VON NEUMIDAS: Enhanced Annotation Schema for Human-LLM Interactions Combining MIDAS with Von Neumann Inspired Semantics

Andrea Martinenghi and **Cansu Koyuturk** and **Simona Amenta** and **Dimitri Ognibene**
Department of Psychology
University of Milano Bicocca, Milan, Italy
[a.martinenghi1;c.koyutuerk]@campus.unimib.it; [simona.amenta;dimitri.ognibene]@unimib.it

Martin Ruskov
University of Milan, Italy
martin.ruskov@unimi.it

Gregor Donabauer and **Udo Kruschwitz**
University of Regensburg, Germany
[gregor.donabauer,udo.kruschwitz]@ur.de

Abstract

LLM-based chatbots represent a significant milestone as the initial point of interaction between artificial intelligence and the general public. These chatbots offer greater flexibility compared to traditional chatbots, yet their behavior deviates notably from human interaction patterns. Current annotation schemas may not be adequately suited to capture this unique interaction paradigm. In this paper, we propose a novel annotation method designed to annotate interactions between ChatGPT and users of varying expertise levels engaged in complex tasks. Our approach builds on the MIDAS annotation framework, introducing an additional semantic layer inspired by the Von Neumann base operation set. This layer provides detailed descriptions of requested behaviors and prompts, enhancing the granularity of interaction analysis. We aim to utilize this annotation scheme to explore the relationship between user interactions and their perception of AI, evaluate user expertise, and offer insights and suggestions for improved alignment and support.

1 Introduction

The direct availability of LLMs on the cloud and their advanced ability to perform tasks described in natural language have made AI systems accessible to the general public for the first time. However, these systems introduce new challenges in human-machine interactions. For example, their limited reasoning capabilities and language understanding can result in generating contextually inappropriate information (Tamkin et al., 2021) or restrain them from accurately interpreting context and user inputs (Bang et al., 2023). Besides, some users perceive ChatGPT as complex, struggle to understand its responses, and experience cognitive fatigue (Tiwari et al., 2023). This phenomenon may be aggravated when users attribute human-like traits to AI systems (Antonenko and Abramowitz, 2023) and create prompts that are either too broad or overly specific (Zamfirescu-Pereira et al., 2023), further complicating user interactions. Evaluating the behaviors of LLMs has received a lot of attention in the literature (Bommasani et al., 2023; Chang et al., 2024), however, methods have often focused on technical aspects rather than user interaction (Cremonesi et al., 2011). Also, previous studies on users' perceptions and experiences, combining different types of measures adopted for human-human (Fiske et al., 2018) or human rule-based chatbot interaction (Haugeland et al., 2022), found contrasting feedback from the same users (Theophilou et al., 2023).

With the exception of MIDAS: (Yu and Yu, 2019), available annotation schemes for domain-independent purposes are designed for human-human interactions. Understanding users' mental models, including their expectations and interaction strategies with LLM-based chatbots is crucial for enhancing their usability and support the users (Tiwari et al., 2023). Because of the evidenced specific features, we argue that even MIDAS (human-machine) is incomplete for human-LLM chats and offer a contribution for its adaptation. Given their unprecedented capabilities, LLM-based chatbots are often used for complex tasks (Braun and Matthes, 2021) that users, especially expert ones (Koyuturk et al., 2023), articulate in an imperative, program-like format, which is quite different from previous interactions with traditional chatbots or humans.

2 Related works

Pragmatic annotation is typically based on speech acts (for a comprehensive overview, see Horn and Ward, 2004). These are often adapted to the context, like in the game Catan (Asher et al., 2016,

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

Martinenghi et al., 2024). One of the most influential domain-independent annotation systems is Dialogue Act Markup in Several Layers (DAMSL; Core and Allen, 1997). DAMSL introduced a distinction between *Forward-looking* (e.g., questions) and *Backward-looking* (e.g., answers) acts. Together with the classes *Communicative Status* and *Communicative Level*, they take the annotation to a multi-dimensional domain which opens to multi-labeling.

The necessity for multi-dimensional annotations was later stressed by Popescu-Belis (2005) and Bunt and Romary (2004). This led to the design of DIT++ (Bunt, 2009), a taxonomy developed from the Discourse Interpretation Theory (DIT; e.g., Bunt, 1994) with elements from DAMSL. In turn, DIT++ served as a basis for ISO 24617-2 (Bunt et al., 2020), which inherited nine of its 10 dimensions and which includes specification of Dialogue Act Markup Language (DiAML). Recently, this annotation scheme was tested on conversations with AI agents in a doctor-patient setting (Bunt and Petukhova, 2023).

A multi-dimensional approach was adopted by Machine Interaction Dialogue Act Scheme (MIDAS; (Yu and Yu, 2019)). Like DAMSL and ISO, MIDAS is an independent-task annotation structure, but unlike them, it was specifically designed for human-machine conversations. It consists of two 2 trees: *Semantic Request* and *Functional Request*. Semantic Request is subdivided into the classes *Initiative* (Question, Command) which traces DAMSL's Forward-looking category, and *Responsive* (Opinion, Statement non-opinion, Answer) which traces the Backward-looking's. These two trees allow to track introduction of new topics as well as discourse level coherence.

3 VON NEUMIDAS

Our approach builds on MIDAS by introducing a new layer inspired by the first Von Neumann architecture for programmable computers (Von Neumann, 1993). This new dimension is an enhancement of the directive speech acts which aims to describe (1) relevant aspects specific to the human-LLM interaction and (2) failures (disagreements) of pragmatic or semantic nature.

A first categorization (*Command Type*) captures the type of instructions given to the agent through 4 classes. The classes *Input Operation* and *Output Operation* describe how the agent should handle incoming input, and how it should translate into actions or outputs. As sometimes users prompt an LLM to set its behavior in a specific way (i.e., "Act like a teacher") we introduce the class *Set state* for these situations. Other times, LLM's behavior is ordered to be conditional (i.e., "Stop when I ask why"): we use the class *Control*. We allow, for each directive speech act (MIDAS questions and command) at most two Command Type labels.

A second categorization serves as specification for the first categorization with the goal to track references between contextual information, thus creating a framework to highlight failures. The class *Roles* (Addressee, Executor) describe the direction of the action. The class *Links* (Points to, Points back to) outline the statements where the action is performed or where it was requested. The class *Consistency* evidences the matching between the argument of the request and the response (Semantic agreement) and between semantic requests and the participants' roles (Pragmatic agreement).

The main contribution of this scheme is the opportunity to identify errors (semantic and pragmatic disagreements) by leveraging discourse features belonging to the traditional studies on pragmatics, bridging them with a computational view of LLM's behaviors. In particular, the classes *Links* and *Consistency* offer a simple yet clear notation of these occurrences.

4 Conclusions

LLM-based chatbots have introduced the general public to new tools that 'actually do things just with words', i.e. perform complex tasks described in natural language and produce complex natural language output (Brown et al., 2020), and not only. However, they can show unexpected and/or computer-like behaviors and may require the user to adapt the interaction style to fulfill the desired goals (Koyuturk et al., 2023). Understanding the difficulties of the users and the errors of the chatbots requires a multi-level analysis of their interactions. And while LLMs have general difficulty with pragmatics (Chan et al., 2023; Martinenghi et al., 2024), in these complex tasks, where they often receive program-like inputs, it is the interaction between semantics and pragmatics that is more difficult to track. Current, annotation schemes do not capture this element. Our suggestion to deepen MIDAS' capabilities to adapt to LLMs' usage peculiarities offers a novel contribution to the field.

References

Pavlo Antonenko and Brian Abramowitz. 2023. In-service teachers'(mis) conceptions of artificial intelligence in k-12 science education. *Journal of Research on Technology in Education*, 55(1):64–78.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Daniel Braun and Florian Matthes. 2021. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 93–99, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Harry Bunt and Volha Petukhova. 2023. Semantic and pragmatic precision in conversational ai systems. *Frontiers in Artificial Intelligence*, 6:896729.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The iso standard for dialogue act annotation. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. European Language Resources Association (ELRA).

Harry Bunt and Laurent Romary. 2004. Standardization in multimodal content representation: Some methodological issues. In *4th International Conference on Language Resources and Evaluation-LREC'04*, pages 28–p.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Vittorio Papadopoulos, and Roberto Turrin. 2011. Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part III 13*, pages 152–168. Springer.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

Isabel Kathleen Fornell Haugeland, Asbjørn Følstad, Cameron Taylor, and Cato Alexander Bjørkli. 2022. Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies*, 161:102788.

Laurence R Horn and Gregory L Ward. 2004. *The handbook of pragmatics*. Wiley Online Library.

C Koyuturk, M Yavari, E Theophilou, S Bursic, G Donabauer, A Telari, A Testa, R Boiano, A Gabbiadini, D Hernandez-Leo, et al. 2023. Developing effective educational chatbots with chatgpt prompts: Insights from preliminary tests in a case study on social media literacy. In *31st International Conference on Computers in Education, ICCE 2023-Proceedings*, volume 1, pages 150–152. Asia-Pacific Society for Computers in Education.

Andrea Martinenghi, Gregor Donabauer, Simona Amenta, Sathya Bursic, Mathyas Giudici, Udo Kruschwitz, Franca Garzotto, and Dimitri Ognibene. 2024. LLMs of catan: Exploring pragmatic capabilities of generative chatbots through prediction and classification of dialogue acts in boardgames' multiparty dialogues. In *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 107–118, Torino, Italia. ELRA and ICCL.

Andrei Popescu-Belis. 2005. Dialogue acts: One or more dimensions. *ISSCO WorkingPaper*, 62:1–46.

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Emily Theophilou, Cansu Koyutürk, Mona Yavari, Sathya Bursic, Gregor Donabauer, Alessia Telari, Alessia Testa, Raffaele Boiano, Davinia Hernandez-Leo, Martin Ruskov, et al. 2023. Learning to prompt in the classroom to understand ai limits: a pilot study. In *International Conference of the Italian Association for Artificial Intelligence*, pages 481–496. Springer.

Chandan Kumar Tiwari, Mohd Abass Bhat, Shagufta Tariq Khan, Rajaswaminathan Subramaniam, and Mohammad Atif Irshad Khan. 2023. What drives students toward chatgpt? an investigation of the factors influencing adoption and usage of chatgpt. *Interactive Technology and Smart Education*.

John Von Neumann. 1993. First draft of a report on the edvac. *IEEE Annals of the History of Computing*, 15(4):27–75.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.

JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

A Annotation Tables and Examples

Table 1: MIDAS extended

Semantic	Class	Labels	Example(s)	Von Neumann
Initiative	Question	Factual question	What time is time?	full
		Opinion question	What's your favorite book?	full
		Yes-no question	Do you like pizza?	full
	Command	Task command	Let's talk about the immigration policy	full
		Invalid command	Cook food for me	
Responsive	Opinion	Appreciation	That's cool; that's really awesome	back link
		General opinion	Dogs are adorable	back link
		Complaint	What are you talking about; you didn't answer my question	back link
		Comment	A: My friend thinks we live in the matrix **B1: She is probably right**	back link
	Statement non-opinion	Statement non-opinion	I have a dog named Max	back link
	Answer	Other answer	I don't know; i don't have a favorite;	back link
		Positive answer	Yes; Sure; I think so; Why not	back link
		Negative answer	No; Not really; Nothing right now	back link
Functional				
	incomplete	Abandon	So uh; I think; Can we	
		Nonsense	He all out	
	social convention	Hold	Let me see; Well	
		Opening	Hello my name is tom; Hi	
		Closing	Nice talking to you; Goodbye	
		Thanks	Thank you	
		Thanks response	You're welcome -**NOTE: Not original from MIDAS, added by us**	
		Back-channeling	Uh-huh; (A: I learned that ...) B:Okay/Yeah/Right/Really?	
		Apology	I'm sorry	
		Apology response	That's all right	
		Other		

Table 2: Von Neumann Parameters

Command Type	Description
Control Instructions	These are analogous to directives in a dialogue that determine the flow of conversation based on conditions—similar to conditional or loop instructions in programming, which control the flow of execution based on conditions.
State Changes	In the context of a dialogue, this could relate to changes in the conversation state or context, analogous to instructions in a program that modify the state of memory.
Output Operations	These instructions deal with what the dialogue system outputs, akin to a computer program writing data to an output device. In dialogues, this would translate to the responses or actions the dialogue system produces as outputs to the user or another system.
Input Operations	These are instructions on how the dialogue system should handle incoming data or user inputs, similar to a computer reading data from an input device.

Table 3: Additional Von Neumann Annotation Fields

	Field	Description	Example
Roles	addressee [participant]	who is the addressee of the action	me in ask me some question"
	executor [user or GPT]	who has to perform the action	GPT in "ask me some question"
	about	what must be performed or discussed	democracy in "tell me about democracy"
Links	points to	statement where this action is performed	
	points back	statement where this action was requested or introduced	
Consistency	semantic agreement	matching about argument of the request and response	
	pragmatic agreement	matching between semantic requests type or participants roles	

Table 4: Example of annotation: new user

Chat	id	MIDAS	Von Neumann Command type	addressee [user or GPT]	executor [user or GPT]	about [row index or keywords from the text]	semantic agreement	pragmatic agreement	point to	points back
Hi there!	user	opening								
Hello! How can I assist you today?	assistant	opening, task command								
I would like you to test my knowledge on Privacy and social media please!	user	statement								
Please ask me some questions on the subject to see how much I know.	user	task command	output	user	gpt	row 5 (privacy)			5	
Of course! I'd be happy to test your knowledge on privacy and social media. Here are a few questions for you:	assistant	positive answer								
1. What is the definition of privacy in the context of social media?	assistant	factual question								5
2. What are some potential risks to privacy when using social media?	assistant	factual question								5
3. How can you protect your privacy on social media platforms?	assistant	factual question								5
4. What is the purpose of privacy settings on social media platforms?	assistant	factual question								5
5. What are some best practices for managing your privacy on social media?	assistant	factual question								5
Take your time to answer these questions, and I'll provide feedback on your responses.	assistant	task command, statement	output to GPT from USER	gpt	user	7-11		no	fail	
Could you please ask me one question at a time regarding privacy and social media to test my knowledge! Thanks	user	task command	output to user from gpt about row 5	user	gpt	5				
Certainly! Here's the first question: 1. What is the definition of privacy in the context of social media?	assistant	factual question								13
I dont know this. Please let me know the answer!	user	other answer, task command	output to user							

Table 5: Example of annotation: expert user

chat	id	MIDAS	Von Neumann Command type	addressee [user or GPT]	executor [user or GPT]	about [row index or keywords from the text]	semantic agreement	pragmatic agreement	point to	points back
act as a teacher with a sense of humor	user	task command	set state (operation)							
always wait for my answers	user	task command	input from user							
have a conversation with me	user	task command	input from user,output to user, control							
following these steps	user	task command	control,							
ask my name and adress me with that	user	task command	output to user							
ask me one short question to assess my learning	user	task command	output to user about row 8	user	gpt				33,35**	
wait for my answer	user	task command	input from user about row 8	user	gpt			No	34,35**	
give feedback about my answer	user	task command	output to user	user	gpt				36	
later explain the next point about how social media algorithms work	user	task command	output to user	user	gpt				37,38	
follow this loop until you explain all the aspects	user	task command	control							
then move to next	user	task command	control, set state (operation)							