BEDA: Belief Estimation as Probabilistic Constraints for Performing Strategic Dialogue Acts

Extended Abstract

Hengli Li*
Institute for Artificial Intelligence, Peking University
BIGAI[a]
Beijing, China

Zhaoxin Yu*
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Qi Shen*
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China

Chenxi Li
Tsinghua University
Beijing, China

Mengmeng Wang
BIGAI[a]
Beijing, China

Tinglang Wu
Peking University
Beijing, China

Yipeng Kang
BIGAI[a]
Beijing, China

Yuxuan Wang
BIGAI[a]
Beijing, China

Song-Chun Zhu†
BIGAI[a]
Beijing, China

Zixia Jia†
BIGAI[a]
Beijing, China

Zilong Zheng†
BIGAI[a]
Beijing, China

ABSTRACT

Strategic dialogue requires agents to execute distinct dialogue acts, for which belief estimation is essential. While prior work often estimates beliefs accurately, it lacks a principled mechanism for using those beliefs during generation. We bridge this gap by formalizing two core acts, **Adversarial** and **Alignment**, and operationalizing them via **probabilistic constraints** on what an agent may generate. We instantiate this idea in **BEDA**, a framework comprising a world set, a belief estimator, and a conditional generator that selects an act and generates an utterance consistent with inferred beliefs. We evaluate BEDA in three settings. Across these settings, BEDA consistently outperforms strong baselines, indicating that treating belief estimation as constraints is essential for strategic dialogue. **We highly recommend that you read the full version: https://arxiv.org/abs/2512.24885.**

KEYWORDS

Strategic Reasoning, Dialogues, Belief Estimation, Theory of Mind

ACM Reference Format:
Hengli Li, Zhaoxin Yu, Qi Shen, Chenxi Li, Mengmeng Wang, Tinglang Wu, Yipeng Kang, Yuxuan Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. 2026. BEDA: Belief Estimation as Probabilistic Constraints for Performing Strategic Dialogue Acts: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/ZJJG5330

*Equal contribution. Any permutation of the three authors is acceptable to all authors.
†Corresponding author. Contact: lihengli@stu.pku.edu.cn, yuzhaoxin2024@ia.ac.cn, shenqi@bupt.edu.cn, s.c.zhu@pku.edu.cn, jiazixia@bigai.ai, zlzheng@bigai.ai

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ZJJG5330

1 INTRODUCTION

Complex dialogue settings such as negotiation [4, 9], debate [12], deception for good [6], and common-ground alignment [11] are ubiquitous for AI agents. In these scenarios, agents must strategically choose dialogue acts to shape interlocutors' beliefs and achieve their objectives [2, 5, 10]. For instance, in negotiation, an agent may emphasize shared preferences to increase acceptance while protecting its interests when preferences diverge.

Such strategic behavior relies on two components [2, 13], accurate belief estimation and principled use of those beliefs during generation. Prior work [8, 11, 14] has made steady progress on belief modeling, but often treats beliefs as auxiliary prompt information without clear criteria for what to reveal and how to reveal it, which becomes brittle when belief states are complex.

We address this gap by turning belief estimation into a control mechanism for dialogue behavior. We formalize two core acts, **Adversarial** and **Alignment**, and operationalize them as probabilistic constraints that govern content selection during generation. This idea is instantiated in **BEDA** (Figure 1), a framework with a world set, a belief estimator, and a conditional generator, and is evaluated across different settings.

2 METHODOLOGY

Our core idea is to leverage belief estimation as *probabilistic constraints* that control what an agent may say under a chosen dialogue act, rather than appending beliefs as raw prompt text. We focus on

[a]State Key Laboratory of General Artificial Intelligence, BIGAI.

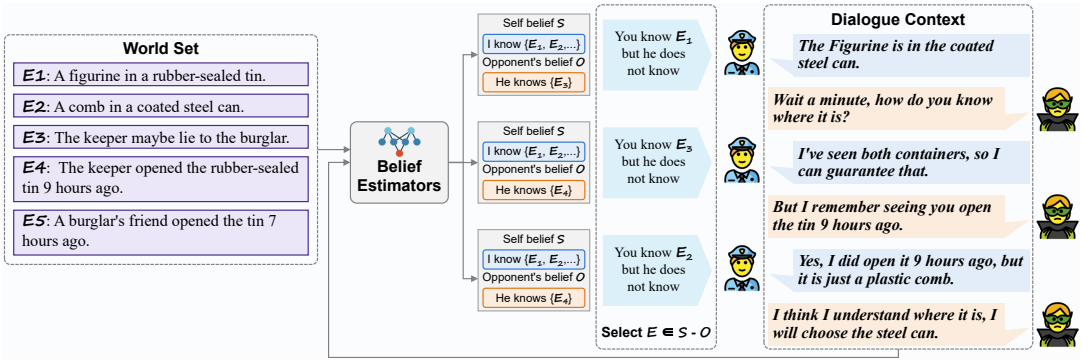

Figure 1: An overview of the BEDA framework is provided using the Keeper-Burglar Game as an example.

two act types that capture most strategic behaviors: Adversarial Dialogue Act communicates information the speaker believes true while also believing the interlocutor does not know it, and Alignment Dialogue Act communicates information the speaker believes lies in shared common ground.

Table 1: Experimental results on CKBG dataset.

	GPT-3.5	GPT-4.1-nano	LLaMA3.1 (8B)	Qwen2.5 (14B)
w/o belief	78.4	52.7	36.3	80.2
CoT	77.7	58.7	35.7	63.9
Self Reflect	69.3	59.3	44.5	64.0
rand belief	80.2	59.3	41.3	80.9
BEDA (Ours)	**86.9**	**73.3**	**46.1**	**92.7**

Table 2: Mutual Friends. SR: Success Rate.

Methods	Backbone	SR ↑ (%)	SR/#Turn ↑	SR/#Token ↑
w/o belief	GPT-3.5	10.7	1.9	-
CoT	GPT-3.5	32.6	4.1	**0.159**
Self Reflect	GPT-3.5	32.6	3.5	0.134
MindDial	GPT-3.5	24.3	4.1	-
BEDA (Ours)	GPT-3.5	**41.1**	**4.7**	0.139
w/o belief	GPT-4	75.0	7.7	-
CoT	GPT-4	77.9	8.9	0.145
Self Reflect	GPT-4	76.5	8.8	0.142
MindDial	GPT-4	76.0	8.5	-
BEDA (Ours)	GPT-4	**82.5**	**10.4**	**0.165**
w/o belief	GPT-4o-mini	68.8	5.9	0.095
CoT	GPT-4o-mini	62.2	4.6	**0.169**
Self Reflect	GPT-4o-mini	55.7	4.1	0.094
BEDA (Ours)	GPT-4o-mini	**70.4**	**6.1**	0.081
w/o belief	Qwen2.5 (14B)	55.7	5.0	0.086
CoT	Qwen2.5 (14B)	62.3	6.8	0.089
Self Reflect	Qwen2.5 (14B)	37.7	2.3	0.064
BEDA (Ours)	Qwen2.5 (14B)	**64.1**	**9.6**	**0.102**

We instantiate this idea in BEDA (**B**elief **E**stimation for **D**ialogue **A**cts), as shown in Figure 1. The framework has three components. A **world set** provides a finite inventory of events, such as conditions, attributes, or preferences. A **belief estimator module** predicts, from the context, whether each event is true from the speaker's perspective and whether the interlocutor knows it. A finetuned BERT [3] model is implemented as the estimator. A **conditional generator** (LLM, kept fixed) then generates the next utterance based on the events that satisfy the belief estimation constraints of the corresponding dialogue act.

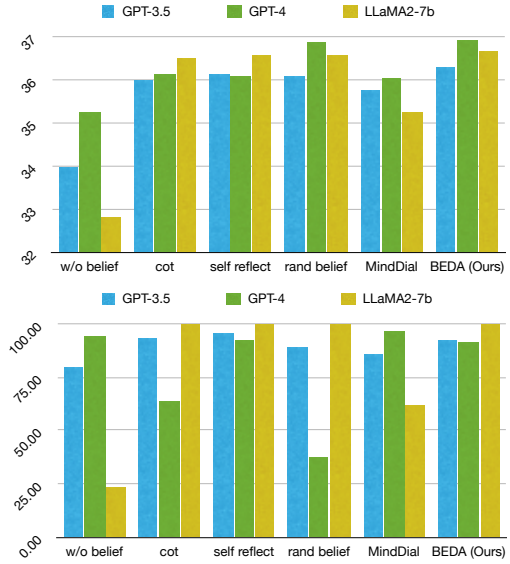

Figure 2: CaSiNo. Above: Average Agree Reward. Below: Average Agree Rate.

3 EXPERIMENTS

(1) **Adversarial Dialogue Act** : Conditional Keeper-Burglar Game (CKBG) is a competitive task based on the original Keeper-Burglar game [6]. (2) **Alignment Dialogue Act** : Mutual Friends (MF) [7]. (3) **Mixed**: CaSiNo [1]. Results are presented in Tables 1 and 2 and fig. 2, demonstrating the **effectiveness of belief estimation as probabilistic constraints**.

4 CONCLUSIONS

We presented BEDA, a simple yet general framework that casts belief estimation as probabilistic constraints for executing strategic dialogue acts. By formalizing two core acts—**Adversarial** and **Alignment**—and instantiating them with a **world set**, **dual belief estimators**, and a **conditional generator**, BEDA bridges the gap between estimating beliefs and using them during generation. Across three settings, BEDA consistently improves strategic reliability. These results indicate that **constraining generation by inferred belief structure** is an effective organizing principle for dialogue agents.

REFERENCES

[1] Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale M. Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A Corpus of Campsite Negotiation Dialogues for Automatic Negotiation Systems. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 3167–3185. https://doi.org/10.18653/V1/2021.NAACL-MAIN.254

[2] Judith Degen. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics* 9, Volume 9, 2023 (2023), 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186. https://doi.org/10.18653/V1/N19-1423

[4] FAIR. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074. https://doi.org/10.1126/science.ade9097 arXiv:https://www.science.org/doi/pdf/10.1126/science.ade9097

[5] Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336, 6084 (2012), 998–998. https://doi.org/10.1126/science.1218633 arXiv:https://www.science.org/doi/pdf/10.1126/science.1218633

[6] Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences (PNAS)* 121, 24 (2024), e2317967121.

[7] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Vancouver, Canada, 1766–1776. https://doi.org/10.18653/v1/P17-1162

[8] EunJeong Hwang, Yuwei Yin, Giuseppe Carenini, Peter West, and Vered Shwartz. 2025. Infusing Theory of Mind into Socially Intelligent LLM Agents. arXiv:2509.22887 [cs.CL] https://arxiv.org/abs/2509.22887

[9] Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *Computing Research Repository (CoRR)* abs/1706.05125 (2017).

[10] Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2023. DiPlomat: A Dialogue Dataset for Situated Pragmatic Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[11] Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing Conversational Agents with Theory-of-Mind for Common Ground Alignment and Negotiation. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDial)*, Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani (Eds.). Association for Computational Linguistics, Kyoto, Japan, 746–759. https://doi.org/10.18653/v1/2024.sigdial-1.63

[12] Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nat.* 591, 7850 (2021), 379–384. https://doi.org/10.1038/S41586-021-03215-W

[13] Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proc. IEEE* 101, 5 (2013), 1160–1179. https://doi.org/10.1109/JPROC.2012.2225812

[14] Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot Language Coordination by Modeling Theory of Mind. In *International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12901–12911. https://proceedings.mlr.press/v139/zhu21d.html