algorithmic reasoning, planning, and hypothetical simulation
- May incorporate external memory, RAG, or tool use (e.g., calculators, search APIs, WolframAlpha)
- Good for: automated theorem proving, scientific discovery, decision support, etc.

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaee^{*†} Iman Mirzadeh^{*} Maxwell Horton Samy Bengio

Keivan Alizadeh
 Mehrdad Farajtabar

Apple

Abstract

Recent generations of frontier language models have introduced Large Reasoning Models (LRMs) that generate detailed thinking processes before providing answers. While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood. Current evaluations primarily focus on established mathematical and coding benchmarks, emphasizing final answer accuracy. However, this evaluation paradigm often suffers from data contamination and does not provide insights into the reasoning traces' structure and quality. In this work, we systematically investigate these gaps with the help of controllable puzzle environments that allow precise manipulation of compositional complexity while maintaining consistent logical structures. This setup enables the analysis of not only final answers but also the internal reasoning traces, offering insights into how LRMs "think". Through extensive experimentation across diverse puzzles, we show that frontier LRMs face a complete accuracy collapse beyond certain complexities. Moreover, they exhibit a counterintuitive scaling limit: their reasoning effort increases with problem complexity up to a point, then declines despite having an adequate token budget. By comparing LRMs with their standard LLM counterparts under equivalent inference compute, we identify three performance regimes: (1) lowcomplexity tasks where standard models surprisingly outperform LRMs, (2) medium-complexity tasks where additional thinking in LRMs demonstrates advantage, and (3) high-complexity tasks where both models experience complete collapse. We found that LRMs have limitations in exact computation: they fail to use explicit algorithms and reason inconsistently across puzzles. We also investigate the reasoning traces in more depth, studying the patterns of explored solutions and analyzing the models' computational behavior, shedding light on their strengths, limitations. and ultimately raising crucial questions about their true reasoning capabilities.

STOP ANTHROPOMORPHIZING INTERMEDIATE TOKENS AS REASONING/THINKING TRACES!

Subbarao Kamb	hampati	Kaya Stech	ly Karthik Valmeekam	Lucas Saldyt	Siddhant Bhambri				
Vardhan Palod	Atharva Gundawar		Soumya Rani Samineni	Durgesh Kalwar	Upasana Biswas				
School of Computing & AI									

Arizona State University

ABSTRACT

Intermediate token generation (ITG), where a model produces output before the solution, has been proposed as a method to improve the performance of language models on reasoning tasks. These intermediate tokens have been called "reasoning traces" or even "thoughts" – implicitly anthropomorphizing the model, implying these tokens resemble steps a human might take when solving a challenging problem. In this paper, we present evidence that this anthropomorphization isn't a harmless metaphor, and instead is quite dangerous – it confuses the nature of these models and how to use them effectively, and leads to questionable research.

Large Action Models

- A parameter-rich neural policy model trained to map from high-dimensional observations and goals to action distributions
- Leverages transformer-based architectures for sequence modeling of action trajectories
- Operates over state-action-return triples (or variations) for temporal credit assignment and long-horizon planning
- Trained via offline reinforcement learning, behavior cloning, or trajectory-level supervision from expert demonstrations or synthetic data
- Can ingest multimodal inputs and output low-level control signals or symbolic action commands
- Supports zero-shot generalization across tasks via goal-conditioning, prompting, or language grounding
- Frequently deployed in embodied agents, robotic manipulation, navigation, game environments, and tool use contexts

Language Models (LMs)

Large Language Models (LLMs)

- Large-scale knowledge of language patterns
- Understands natural language
- Excels at generating coherent text
- Efficient for translation Q&A, etc.
- No deliberate, iterative reasoning
- No direct interaction with environment
- Tends to hallucinate

GPT-3.5, GPT-4, Claude, LLaMa, BERT, Qwen, Grok, Gemini 1

Examples

Legend

Strengths Limitations

Large Reasoning Models (LRMs)

- Understands natural language
- Deep, explicit reasoning
- Strong at planning and problem-solving
- Often uses explicit structure of reasoning

- Typically does not act on the environment - May require extensive computational resources

- Can be slower

OpenAl o1/o3, DeepSeek R1, Gemini 2.0, QwQ

Large Action Models (LAMs)

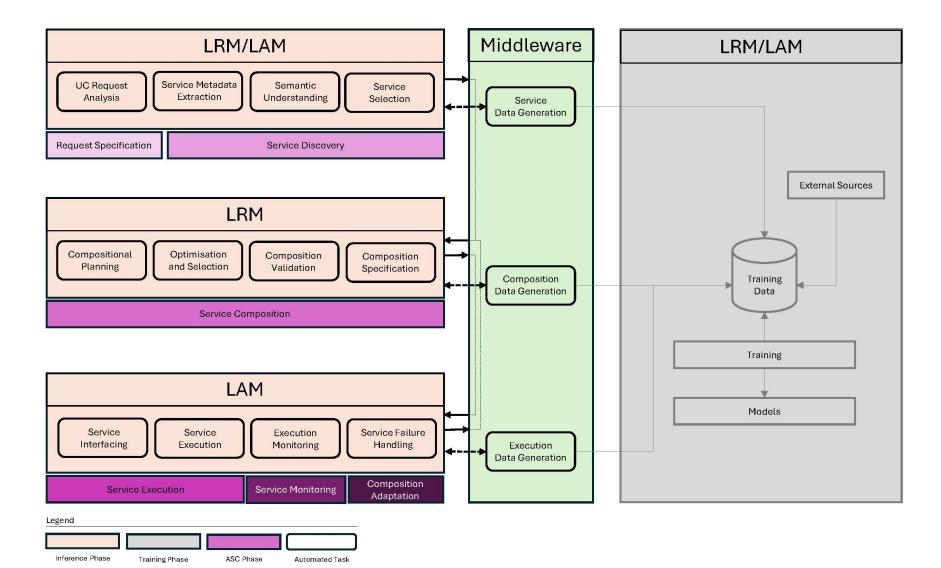
- Understands multimodal inputs
- Directly executes actions in environments
- Integrates sensing with action outputs
- Can handle tasks in real time

- Often weak at (high-level) reasoning

- Requires specialised data (e.g., action logs)

- Can be difficult to train for safety and reliability

Google RT-1/RT-2, DeepMind Gato, Rabbit R1, CogAgent, ScreenAI, xLAM



Concluding remarks

Reflection on Initial Steps

- LLM, LRM, LAM can cover various aspects of Service Composition
- Promising technologies with some known and yet unknown limitations
- See you at SummerSOC 2026 for more



Thank you! Vielen Dank! Grazie! Merci! Bedankt!



Marco Aiello

多謝!

E-Mail marco.aiello@iaas.uni-stuttgart.de Telefon +49 (0) 711 685-88471 www.iaas.uni-stuttgart.de/en/department-service-computing/ or aiellom.it

Universität Stuttgart Service Computing Dep — IAAS Universitätsstr. 38, 70569 Stuttgart

