Survey on Computational Approaches to Implicature

Kaveri Anuranjana
IIIT Hyderabad
kaveri.a@research.iiit.ac.in

Srihitha Mallepally*
IIIT Hyderabad
srihitha.mallepally@students.iiit.ac.in

Sriharshitha Mareddy*
IIIT Hyderabad
sriharshitha.mareddy@students.iiit.ac.in

Amit Shukla
IISER Bhopal
amit20@iiserb.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

This paper explores the concept of solving implicature in Natural Language Processing (NLP), highlighting its significance in understanding indirect communication. Drawing on foundational theories by Austin, Searle, and Grice, we discuss how implicature extends beyond literal language to convey nuanced meanings. We review existing datasets, including the Pragmatic Understanding Benchmark (PUB), that assess models' capabilities in recognizing and interpreting implicatures. Despite recent advances in large language models (LLMs), challenges remain in effectively processing implicature due to limitations in training data and the complexities of contextual interpretation. We propose future directions for research, including the enhancement of datasets and the integration of pragmatic reasoning tasks, to improve LLMs' understanding of implicature and facilitate better human-computer interaction.

1 Introduction

Humans have long been fascinated by how language conveys meaning through words, intentions, and implications, prompting philosophers and linguists to explore its role in communicating information and performing actions.

(Austin, 1975) introduced the concept of speech acts, stating that utterances can perform actions beyond merely stating facts. He distinguished between locutionary acts, illocutionary acts, and perlocutionary acts. A **locutionary act** is the act of making an utterance with a specific sense and reference, an **illocutionary act** is the intentional action performed by the speaker through the utterance, and a **perlocutionary act** is the effect the utterance has on the listener.

(Searle, 1975) expanded speech act theory by categorizing illocutionary acts into five types: assertives (statements of fact), directives (requests or commands), commissives (promises), expressives (emotional expressions), and declarations (statements that change reality, such as making a decision official). He also introduced indirect speech acts, where the intended meaning differs from the literal words. This is crucial to implicature, as it highlights how speakers can imply actions or intentions indirectly, such as saying, "It's getting late," to suggest it's time to leave without explicitly stating it.

Implicature introduced by (Grice, 1975), refers to the implied meanings conveyed by speakers that extend beyond literal interpretations. It relies on the *cooperative principle* and *conversational maxims*, where violations lead to additional meanings, i.e., implicatures or pragmatic effects.

1.1 Gricean Theory of Implicature

Grice's theory distinguishes between two types of implicature: conventional and conversational. Conventional implicatures are tied to specific words, while conversational implicatures depend on context and the cooperative principle.

Implicature, as introduced by (Grice, 1975), refers to the implied meaning that a speaker conveys in an utterance, often extending beyond its literal interpretation. Grice's work laid the foundation for understanding how speakers communicate additional information implicitly, which is crucial for successful human communication and a deeper understanding of pragmatics. Under the **cooperative principle**, Grice posits that speakers facilitate a conversation by being cooperative agents — being truthful, informative, relevant, and clear. The meaning of an utterance is thus inferred based on the *assumption* that speakers generally follow these conversational **maxims**: *quantity, quality, relation*, and *manner*. When a speaker **violates** or contradicts one of these maxims, they do so to convey an additional meaning beyond the literal content. This results in pragmatic effects, such as humor or sar-

* Equal contribution.

casm, but more commonly, it leads to implicature.

2 Implicature in Computational Linguistics

Understanding implicature is essential in NLP tasks like dialogue systems. Despite the success of LLMs, their ability to handle implicature remains largely unexplored. This survey reviews current approaches, datasets, and future directions in the field.

2.1 Datasets for Implicature

Multiple datasets have been created to evaluate implicature understanding, but most have limited scope. Examples are listed below:

- **Pragmatic Understanding Benchmark (PUB)**: A benchmark (Sravanthi et al., 2024) designed to assess pragmatic reasoning across multiple dimensions, including implicature. The PUB tasks pertaining to implicature include:
  - **CIRCA**: A dataset designed to assess conversational implicature by providing dialogue snippets and context-based questions, enabling models to infer implied meanings from conversations. 2.5K label-balanced samples from the dataset's 4.3K samples were considered for the benchmark.
  - **GRICE**: This dataset focuses on implicature recovery tasks, evaluating models' abilities to identify and interpret implied meanings based on conversational maxims and context.
  - **FigQA**: A dataset that integrates language and visual understanding, assessing models' capabilities to reason about implicatures within visual contexts, using images paired with descriptive questions.
  - **FLUTE**: Aimed at understanding linguistic context, FLUTE evaluates how well models can identify implicatures that arise during conversational exchanges, emphasizing the importance of context in meaning interpretation.
  - **IMPRESS**: This dataset tests the ability to recognize and interpret implicatures across various communicative scenarios, enhancing models' understanding of nuanced language use and social dynamics. (Sravanthi et al., 2024) select 2100 (together with NOPE dataset) from the dataset's 25K datapoints.
  - **NOPE**: A dataset focused on evaluating models' performance in understanding negation and its implications in conversational contexts, providing insights into how negation interacts with implicature. (Sravanthi et al., 2024) select 2100 (together with IMPRESS dataset) fromu the dataset's 2732 datapoints.
  - **CircaPlus**: A newly annotated dataset containing 2,500 human-written implied meanings derived from the CIRCA dataset, focusing on indirect responses to enrich the study of conversational implicature.
  - **DialogAssumptions**: This dataset includes 2,500 pairs of expert-annotated presuppositions based on dialogues from the Dailydialog dataset, specifically addressing presuppositions in conversational contexts where trigger words are absent.
  - **MetoQA**: A novel dataset comprising 1,100 multiple-choice questions based on metonymy, exploring the pragmatic implications of using closely associated terms in language, contrasting with metaphorical expressions.
- **Conversational Implicature Dataset**: This dataset (George and Mamidi, 2020) provides conversational contexts where the model is tasked with generating the implied meaning. It offers a more naturalistic setting for evaluating a model's ability to infer implicatures. This dataset was also included in (Srivastava et al., 2023) and converted into a binary classification task to avoid challenges associated with evaluating generative models.

2.1.1 Language Reasoning Tasks and Their Relation to Implicature

Recent advancements in NLP require LLMs capable of complex reasoning and contextual understanding, similar to the challenges presented by implicature. The **Massive Multitask Language Understanding** (MMLU) benchmark (and recently, MMLU-Pro+ (Taghanaki et al., 2024)) evaluates

model performance across various tasks, including commonsense reasoning and reading comprehension (Hendrycks et al., 2021). It demands that models infer underlying implications and relationships beyond explicit content. Similarly, **HellaSwag** (Zellers et al., 2019) challenges models with context-rich scenarios requiring commonsense reasoning to select the most plausible narrative continuation, emphasizing the complexity of language understanding crucial for grasping implicature.

Tasks such as the **Abstraction and Reasoning Corpus (ARC)** (Chollet, 2019) challenge models with problems that require creative problem-solving and reasoning skills. Unlike simpler NLI tasks that focus on explicit entailment, ARC demands the flexible manipulation of underlying concepts, pushing the boundaries of reasoning capabilities. Similarly, the **MultiWOZ** dataset (Budzianowski et al., 2018) demonstrates the practical significance of implicature in conversational contexts, where accurately interpreting indirect user intents is crucial for effective dialogue management. Together, these tasks underscore the importance of developing models equipped for nuanced reasoning and contextual comprehension—skills critical for handling implicature in natural language understanding.

There is an urgent need for datasets that formalize implicature based on Gricean theory, as these could provide more nuanced insights into conversational dynamics and speaker intentions. Such datasets would enhance NLP models' ability to process indirect communication effectively, bridging the gap between theoretical pragmatics and practical applications in dialogue systems.

3 Advances in Computational Implicature

Recent advances in LLMs have enabled the tackling of complex NLP tasks involving reasoning and inference, yet implicature continues to pose a significant challenge even for the most advanced models, as highlighted by the following key LLMs and their performance on implicature tasks (scores in percentages):

- **GPT-4**: GPT-4 achieves accuracy of 81.8 on the conversational implicature dataset by (Sravanthi et al., 2024), as highlighted in (Ruis et al., 2023). This improvement is attributed to instruction-level fine-tuning, enhancing its ability to infer meaning beyond literal text. While GPT-4 outperforms earlier models, it still falls short of human-level understanding, indicating that scaling alone is insufficient for resolving complex implicatures. Notably, using accuracy as a metric is further justified by (Srivastava et al., 2023) who demonstrated the effectiveness of binary classification for evaluating implicature tasks.

- **GPT-3.5**: While GPT-3.5 performs well on various NLP tasks, achieving an accuracy of 78.13 on the GRICE benchmark and 48.86 on IMPRESS, its ability to generate and recognize implicatures is limited. Evaluations using the Pragmatic Understanding Benchmark (PUB) reveal that it often struggles to produce the depth of reasoning required for meaningful implicatures in conversational contexts.

- **LLaMA-2**: Despite offering scalability, LLaMA-2's performance on implicature tasks is moderate, with accuracies of 56.26 on GRICE and 49.29 on IMPRESS for the 7 billion parameter model and 75.91 and *55.09 for the 70B model, respectively. Even with a large number of parameters, these models fail to consistently generate informative and contextually accurate implicatures in PUB assessments, indicating that scale alone is insufficient for understanding implicature.

- **FlanT5-XXL**: Instruction fine-tuned models like FlanT5-XXL excel in implicature tasks, achieving an accuracy of 82.9 on GRICE and 64.12 on IMPRESS, likely due to their training on reasoning-centric objectives. PUB results indicate that these models outperform larger, untuned alternatives, highlighting the importance of instruction-based fine-tuning for effectively tackling pragmatic reasoning challenges. Notably, human performance on these tasks is significantly higher, with accuracies of 93.67 on GRICE and 57.91 on IMPRESS, underscoring the gap between current LLM capabilities and human understanding.

3.1 Challenges in Evaluating Implicature

Training Data Gaps: Most implicature datasets are small (PUB benchmark generally contains 2.5K samples for each task) compared to extensive datasets used for reasoning tasks such as ARC and HellaSwag (14.3M and 10K samples, respectively). This limits model performance due to insufficient

examples of diverse implicature contexts. Expanding these datasets to include real-world conversational implicatures is crucial for enhancing model performance. Additionally, datasets formalizing the implicature according to Gricean theory could provide deeper insights into conversational dynamics and speaker intentions, ultimately improving the understanding of NLP models of indirect communication.

Complexity of Contextual Interpretation: Implicature is highly context-dependent, making standardized evaluation metrics challenging to establish. Different conversational settings can yield varying implications, complicating the assessment of a model's understanding. Moreover, inferring meaning from nonliteral language requires a nuanced grasp of social norms, cultural references, and pragmatic cues, which are often difficult to quantify.

(Sravanthi et al., 2024) report that models struggle with indirect response interpretation and NLI pragmatic inferences, making implicature particularly challenging among other pragmatic tasks. Furthermore, the subpar performance of LLMs in pragmatics tasks stems more from deficiencies in pragmatic reasoning than from a lack of world knowledge, as evidenced by poor results in tasks like Deixis, which do not rely on such knowledge.

4 Future Scope of Implicature

To improve the understanding of implicature, future models should focus on tasks that involve pragmatic reasoning, integrating world knowledge and social norms. This approach will enable more effective implied communication and enhance human-computer interaction. Future developments could involve expanding current models, which often have limited capacities, through crowd-sourcing efforts and by introducing upgraded versions of existing datasets, as demonstrated with CircaPlus and the Circa dataset. Additionally, utilizing established benchmarks such as HellaSwag, ARC, and MMLU for data augmentation could further enhance these models.

Broadening the scope of current implicature datasets to cover a wider range of conversational contexts and diversifying the types of implicature represented will create more realistic benchmarks for model evaluation. It is essential to incorporate data from real-life conversations, particularly in naturalistic settings, to improve model generalization and ensure that NLP systems can effectively navigate the complexities of human language across various scenarios.

5 Conclusion

In conclusion, the study of implicature remains a critical area within computational linguistics, presenting unique challenges that current models struggle to address effectively. Despite advancements in large language models (LLMs), their limited ability to grasp the nuances of implicature underscores the need for more comprehensive datasets and refined training methodologies. By integrating insights from Grice's theory and focusing on pragmatic reasoning tasks, future research can help bridge the gap between human and machine understanding of indirect communication. Enhancing models with diverse conversational contexts will improve their performance in implicature tasks and enhance the overall effectiveness of NLP systems in real-world applications.

References

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

François Chollet. 2019. On the measure of intelligence.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.

HP Grice. 1975. Logic and conversation. *Cole and Morgarn*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms.

JR Searle. 1975. Indirect speech acts. *Syntax and semantics*, 3.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixi-

ang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Saeid Asgari Taghanaki, Aliasgahr Khani, and Amir Khasahmadi. 2024. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. *arXiv preprint arXiv:2409.02257*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?