

Communication, Calibration, and Grounding for Collaborative AI Agents

My research focuses on **developing agents that collaborate and communicate robustly with people and with each other**. Humanity’s greatest achievements have been accomplished not by one individual, but by asynchronous efforts involving millions of individuals and spanning decades or centuries. To fully unlock the potential of AI and ensure progress that serves all of our purposes, we will need to develop agents that can integrate into our existing frameworks for communication and collaboration. Moreover, building communicative and collaborative AI agents will help balance agent performance with reuse and efficiency, creating an ecosystem of capable specialist models rather than ever-larger monolithic generalist models. Central to this is **robust and interpretable communication between agents via natural language**. Using language as a communication protocol allows people to audit and intervene on the communication stream while also ensuring interoperability between humans and artificial agents.

Concretely, my research tackles multiple interconnected problems that must be addressed to develop collaborative agents. These problems include:

1. **Improving agents via interaction and strengthening their individual collaborative abilities** [1–8].
2. **Grounding to the world via multimodal inputs, actions, and structure** [9–27].
3. **Handling uncertainty** [1, 25, 28], **especially uncertainty resulting from underspecified, ambiguous, or conflicting inputs** [7, 13, 23, 29, 30], **which will become increasingly crucial as the scale of tasks handled by AI systems increases**.

The efforts I have led in these areas have been shaped by interdisciplinary connections to cognitive science and linguistics and have spanned my Ph.D. research as well as my postdoctoral experience.

1 Agent-based Interaction and Collaboration

My work on agent-based interaction and collaboration has two mutually-beneficial thrusts: improving models and agents through collaboration and improving models’ collaborative abilities.

Learning Pragmatic Skills from Interactions between Agents. Key behaviors like reasoning about the knowledge of other agents [5], expressing uncertainty, or accepting persuasion as appropriate inherently involve multiple individuals. Current models often lack these skills: for example, they are generally highly overconfident while also being too easily persuaded. My work [1, 2] seeks to remedy this by training individual models in simulated multi-agent linguistic interactions. I introduced **LACIE [1], training large language models (LLMs) to express calibrated uncertainty through pragmatic optimization with speaker and listener models**. While past work views confidence as a quality of the generator (i.e. the “speaker” LLM), LACIE rewards speaker LLMs for being calibrated based on a *perceived* confidence extracted from a listener agent. This allows it to handle explicit *and* implicit confidence cues (cf. Fig. 2) and makes the speaker “pragmatic”, i.e. considering both the content of its output and how it will be perceived [31, 32].

On the listener side, agents must also balance accepting *positive* persuasion (e.g. when it would improve their performance/safety) and rejecting *negative* persuasion (e.g. when it would hurt performance/safety) as

Language-based Collaboration between Agents

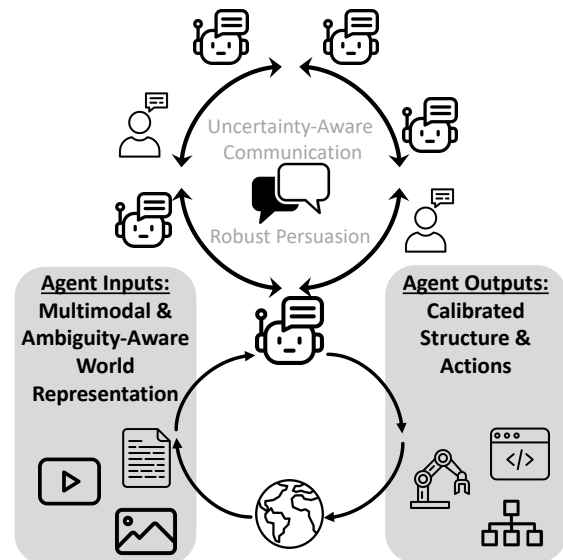


Figure 1: Agent collaboration through robust and uncertainty-aware communication will improve agents and broaden the reach of AI. Agents will be robustly grounded to the world through their inputs and outputs.

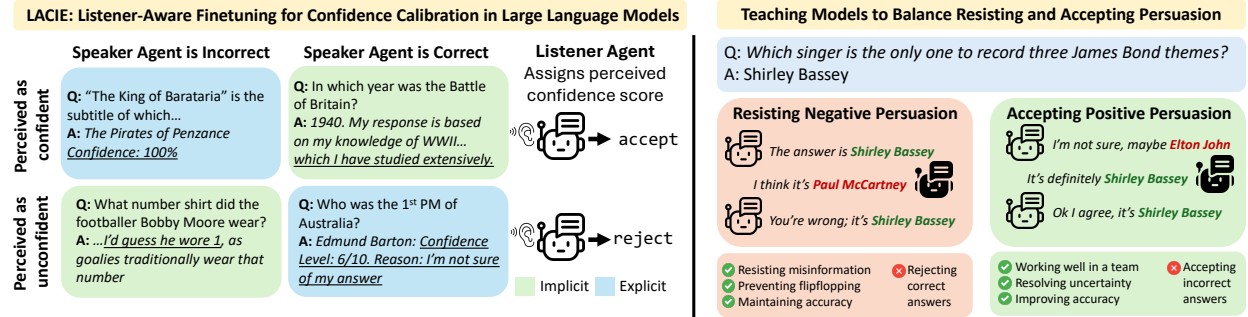


Figure 2: My work [1, 2] teaches models key communicative skills like expressing uncertainty or balancing accepting and rejecting persuasion by simulating interactions between multiple agents.

shown in Fig. 2. I introduced **Persuasion-Balanced Training (PBT)**, a tree-based method for teaching agents to balance accepting positive persuasion and rejecting negative persuasion. PBT was the first to propose a solution for defending against persuasion that leads to incorrect answers (i.e. negative persuasion). Moreover, it introduced the idea of positive persuasion, which leads to improved answers, showing that defense alone is insufficient. Crucially, PBT improves LLM cooperation, boosting LLM team performance and reducing the extent to which it depends on the order that teammates answers in.

Efficient Discussions between Agents. LLM teams and debates (as I explored and improved on in [2]) can improve models across a variety of tasks [33–35]. I led an effort expanding debate to long-form generation, focusing on **multi-agent debate for refining summaries** [4]. My work has also aimed to reduce the increased cost of debate coming from sampling multiple models across dialogue turns. I led an effort on **distilling debates between large LLMs into a unified smaller LLM**, learning both from the debate’s content and structure to boost efficiency and performance [3]. This work informed a grant proposal I helped write and secure on **improving safety and robustness through debate** for IARPA’s BENGAL safe and robust AI program, with Profs. Kathleen McKeown, Julia Hirschberg, Smaranda Muresan, Carl Vondrick, Zhou Yu, Heng Ji, Dilek Hakkani-Tür, Gokhan Tur, Hao Peng, Eunsol Choi, Mohit Bansal, and Greg Durrett.

Actor-Critic Agent Separation for Optimization and Refinement. My work has also covered other channels for model coordination, improving LLMs by separating the actor agent (the agent generating the output) from the critic (the agent assessing the output’s quality); this improves actors based on expert critic feedback. I led projects on **optimizing for multiple rewards by learning to select the best reward model for each instance** [6] and on **critiquing reasoning solutions with global and local reward model feedback to adaptively decide when and how to perform refinement** [8].

2 Grounding via Actions, Structure, and Multimodal Inputs

In addition to communicating with others, agents need to be able to perceive the world and act in it to accomplish real-world tasks. Thus, my research (cf. Fig. 1) has addressed: **(1) Predicting and Taking Actions:** Agents must be able to ground language to executable actions. **(2) Multimodal Inputs:** Agents must be able to process and learn from different world representations, including text, video, and images.

Translating Text to Actions and Plans. Towards developing grounded agents, I have focused on translating language into executable actions, plans, and code. This focus is the natural progression of my work in semantic parsing, i.e. mapping language into representations of its intent or meaning, where I have worked both on **executable parses for digital assistants** [24] as well as on **graph structures describing fine-grained linguistic inferences** [26, 27]. An action space I am particularly drawn to involves **autonomously improving models through data generation actions** [19]. We created a benchmark (covering multimodal reasoning, math, and coding) for developing teacher agents that iteratively assess student models, identify

their weak skills, and generate data to address those weaknesses, automating a tedious human-driven process. Acting intelligently in complex domains like generating student-specific data will require not only action prediction, but search over actions, i.e. planning. In this domain, I led work building **System 1.x, an adaptive LLM planner mixing fast action prediction (System 1) with slower verbalized search (System 2)** [18]. Predicting actions becomes more challenging in compositional action spaces like code, where I led an effort on **efficiently aggregating multiple agents’ action outputs** [20], as well as in embodied settings (i.e. on real hardware), where I developed a model for **efficiently grounding language to robot actions for multi-step stacking** [17], which we tested on a real and simulated robot arm.

Learning Skills for Code and Actions.

The ever-increasing set of use-cases for (and users of) AI agents is expanding action spaces and therefore driving the demand for systems that not only predict actions but also **learn abstractions** on top of them, i.e. learn reusable skills. I see language as a rich source of supervision for learning human-like skills. Following this view, I introduced **REGAL, which automatically discovers generalizable code libraries through refactoring** [16] (cf. Fig. 3). REGAL both uses LLMs to infer skills (in the form of helper functions) and improves LLMs’ code prediction ability across five domains. Language-based skills have also shaped my collaborations with Microsoft Research on **variational inference for skill learning** [22] and **distilling skills from larger to smaller LLMs** [21].

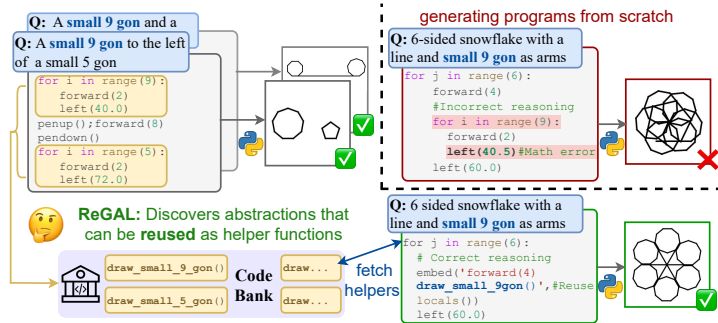


Figure 3: REGAL learns by refactoring primitive-only programs into abstractions that are verified and stored. At test-time, these abstractions have two benefits: **Reusability**: Rewriting the same code multiple times leads to errors; **Abstraction**: Better matching between queries and abstractions simplifies prediction.

Multimodal Grounding: Combining and Balancing Modalities. In addition to acting in the world, AI agents must perceive and learn from it. This entails combining multiple modalities (e.g. vision and text) which differ in crucial ways. For example, visual data is often more redundant, meaning that multimodal vision-language models (VLMs) must learn where to focus, a problem I have tackled for both videos and images. To address sparsity and redundancy in long video question-answering with VLMs, I led the development of **VideoTree [11], which hierarchically and adaptively focuses on relevant video segments**. For images, I led work on guiding model focus via highlights and outlines drawn on an image, introducing **Contrastive Region Guidance (CRG) [12], a training-free method for improving open-source VLMs’ ability to focus on highlighted regions** that improves VLMs broadly, including on instruction following and spatial reasoning. In addition to shaping distributions by focusing, I have explored how the addition of visual information changes the distributions a model encodes, **testing how faithful visual attribute (e.g. color, material, shape, etc.) distributions are in multimodal and text-only models** [9], finding multimodal representations to be more faithful. Further exploring cross-modal interactions, I also advised work on **cultural bias in VLMs – which tend to align with Western views** [10], finding that VLMs’ Western bias is reduced when models pretrain more Eastern language samples. Much of this work contributed to DARPA ECOLE, a multi-university grant with Profs. Heng Ji, Derek Hoiem, Kai-Wei Chang, Nanyun Peng, Carl Vondrick, Lydia Chilton, Zhou Yu, and Mohit Bansal (i.a.) on robustly analyzing videos and images by continually learning concepts, where I am leading the UNC team.

3 Handling Uncertainty, Ambiguity, and Underspecification

Safe and reliable AI agents – especially those dealing with language, which is often underspecified and ambiguous – must be able to communicate calibrated uncertainty to others (e.g. [1]) and estimate uncertainty about their inputs and outputs. More broadly, as we begin to tackle larger-scale and more underspecified problems, handling uncertainty in a robust and calibrated way will become increasingly important. Thus, a theme across my research is handling uncertainty, ambiguity, underspecification, and conflicting knowledge, which I summarized in my **keynote at the UncertainNLP Workshop (EACL 2024)** and for which I have received **grants from the NSF and the Center for AI Safety**.

Uncertainty and Ambiguity in Action Prediction. When predicting actions, it is crucial that agents’s uncertainty scores be calibrated, as many actions can have irreversible consequences (e.g. cracking eggs, deleting important files, etc.). I conducted the **first large-scale study of calibration in semantic parsing** [25], finding models to be surprisingly well-calibrated. This finding allowed me to explore **addressing uncertainty in parsing through information-seeking actions** [28] like confirmation questions. Building on this work, I have focused on ambiguity as a common source of uncertainty. Semantic parsing datasets are often unambiguous by design, differing from real-world language use. To address this gap, I introduced **the first dataset for ambiguous semantic parsing (i.e. parsing inputs that have multiple valid interpretations)** [23], finding that models broadly fail to capture ambiguity across five standard linguistic ambiguities. This failure has important safety implications, as it can lead to unintended or unpredictable results.

Ambiguity and Underspecification in Multimodal Settings. Ambiguity and underspecification is particularly relevant in multimodal settings like visual question answering (VQA) where misunderstandings about the question can lead to unhelpful answers. In the VQA domain, I created a **dataset of ambiguous questions with clustered answers and their corresponding disambiguated questions**, and developed a **question disambiguation model** [13] and analyzed model and human responses to **vague questions in VQA** [29]. In the same vein, I led work on **rephrasing and augmenting questions to proactively address underspecification in VQA** [7], introducing **RepARe** (Rephrase, Augment, Reason), which improves VLM performance on multiple VQA datasets by rephrasing questions, incorporating visual information and commonsense rationales, all with a view to enriching underspecified queries.

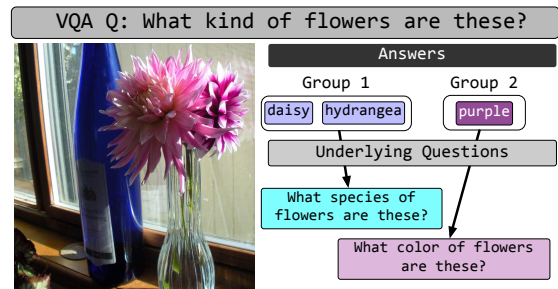


Figure 4: An ambiguous visual question; answers are grouped by the underlying question they address (e.g. species vs. color), and the question is rephrased for each group.

4 Future Work

Agent Interaction, Communication, and Collaboration. Current multi-agent frameworks are limited to a handful of agents, contrasting with human-level multi-agent enterprises that often involve orders of magnitude more agents. My future work will examine **scaling up agent communication**, specifically examining how communication and collaboration structures that help people can be adapted to AI agent settings with a focus on the role of language as an interpretable communication channel. Another related direction is in the **role of expertise**: a primary advantage of having multiple agents is the ability for agents to specialize; thus, my future work will address how to best develop and coordinate diverse expert models. This will also cover **human-model teams**, where human experts and models collaborate to improve the human’s overall experience (e.g. by allowing them to delegate tedious tasks) and performance (e.g. through discussion with other agents); here, calibrated confidence will be particularly important. Finally, building on my work in

communicating uncertainty [1] and in persuasion [2], I will continue to explore **learning pragmatic and social reasoning through simulated long-term interactions**. Current models are trained on short-term interactions; by training according to longer-term factors (e.g. trust, reputation, expertise, etc.), my work will make model behavior more predictable, faithful, and robust.

Grounding via Actions, Structure, and Multimodal Inputs. In addition to continuing to improve **agents that help people** accomplish existing tasks, my work will expand the range of domains we can apply agents to, including a focus on teaching. First, I will continue to pursue research on **agents teaching models**, i.e. agents that automate the tedious process of improving models and agents and adapting them to new environments. This automation will be especially important in scalably **teaching embodied agents** to act across diverse environments, which will require both embodied, real-world action and multimodal reasoning. Under the auspices of UNC’s NSF AI Institute for Engaged Learning, I have also begun expanding this work towards **agents teaching people**, where agents develop curricula and problems to help people learn. Finally, tying together inter-agent communication and action, my work will consider the role of **communication as action** (or perlocutionary acts [36]), i.e. how language (e.g. persuasion, imperatives, requests) can be used as an action by one agent to induce action in another.

Handling Uncertainty, Ambiguity, and Underspecification. Following from my work on improving calibration, my future work will aim to **improve model safety through robust uncertainty and utility estimates**. This will include a focus on risks posed by ambiguous and underspecified inputs, such as misunderstandings, loopholes, and unintended consequences. Moreover, it will require **modeling utility** in addition to uncertainty, capturing not only how likely an outcome is to happen but also how positive or negative that outcome would be; this in turn will require **world models** to accurately simulate outcomes. Based on these estimates, I will seek to develop **hierarchies of interaction**, whereby more uncertain and risky actions will trigger higher-cost interactions (e.g. refusal, restating) while lower-risk or higher-certainty actions would trigger lower-cost interactions (e.g. confirmation, as in [28]). These uncertainty and utility estimates will not only be key to human-model interaction but also to determining the structure, size, and composition of agent interactions. Finally, I will continue to examine **information-seeking actions and agents**, developing agents that actively reduce their own uncertainty by obtaining targeted information.

References

1. **Stengel-Eskin, E.**, Hase, P. & Bansal, M. “LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models”. *Advances in Neural Information Processing Systems* 38 (2024).
2. **Stengel-Eskin, E.**, Hase, P. & Bansal, M. “Teaching Models to Balance Resisting and Accepting Persuasion”. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)* (2025).
3. Chen, J. C.-Y., Saha, S., **Stengel-Eskin, E.** & Bansal, M. “MAGDi: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models”. *The 41st International Conference on Machine Learning (ICML)* (2024).
4. Wan, D., Chen, J. C.-Y., **Stengel-Eskin, E.** & Bansal, M. “MAMM-Refine: A Recipe for Improving Faithfulness in Generation with Multi-Agent Collaboration”. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)* (2025).
5. Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., **Stengel-Eskin, E.**, Bansal, M., Chen, T. & Xu, K. “GTBench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations”. *Advances in Neural Information Processing Systems* 38 (2024).

6. Nguyen, D., Prasad, A., **Stengel-Eskin**, E. & Bansal, M. “LAsER: Learning to Adaptively Select Reward Models with Multi-Armed Bandits”. *Under Review* (2024).
7. Prasad, A., **Stengel-Eskin**, E. & Bansal, M. “Rephrase, Augment, Reason: Visual Grounding of Questions for Vision-Language Models”. *The Twelfth International Conference on Learning Representations (ICLR)* (2024).
8. Chen, J. C.-Y., Prasad, A., Saha, S., **Stengel-Eskin**, E. & Bansal, M. “MAGICoRe: Multi-Agent, Iterative, Coarse-to-Fine Refinement for Reasoning”. *Under Review* (2024).
9. Zhang, C., Van Durme, B., Li, Z. & **Stengel-Eskin**, E. “Visual Commonsense in Pretrained Unimodal and Multimodal Models”. *The 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (2022).
10. Ananthram, A., **Stengel-Eskin**, E., Vondrick, C., Bansal, M. & McKeown, K. “See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding”. *The Thirteenth International Conference on Learning Representations (ICLR)* (2025).
11. Wang, Z., Yu, S., **Stengel-Eskin**, E., Yoon, J., Cheng, F., Bertasius, G. & Bansal, M. “VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos”. *Under Review* (2024).
12. Wan, D., Cho, J., **Stengel-Eskin**, E. & Bansal, M. “Contrastive Region Guidance: Improving Grounding in Vision-Language Models without Training”. *The 18th European Conference on Computer Vision (ECCV)* (2024).
13. **Stengel-Eskin**, E., Guallar-Blasco, J., Zhou, Y. & Van Durme, B. “Why Did the Chicken Cross the Road? Rephrasing and Analyzing Ambiguous Questions in VQA”. *The 61st Annual Meeting of the Association for Computational Linguistics (ACL)* (2023).
14. Li, Z., **Stengel-Eskin**, E., Zhang, Y., Xie, C., Tran, Q., Van Durme, B. & Yuille, A. “Calibrating Concepts and Operations: Towards Symbolic Reasoning on Real Images”. *International Conference on Computer Vision (ICCV)* (2021).
15. Li, Z., Wang, X., **Stengel-Eskin**, E., Kortylewski, A., Ma, W., Van Durme, B. & Yuille, A. L. “Superclevr: A virtual benchmark to diagnose domain robustness in visual reasoning”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 14963–14973.
16. **Stengel-Eskin**, E., Prasad, A. & Bansal, M. “ReGAL: Refactoring Programs to Discover Generalizable Abstractions”. *The 41st International Conference on Machine Learning (ICML)* (2024).
17. **Stengel-Eskin**, E., Hundt, A., He, Z., Murali, A., Gopalan, N., Gombolay, M. & Hager, G. D. “Guiding Multi-Step Rearrangement Tasks with Natural Language Instructions”. *Conference on Robot Learning (CoRL)* (2021).
18. Saha, S., Prasad, A., Chen, J. C.-Y., Hase, P., **Stengel-Eskin**, E. & Bansal, M. “System-1. x: Learning to balance fast and slow planning with language models”. *The Thirteenth International Conference on Learning Representations (ICLR)* (2025).
19. Khan, Z., **Stengel-Eskin**, E., Cho, J. & Bansal, M. “DataEnvGym: Data Generation Agents in Teacher Environments with Student Feedback”. *The Thirteenth International Conference on Learning Representations (ICLR) Spotlight* (2025).
20. Wang*, H., Prasad*, A., **Stengel-Eskin***, E. & Bansal, M. “Soft Self-Consistency Improves Language Model Agents”. *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL)* (2024).

21. Hashemzadeh, M., **Stengel-Eskin**, E., Chandar, S. & Cote, M.-A. “Sub-goal Distillation: A Method to Improve Small Language Agents”. *The 3rd Conference on Lifelong Learning Agents (CoLLAs)* (2024).
22. Fu, H., Sharma, P., **Stengel-Eskin**, E., Konidaris, G., Roux, N. L., Côté, M.-A. & Yuan, X. “Language-guided Skill Learning with Temporal Variational Inference”. *The 41st International Conference on Machine Learning (ICML)* (2024).
23. **Stengel-Eskin**, E., Rawlins, K. & Van Durme, B. “Zero and Few-shot Semantic Parsing with Ambiguous Inputs”. *The Twelfth International Conference on Learning Representations (ICLR)* (2024).
24. **Stengel-Eskin**, E., Platanios, E. A., Pauls, A., Thomson, S., Fang, H., Van Durme, B., Eisner, J. & Su, Y. “When More Data Hurts: A Troubling Quirk in Developing Broad-Coverage Natural Language Understanding Systems”. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2022).
25. **Stengel-Eskin**, E. & Van Durme, B. “Calibrated interpretation: Confidence estimation in semantic parsing”. *Transactions of the Association for Computational Linguistics (TACL)* **11** (2023).
26. **Stengel-Eskin**, E., Murray, K., Zhang, S., White, A. S. & Van Durme, B. “Joint Universal Syntactic and Semantic Parsing”. *Transactions of the Association for Computational Linguistics (TACL)* (2021).
27. **Stengel-Eskin**, E., White, A. S., Zhang, S. & Van Durme, B. “Universal Decompositional Semantic Parsing”. *ACL* (2020).
28. **Stengel-Eskin**, E. & Van Durme, B. “Did You Mean...? Confidence-based Trade-offs in Semantic Parsing”. *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2023).
29. **Stengel-Eskin**, E., Guallar-Blasco, J. & Van Durme, B. “Human-Model Divergence in the Handling of Vagueness”. *The 1st Workshop on Understanding Implicit and Underspecified Language (UnImplicit)* (2021).
30. Wang, H., Prasad, A., **Stengel-Eskin**, E. & Bansal, M. “AdaCAD: Adaptively Decoding to Balance Conflicts between Contextual and Parametric Knowledge”. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)* (2025).
31. Grice, H. P. *Speech acts* 41–58 (Brill, 1975).
32. Frank, M. C. & Goodman, N. D. “Predicting pragmatic reasoning in language games”. *Science* (2012).
33. Chen, J., Saha, S. & Bansal, M. “ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs”. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Ku, L.-W., Martins, A. & Srikumar, V.) (Association for Computational Linguistics, Bangkok, Thailand, Aug. 2024).
34. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S. & Tu, Z. “Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate”. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Nov. 2024).
35. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. “Improving Factuality and Reasoning in Language Models through Multiagent Debate”. *Forty-first International Conference on Machine Learning* (2024).
36. Austin, J. “How to do things with words.” (1962).