ALICEA: Accountable Collaborative Conversations with Embodied Agents

Shiwali Mohan, Yonghui Fan, Sachin Grover, Wiktor Piotrowski

Executive Summary

ALICEA is an embodied conversational reasoning system designed for accountable, natural language interactions with an embodied AI agent that pursues goal-oriented tasks in complex environments (e.g., a ground robot participating in a human-robot rescue mission). Current conversational systems are *frictionless*: they represent conversational reasoning knowledge as a probability distribution over the next token prediction and generate content regardless of correctness, applicability, or consequentiality. ALICEA employs statistical inference and deliberate, logical reasoning while communicating with a human partner. It is accountable; it employs embodied world reasoning to ensure its responses and behavior are correct, applicable to the current situation, and aligned with causal reasoning. It is built on three innovations: (1) System 1/System 2 organization of conversational knowledge inspired by Kahnemann's dual process model [1] that balances the need for flexible natural language interaction with deliberate, contextual reasoning; (2) An intentional discourse model based on Collaborative Discourse Theory that introduces friction in human-agent dialog when interactions become incongruent with the current situation to encourage further reflection and resolution; and (3) Novel methods for behavioral and situational accountability based on causal reasoning (planning theory and Bayesian inference) that recognize when linguistic beliefs become incongruent with what currently exists or is plausible in the world. ALICEA will advance human language technologies by providing a framework, algorithms, and a taxonomy for intentional, collaborative human-agent/robot interaction such that interactions are aligned with embodied world reasoning.

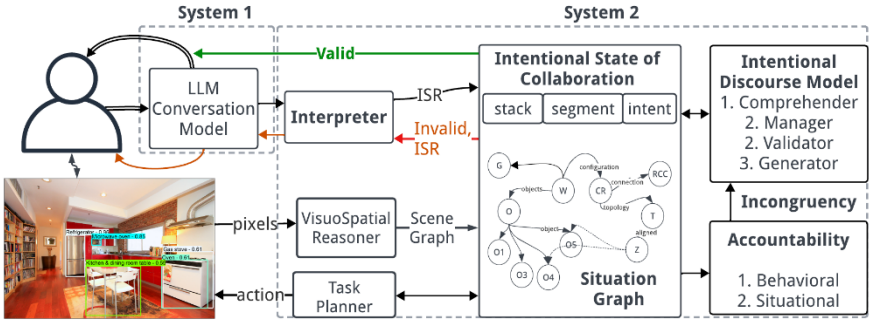


Figure 1. *A notional diagram of the ALICEA framework.*

Expected Scientific and Technical Impact

ALICEA (*Figure 1*) augments an LLM-based conversational model (e.g., DollyV2 [2]) with deliberate and explicit reasoning about the human-agent shared situated task context. We will:

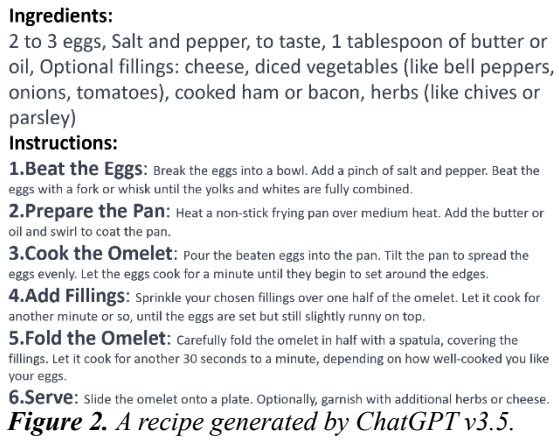
- Develop System 1/System 2 organization of conversational reasoning (Section 3.1). Inspired by Kahnemann's dual process model [1], this organization of knowledge enables ALICEA to generate responses/utterances in an accountable fashion, such that they are correct, meaningful, informative, and advance human-agent task collaboration. Utterances generated by ALICEA System 1 - an LLM conversation model – are further evaluated by explicit and deliberate reasoning in ALICEA System 2 about the current situation and the state of task execution.

- Adopt an intentional discourse model (built upon Collaborative Discourse Theory [3]–[5], Section 3.2)**A**. This model augments LLM conversational models, enabling ALICEA to explicitly represent each partner's intent, perspective, & salient information pertinent to task collaboration. With it, ALICEA can introduce friction in a dialog when its and its partner's beliefs (or assumptions) become incongruent to encourage further introspection and reflection.
- *Invent techniques to ensure behavioral and situational accountability in System 2* (Section 3.3). Causal reasoning mechanisms built upon planning theory and Bayesian inference enable ALICEA to identify incongruencies in conversational beliefs and the current situational and behavioral (task-related) beliefs. Additionally, these techniques reveal costs, implicit assumptions, and generate information that is useful o introduce appropriate friction in human-machine dialog.

We will develop a proof-of-concept of ALICEA in a simulated embodied AI domain (e.g., AI-Thor [6], Habitat [7], Webots [8]) supported by several publicly available datasets for home-related tasks (e.g., ALFRED [9], AlexaTeach [10]). We envision ALICEA as a domain-independent framework that implements general processes for embodied conversational reasoning. Additionally, we will develop a taxonomy of conversational friction (Section 3.4) that identifies a set of incongruencies that can arise in a conversational embodied AI system and corresponding policies for resolving those by advancing human-agent dialog. ALICEA will be 20%-40% more successful in task execution when compared to frictionless baseline agents operating in a carefully designed set of collaborative task execution experiments.

Technical Approach & Plan

Consider a human-embodied agent task collaboration scenario in a kitchen-like environment where the human and the AI system are embedded in a shared situation and interact using natural language. The human asks the robot, *"Can you make omelets for breakfast?"* In response to this request, *frictionless* LLM-based agents [11] and [12] query a language model to access a recipe and follow the actions described in it. However, operationalizing task procedures written for humans in the world is significantly more challenging. Consider the recipe generated by ChatGPT in *Figure 2*. First, the recipe yields one omelet and therefore, by following this recipe, the agent makes an *implicit assumption* that only a single omelet is desired, which may not be sufficient for what the human partner requested. Second, the recipe includes optional fillings, an *unknown criterion* for the task, that the agent must inquire about to execute successfully. Finally, executing this recipe in an embodied context requires the agent to ground linguistic instructions in its environment, find relevant objects and materials, and apply a skill to execute the action described. This process itself may pose challenges. For example, the agent must first locate eggs. It asks ChaptGPT, *"Where are eggs in the kitchen?"* which suggests that



Ingredients:

2 to 3 eggs, Salt and pepper, to taste, 1 tablespoon of butter or oil, Optional fillings: cheese, diced vegetables (like bell peppers, onions, tomatoes), cooked ham or bacon, herbs (like chives or parsley)

Instructions:

1. **Beat the Eggs**: Break the eggs into a bowl. Add a pinch of salt and pepper. Beat the eggs with a fork or whisk until the yolks and whites are fully combined.
2. **Prepare the Pan**: Heat a non-stick frying pan over medium heat. Add the butter or oil and swirl to coat the pan.
3. **Cook the Omelet**: Pour the beaten eggs into the pan. Tilt the pan to spread the eggs evenly. Let the eggs cook for a minute until they begin to set around the edges.
4. **Add Fillings**: Sprinkle your chosen fillings over one half of the omelet. Let it cook for another minute or so, until the eggs are set but still slightly runny on top.
5. **Fold the Omelet**: Carefully fold the omelet in half with a spatula, covering the fillings. Let it cook for another 30 seconds to a minute, depending on how well-cooked you like your eggs.
6. **Serve**: Slide the omelet onto a plate. Optionally, garnish with additional herbs or cheese.

Figure 2. *A recipe generated by ChatGPT v3.5.*

eggs are typically kept in the fridge. When it opens the fridge, it doesn't find any, a *situational inconsistency*, which must be addressed through further interactions with the human (who remembers that the eggs are in the shopping bag).

We define *incongruencies* – implicit assumption, unknown criterion, situational inconsistency, etc. - in an embodied conversational agent as divergent, incomplete, or inconsistent situational & task-related beliefs that lead to misalignment with those of its human partner. They can arise for two reasons. First, the embodied agent's situational, task-related beliefs may differ from what are implicitly encoded in an LLM as a part of its statistical training paradigm. Such *intrinsic* incongruency is natural in conversational embodied agents – LLMs are trained over large corpora of text, agnostic of any specific situational contexts. Second, both parties operate with partial observability; the human may not fully know the complete environmental state or fully understand the agent's capacities; the agent may not represent the situation and task goals in a way that is aligned with how a human understands them. Such *extrinsic* incongruency is natural as well – in any collaborative scenario, parties operate with an incomplete understanding of the situation as well as with each other's capabilities, intentions, and expectations.

ALICEA enables embodied agents to handle intrinsic and extrinsic incongruencies by introducing friction in the human-agent dialog. It applies causal reasoning about the world at the behavior level (using task planning theory) and situation level (using Bayesian inference) to identify incongruencies. Upon detecting incongruencies, ALICEA introduces friction through dialog acts such as verification (*My recipe yields one omelet. Is that OK?*), questions (*What do you want on your omelet?*), reports (*No eggs in the fridge*), etc. These utterances trigger further deliberation in the human partner. Upon reflection, the partner provides confirmation (*Yes. That's fine*), answers (*Only cheese*), and elaborations (*Just bought some, they are in my shopping bag*). ALICEA integrates these responses in its current situational, task-related beliefs to advance in task execution. By ensuring that 1) utterances are representative of what exists or is possible in the world, 2) by ensuring that critical aspects of the task are discussed a priori, and 3) by adaptively incorporating a partner's beliefs, ALICEA enables *accountability* in embodied agents.

3.1. System 1/System 2 Organization of Embodied Conversational Reasoning

ALICEA is built upon insights from Kahneman's dual process model [1] that we extended to levels of learning in general autonomous agents [13]. The theory posits that human thinking and problem-solving arise from the confluence of two different processes: System 1, an implicit, uncontrolled process that performs associative inference, and System 2, an explicit, controlled process that performs deliberate, causal reasoning. ALICEA organizes reasoning about human-agent collaborative dialog in a similar fashion. ALICEA's System 1 - a LLM conversation model, quickly proposes plausible responses to a human utterance. Then, System 2 validates the content of those responses against current situational task-related beliefs and assesses incongruencies.

System 1. ALICEA's LLM-based conversation model, System 1 in *Figure 1*, serves two purposes. It serves as the interface between the human and the agent. It also serves as a knowledge source for the agent, who queries it when needed (e.g., to locate ingredients). It will be built using standard implementations of LLM conversation systems (e.g., DOLLY V2 [2], LLAMA 2 [14]). We will adapt it for our domain using in-context learning [15] and finetuning [16] techniques on relevant datasets (e.g., ALFRED [9], AlexaTeach [10]). Learning in such systems culminates into a probability distribution over the next token prediction from which the token is sampled during the generation process. This probability distribution is influenced by language patterns and grammar, types of question-answers in curated datasets, and human preference for certain responses that were part of its training paradigm [17]. Given its numerical nature, this distribution cannot be inspected further. The LLM model itself is frictionless; it samples from the distribution with no deliberate or explicit reasoning about the truth, applicability, assumptions, inconsistencies, or communicative goals and, consequently, so are its responses. ALICEA augments System 1 with explicit, deliberate processing in System 2 to address these gaps and introduce friction.

System 2. As shown in *Figure 1*, System 2 accepts the human utterance and the natural language response generated by System 1 as input. Each utterance is processed by an LLM-based Interpreter, which translates both into *intentional structure representation* (ISR), which is processed further by the intentional discourse model (further detailed in Section **3.2**) to track the intentional state of the dialog. System 2 maintains situational and task-related beliefs as *a situation graph* built upon inference in visuo-spatial reasoning and task planning subsystems. The *situation graph* represents what is currently true, given evidence perceived in the world and what the agent encodes as task knowledge. This graph forms a set of beliefs against which conversational beliefs are validated, which leads to the identification of incongruencies (further detailed in Section **3.3**).

3.2. Managing Conversational Friction with Intentional Discourse Model

ALICEA incorporates an intentional discourse model (IDM) built upon a well-developed theory of collaboration – Collaborative Discourse Theory (CDT [3]–[5]). CDT introduces formal constructs that enable ALICEA to model intentional and attentional states in human-agent dialog. It is effective in managing collaborative human-robot discourse [4], [18] and Interactive Task Learning [19]–[21]. In ALICEA, we employ CDT to manage friction in a dialog in response to incongruencies assessed in the embodied world-reasoning subsystem (Section **3.2**). In a key advancement, we embed CDT within LLM-based interpretation & generation mechanisms, making it robust to diversity and variation in human language. IDM operates as follows:

Interpreter translates a natural language utterance from the human or generated by the System 1 LLM into Intent Structure Representation (ISR) and vice-versa. An ISR consists of a recognized intent, a semantic interpretation in a formal language (e.g., first-order logic), and a

Human: "Can you make some omelets?"
ISR1: {**intent-that:**(agent, achieve-goal),
content: exists ?o and omelet(?o),
embedding: $\vec{X1}$} where $\vec{X}$ is an embedding vector.

System 1 LLM potential response: *"How many would you like?"*
ISR2: {**intent-to:**(self, know-info)
content: number-of(?o) and omelet(?o)
embedding: $\vec{X2}$}

System 1 LLM suggestion: *"The eggs are typically in the fridge."*
ISR3: {**intent-to:**(self, assert)
content: egg(?e) and in(fridge, o)
embedding: $\vec{X3}$}

System 1 LLM action: "*Beat eggs*"
ISR4: {**intent-to:**(self, execute-action),
content: beat(?e) and egg(?e),
embedding: $\vec{X4}$}

Figure 3. *Examples ISRs*

corresponding embedding vector computed by the LLM. Various natural language and ISR pairs are shown in *Figure 3*. To build this module, we will leverage the pattern-based string manipulation capabilities of LLMs, which are effective in translating natural language into structured representations (PDDL [22], first-order logic [23]), as well as in our work on morphological transformation [24]. We will explore both in-context learning with designed prompts and finetuning [16].

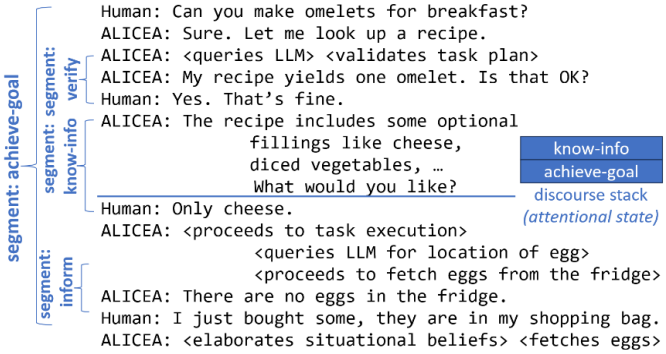


Figure 4. *Intent, segments, and discourse stack*

Manager organizes linguistic exchanges (and ISRs) in the dialog using CDT constructs. Dialog turns (utterances and actions) are organized into *segments* that fulfill a certain purpose or *intent*. A participant who initiates a segment does so with a certain intent. In *Figure 4*, the human initiates a segment by asking the agent, "*Can you make omelets for breakfast?*" the intent is for the agent to achieve a specific goal state upon which the segment is terminated. The segments are organized onto a *discourse stack*, pushed when initiated and popped when terminated, and are hierarchically organized. For example, while pursuing its goal, the agent may introduce another segment by asking the human, *"The recipe includes some optional fillings. What would you like?"* to gather the information required to achieve its goal. This new segment achieves an information subgoal necessary for the original goal. The discourse stack represents *the attentional state* of dialog – it captures the current joint goals the participants are pursuing (with assumptions they are making). The objects and actions in the discourse stack also are salient for the ongoing conversation. ALICEA advances the intentional and attentional state of dialog by processing ISR's onto and off the discourse stack in accordance with conversational policies (section **3.4**).

Comprehender. Next, the ISRs are grounded into the situation graph that is compiled and maintained by the embodied world reasoning module. Comprehension of ISRs is performed using our Indexical Model [25], [26] that posits that language is an *index* in the situation graph. It frames comprehension as a search problem instantiated over the situation graph and task knowledge, and composition of elements under pragmatic and semantic constraints to generate grounded actions/task instantiations. E.g., the comprehender finds the referent for egg(?e) (E1, a specific entity of egg type) and beat (a known skill) to instantiate beat(E1).

Validator. Once an ISR has been grounded into the situation graph, the information contained in it is validated against what the graph entails and the reasoning knowledge the agent possesses. Validation of the ISR occurs in two parallel accountability processes (Section **3.3)** and results in identified incongruencies and information expected to resolve them. Consider the human utterance and the potential response in Figure 4 above. Behavior accountability (Section **3.3.1**) finds that the retrieved recipe yields one omelet – an implicit assumption to be verified with the human. The potential response is considered valid, and System 1 is allowed to proceed with the response. However, consider the case where System 1 proposes a response: *"Alright.*

Proceeding to make some omelets." In this case, the validation process evaluates it as invalid because the available recipe only yields one omelet, and this *implicit assumption* incongruency has not been verified. In this case, System 1's proposed response is rejected, and the generator is invoked.

Generator. The generator applies a logical, rule-based process to generate an ISR corresponding to the validator's finding. For instance, if the validator reports that the recipe only yields one omelet, the generator creates an ISR: {intend-to: (self, report), content: yield(recipe, omelet, n=1), embedding=None; intent:verify, content:proceed, embedding = None}. This ISR is passed to the interpreter to generate a corresponding natural language response *"My recipe yields one omelet. Is that OK?"* The resultant human response is integrated into the dialog state by the dialog manager and in accordance with conversational policies (section **3.4**)

3.3. Accountability with Causal Embodied World Reasoning

A core part of System 2 reasoning is a world-reasoning subsystem that reasons about the world, can determine a sequence of actions corresponding to a goal, and execute them to achieve the goal. This system builds and elaborates an integrative situation graph to represent its *situational* beliefs about the environment and the status of its internal reasoning. These beliefs are consistent with what is observed in the environment and what is possible, given knowledge of environmental transitions and goals. We build upon our previous work embodied agents in DARPA GAILA [27] & SAIL-ON [28], [29] to build an embodied world reasoner as follows:

Pixels (or other sensory information) are processed by a vision transformer [30] to localize objects and extract their visual embeddings. Each object is further categorized [31], its properties identified, and its state estimated. Next, metric information about objects (bounding volumes, distances) are processed by applying domain-independent spatial calculi (e.g., Allen Interval Algebra [32]) to generate a relational situation graph representing the current, high-level state of the world.

Task planning searches for a sequence of actions that can be performed in the current world state (represented as the situation graph) to reach a goal. To handle mixed discrete and continuous environments, it employs Nyx [33], our PDDL+ planner. In addition to the current state and the goal as a problem file, the planning process reasons with a *domain* that contains environment transition models under actions and exogenous events and continuous processes.

3.2.1. Behavior Accountability with Task Plan Reasoning

ALICEA's behavioral accountability module evaluates LLM-generated task guidance against its current knowledge of the structure and dynamics of the environment. Through this evaluation, the module identifies *behavioral incongruencies* – implicit assumptions about task performance, costs of task performance, identification of critical actions and decision points – and generates information IDM uses to introduce friction in the dialog to resolve those incongruencies.

General plan deliberation evaluates if an existing/ recommended plan can be applied to the current state to achieve the expected goal. Plan validation tools such as VAL [34] are widely employed to validate plans constructed by PDDL/+ planners. As shown in Figure 5, given the current state as a problem file, VAL first 1) grounds the domain – it instantiates the actions in the recipe with various objects, and then 2) applies them in the current state sequentially to simulate how the environment evolves under the action and an internal model of environmental transition (domain). This process reveals various implicit considerations. In our omelet recipe scenario, instantiation of *add fillings* fails because no entity has been categorized as a *'filling,'* identifying the *unknown criterion* incongruency. Similarly, the goal condition (= (number_of_omelets) n) also cannot be instantiated because the number of omelets requested is unclear. These incongruencies are returned to the Validator (section **3.2**) and are used by IDM to advance the dialog (and acquire elaborative and corrective information from the human partner).

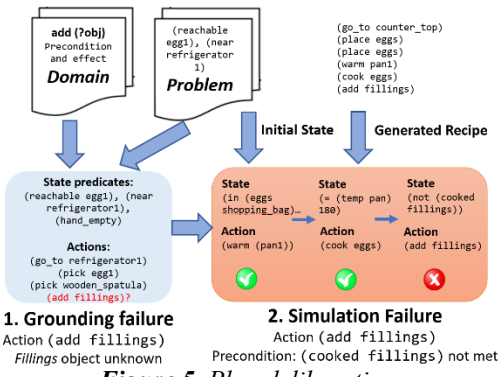


Figure 5. *Plan deliberation*

Other incongruencies are identified through plan simulation. Consider the case where the agent's internal model of *add filling* encodes that fillings must be *sauteed* as a pre-condition. Plan simulation will fail for *adding fillings* because this specific pre-condition is unmet. This *unmet condition* incongruency can be expressed to the human, and confirmation is sought that they want the fillings to be sauteed. Similarly, *implicit assumption* congruency can be identified during simulation when what is obtained by executing the recipe differs from what the human requested. In addition to validating the plan, VAL also computes various metrics such as makespan (duration of the plan), time/cost considerations, number of actions, number of goal conditions, etc., that can be used to reveal assumptions and consequences to the human partner.

Critical action validation identifies actions critical to the task's success so that these can be brought to the attention of the human partner for additional guidance and verification. To identify such actions, ALICEA relies on *landmarks* [35] in planning theory: state descriptors (edges in the situation graph) or actions that are constituents of every plan for a goal. A critical action is the only action that achieves a specific, critical goal condition. Consider various ways of making an omelet. While ways of preparing eggs, the utensils used, fillings add can differ in different plans, cooking eggs (*Figure 6*) is crucial to all. Once identified, such critical actions are highlighted to the human partner by the IDM to ensure that they are executed in an accountable fashion – i.e., asking the human to verify the temperature the eggs should be cooked at in this case.

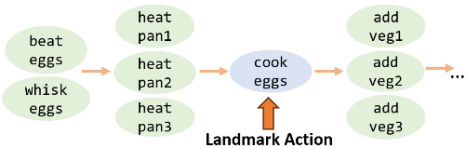
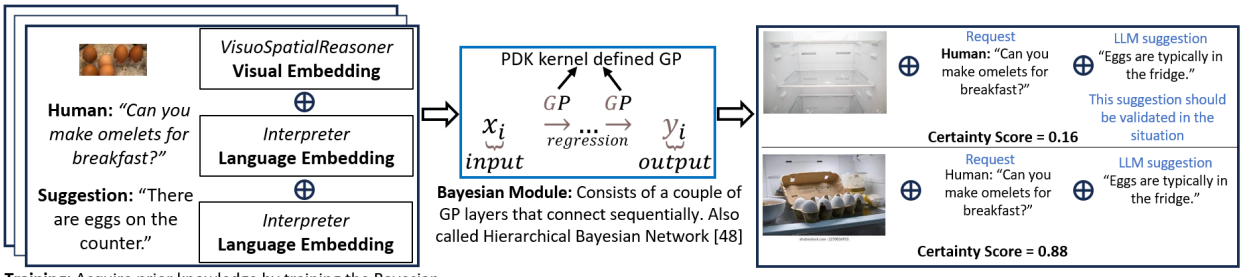


Figure 6:. *Landmark action analysis*

While the computational effort to ground the domain and simulate the plan is negligible, extracting the complete set of landmarks is PSPACE-complete, though in practice only a subset of landmarks is required and which can be extracted quickly, in real-time.

3.2.2 Situational Accountability with Bayesian Inference

Consider the case when following the omelet recipe, the agent queries System 1: *"Where are eggs in the kitchen?"* and the LLM responds: *"Eggs are typically in the fridge."* ALICEA's situational accountability module evaluates LLM's suggestions given visual evidence in the current situation to identify situational incongruencies and provides them to the Validator (section **3.2**). It is built with a deep network-based Bayesian reasoning module that computes a certainty score. The architecture (*Figure 7*) accepts as input X, a composition of three embeddings: visual embeddings (extracted by the *VisuoSpatial reasoner*), language embeddings



for the overall request, and LLM's suggestion (both extracted by the *Interpreter*). The architecture contains a couple of Gaussian Process (GP) layers: $GP(m, K)$, where m and K are corresponding mean and covariance functions. In each GP, we use the Periodic Diffusion Kernel (PDK) [36] to compute the covariance. GPs are sequentially connected, generating a Hierarchical Bayesian network (HBN [37]). The output, observation variable Y is a binary vector, where 0 means high uncertainty (i.e., visual evidence doesn't support the suggestion) and 1 means low uncertainty. We will use the doubly stochastic evidence lower bound (ELBO) – a measure of similarity between different distributions - as the objective function [37, 38]. For training, we will prepare a dataset of triples: (visual embeddings, request, and suggestion) based on ALFRED [9], AlexaTeach [10], and responses from ChatGPT. Gaussian Process-based Bayesian learning is efficient and can learn with a small dataset (~200 samples). Its computational complexity is $O(n^3)$, where n is the dimension of kernel. In our preliminary experiments, the Gaussian inference can be done in real-time (in a couple of milliseconds) when the input feature vector is sized at 2048×50 [36].

With the proposed reasoning mechanism and using different thresholds for the certainty score, we can identify various visual incongruencies useful in introducing relevant friction in the conversation. If the reasoner is highly certain, the task (and dialog) can commence without friction. If the reasoner is somewhat certain, it implies that the suggestion is not exactly applicable, but it may still work. For example, if instead of frying pan, a baking pan is found. In

this case, ALICEA can verify if proceeding with what is found is okay. In extreme cases, when the certainty is low (no eggs in the fridge), ALICEA gathers more information from humans.

3.4 A Taxonomy of Conversational Friction

For ALICEA, we will develop a suite of conversational policies employed by the IDM (section 3.2) that introduce friction when various behavioral and situational incongruencies are identified. These policies control when and to what level friction must be introduced and use the information generated by the accountability modules. The policies are represented classically as $(s, a \rightarrow value)$ where the state (s) is comprised of the type of incongruency identified and action (a) is a specific type of dialog act (question, verification, report etc.) generated based on the incongruency. The dialog act introduces friction by initiating a segment on the discourse stack, the intent of which is to resolve the incongruency. IDM incorporates human responses in the situation graph, advancing task execution. We will explore hand-developed policies and those learned through supervision [39]. To explore various policies, we will develop a taxonomy of conversational friction (example below) that includes incongruency, its type, and an appropriate conversational policy that resolves it. This will be extended during the program as we validate conversational policies for friction.

Incongruency	Example	Friction
Implicit assumption	LLM-recommended recipe yields one omelet but human needs several.	ALICEA: *"My recipe yields one omelet. Is that OK?"* Human: *"No. I need omelets for 4 people"* ALICEA: `updates goal to 4 omelets.`
Unmet condition	LLM recommends adding fillings; internal model requires sauteed.	ALICEA: "I will sauté the mushrooms." Human: "Yes. Sounds good" ALICEA: `updates plan`
Critical action identified	All plans require the eggs to be cooked.	ALICEA: *"What temperature should I cook eggs to?"* Human: *"Around 160 C"* ALICEA: `updates action effects`
Certainty score low (<0.15)	No eggs are found in the fridge, as suggested by LLM	ALICEA: *"There are no eggs in the fridge."* Human: *"They are in my shopping bag"* ALICEA: `updates current state`
Certainty score medium (>0.15, <0.85)	Frying pan not found as suggested by LLM, but a baking pan is found.	ALICEA: "Didn't find any frying pan, using a baking pan." Human: "That's alright." ALICEA: `assigns baking pan as a frying pan`
Certainty score high (>0.85)	All required items found at their suggested places	No friction.

Bibliography

[1] D. Kahneman, *Thinking, Fast and Slow*. MacMillan, 2011.

[2] M. Conover *et al.*, "Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM." Accessed: Jun. 30, 2023. [Online]. Available: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

[3] B. J. Grosz and C. L. Sidner, "Attention, Intentions, and the Structure of Discourse," *Comput. Linguist.*, vol. 12, no. 3, pp. 175–204, 1986.

[4] C. Rich, C. L. Sidner, and N. Lesh, "Collagen: Applying Collaborative Discourse Theory to Human-Computer Interaction," *AI Mag.*, vol. 22, no. 4, pp. 15–15, 2001.

[5] B. J. Grosz and S. Kraus, "The evolution of SharedPlans," in *Foundations of rational agency*, Springer, 1999, pp. 227–262.

[6] E. Kolve *et al.*, "AI2-THOR: An Interactive 3D Environment for Visual AI." arXiv, Aug. 26, 2022. doi: 10.48550/arXiv.1712.05474.

[7] M. Savva *et al.*, "Habitat: A Platform for Embodied AI Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9339–9347.

[8] O. Michel, "Cyberbotics ltd. Webots™: Professional Mobile Robot Simulation," *Int. J. Adv. Robot. Syst.*, vol. 1, no. 1, p. 5, 2004.

[9] M. Shridhar *et al.*, "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [Online]. Available: https://arxiv.org/abs/1912.01734

[10] A. Padmakumar *et al.*, "Teach: Task-driven embodied agents that chat," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2017–2025.

[11] brian ichter *et al.*, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," in *Proceedings of The 6th Conference on Robot Learning*, K. Liu, D. Kulic, and J. Ichnowski, Eds., in Proceedings of Machine Learning Research, vol. 205. PMLR, Dec. 2023, pp. 287–318. [Online]. Available: https://proceedings.mlr.press/v205/ichter23a.html

[12] N. Shinn, F. Cassano, A. Gopinath, K. R. Narasimhan, and S. Yao, "Reflexion: language agents with verbal reinforcement learning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=vAElhFcKW6

[13] J. Laird and S. Mohan, "Learning fast and slow: Levels of learning in general autonomous intelligent agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[14] "Llama 2: Open Foundation and Fine-Tuned Chat Models." 2023.

[15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Comput Surv*, vol. 55, no. 9, Jan. 2023, doi: 10.1145/3560815.

[16] H. Liu *et al.*, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 1950–1965, 2022.

[17] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, "Sources of Hallucination by Large Language Models on Inference Tasks," *ArXiv Prepr. ArXiv230514552*, 2023.

[18] J. Rickel, N. Lesh, C. Rich, C. L. Sidner, and A. Gertner, "Collaborative discourse theory as a foundation for tutorial dialogue," in *6th International Conference on Intelligent Tutoring Systems*, Springer, 2002, pp. 542–551.

[19] S. Mohan, A. Mininger, J. Kirk, and J. Laird, "Acquiring Grounded Representations of Words with Situated Interactive Instruction," *Adv. Cogn. Syst.*, vol. 2, pp. 113–130, Dec. 2012.

[20] S. Mohan, J. Kirk, and J. Laird, "A computational model for situated task learning with interactive instruction," in *International Conference on Cognitive Modeling*, 2013.

[21] P. Ramaraj, C. L. Ortiz, and S. Mohan, "Unpacking Human Teachers' Intentions for Natural Interactive Task Learning," in *2021 30th IEEE International Conference on Robot & Human*

Interactive Communication (RO-MAN), Vancouver, BC, Canada: IEEE Press, 2021, pp. 1173–1180. doi: 10.1109/RO-MAN50785.2021.9515448.
[22] Y. Xie, C. Yu, T. Zhu, J. Bai, Z. Gong, and H. Soh, "Translating natural language to planning goals with large-language models," *ArXiv Prepr. ArXiv230205128*, 2023.
[23] Y. Yang, S. Xiong, A. Payani, E. Shareghi, and F. Fekri, "Harnessing the Power of Large Language Models for Natural Language to First-Order Logic Translation." arXiv, May 24, 2023. Accessed: Dec. 11, 2023. [Online]. Available: http://arxiv.org/abs/2305.15541
[24] O.-M. Sulea and S. Young, "Unsupervised Inflection Generation Using Neural Language Modeling," *ArXiv Prepr. ArXiv191201156*, 2019.
[25] S. Mohan, A. Mininger, and J. Laird, "Towards an indexical model of situated language comprehension for cognitive agents in physical worlds," *Adv. Cogn. Syst.*, 2016.
[26] A. M. Glenberg and D. A. Robertson, "Indexical understanding of instructions," *Discourse Process.*, vol. 28, no. 1, pp. 1–26, 1999, doi: 10.1080/01638539909545067.
[27] S. Mohan, M. Klenk, M. Shreve, K. Evans, A. Ang, and J. Maxwell, "Characterizing an Analogical Concept Memory for Architectures Implementing the Common Model of Cognition," *Adv. Cogn. Syst.*, 2020.
[28] S. Mohan *et al.*, "A Domain-Independent Agent Architecture for Adaptive Operation in Evolving Open Worlds," *ArXiv Prepr. ArXiv230606272*, 2023.
[29] W. Piotrowski *et al.*, "Learning to Operate in Open Worlds by Adapting Planning Models," in *International Conference on Autonomous Agents and Multi-Agent Systems*, 2023.
[30] A. M. Rekavandi, S. Rashidi, F. Boussaid, S. Hoefs, E. Akbas, and M. bennamoun, "Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art." arXiv, Sep. 09, 2023. doi: 10.48550/arXiv.2309.04902.
[31] G. Cheng *et al.*, "Towards Large-Scale Small Object Detection: Survey and Benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13467–13488, Nov. 2023, doi: 10.1109/TPAMI.2023.3290594.
[32] Y. Gatsoulis *et al.*, "Qsrlib: a software library for online acquisition of qualitative spatial relations from video," 2016.
[33] I. Matei *et al.*, "System Resilience through Health Monitoring and Reconfiguration," *ACM Trans. Cyber-Phys. Syst.*, Nov. 2023, doi: 10.1145/3631612.
[34] R. Howey, D. Long, and M. Fox, "VAL: Automatic plan validation, continuous effects and mixed initiative planning using PDDL," in *16th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, 2004, pp. 294–301.
[35] S. Richter and M. Westphal, "The LAMA Planner: Guiding Cost-Based Anytime Planning with Landmarks," *J. Artif. Intell. Res.*, vol. 39, pp. 127–177, 2010.
[36] Y. Fan, "Solving SPDEs for Multi-Dimensional Shape Analysis," PhD Thesis, Arizona State University, 2021.
[37] Y. Fan and Y. Wang, "Geometry-aware hierarchical Bayesian learning on manifolds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1786–1795.
[38] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2. MIT press Cambridge, MA, 2006.
[39] A. Sonabend, J. Lu, L. A. Celi, T. Cai, and P. Szolovits, "Expert-supervised reinforcement learning for offline policy learning and evaluation," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 18967–18977, 2020.
[40] M. Johnson, K. Hofmann, T. Hutton, and D. Bignell, "The Malmo Platform for Artificial Intelligence Experimentation.," in *International Joint Conference on Artificial Intelligence*, 2016, pp. 4246–4247.