Which course? *Dis*course!
Teaching Discourse and Generation in the Era of LLMs

Junyi Jessy Li[1] **Yang Janet Liu**[2] **Kanishka Misra**[1]
Valentina Pyatkin[3] **William Sheffield**[1]
[1]The University of Texas at Austin
[2]University of Pittsburgh
[3]Allen Institute for AI
{jessy,kmisra,sheffieldw}@utexas.edu, jal787@pitt.edu, valpyatkin@gmail.com

Abstract

The field of NLP has undergone vast, continuous transformations over the past few years, sparking debates going beyond discipline boundaries. This begs important questions in education: how do we design courses that bridge sub-disciplines in this shifting landscape? This paper explores this question from the angle of discourse processing, an area with rich linguistic insights and computational models for the intentional, attentional, and coherence structure of language. Discourse is highly relevant for open-ended or long-form text generation, yet this connection is under-explored in existing undergraduate curricula.

We present a new course, "**Computational Discourse and Natural Language Generation**". The course is collaboratively designed by a team with complementary expertise and was offered for the first time in Fall 2025 as an upper-level undergraduate course, cross-listed between Linguistics and Computer Science.[1] Our philosophy is to deeply integrate the theoretical and empirical aspects, and create an exploratory mindset inside the classroom and in the assignments. This paper describes the course in detail and concludes with takeaways from an independent survey as well as our vision for future directions.

1 Introduction

Natural Language Processing (NLP) as a field today is one filled with intensely rapid developments and clashing view points: "Scaling is powerful" (Kaplan et al., 2020; Hoffmann et al., 2022) vs. "Scaling is not all you need" (Li et al., 2025a; Marcus, 2025; Hooker, 2025); "LLMs cannot possibly tell us anything about language" (Chomsky et al., 2023; Bolhuis et al., 2024) vs. "Actually, they can" (Piantadosi, 2024; Futrell and Mahowald, 2025); "LLMs cannot model meaning because they are not grounded" (Bender and Koller, 2020) vs. "But grounding is not always needed" (Piantadosi and Hill, 2022; Pavlick, 2023); "LLMs can be proxies for human annotators" (Gilardi et al., 2023; Calderon et al., 2025) vs. "No, that's risky" (Baumann et al., 2025; Wang et al., 2025), to name a few. These debates no longer focus on NLP alone; they recontextualize how we think about linguistics and many other disciplines. Thus, in order for the next generation of AI practitioners to grasp the cutting edge, to critically engage in these debates, and to contribute new insights, we believe it is crucial to create curricula sitting at the intersection of different (sub)areas.

This paper presents a new course at the intersection of discourse processing and natural language generation (NLG). LLMs increasingly perform tasks that require reasoning over lengthy contexts and generating long-form responses, from conversational agents to deep research. Yet, much of what enables these systems to produce meaningful discourse remains poorly understood. In addition, recent advances in long-context LLMs have extended the ability of AI systems to process and generate text across thousands of tokens, raising new questions about how coherence, relevance, and discourse structure are maintained over several spans of text. Despite impressive surface fluency, such models often struggle with maintaining global consistency, logical flow, and discourse-level planning. These challenges highlight the need for a deeper understanding of discourse as a fundamental component of language intelligence.

However, courses about (computational) discourse[2] are far and between. Existing curricula also depict distinct partial views of discourse depending on the discipline: for CS students, it may look like a collection of tasks to be solved like topic segmenta-

[1] Website: https://jessyli.com/courses/lin353d
Authors listed alphabetically.

[2] Note that "discourse" here refers to language processing beyond the sentence boundaries (Stede, 2012), and not critical discourse theory.

Proceedings of the Seventh Workshop on Teaching Natural Language Processing (TeachNLP 2026), pages 45–58
March 29, 2026 ©2026 Association for Computational Linguistics

tion, discourse parsing, coreference resolution, and coherence modeling (Jurafsky and Martin, 2026; Eisenstein, 2019), while for Linguistics students it could be taking concepts from their theoretical classes and understanding their interaction with computational tools. These views, while valid and useful, do not align well with the reality of today, where multi-sentential processing and generation implicitly but inevitably involve discourse-sensitive capabilities that are not always squarely covered by well-defined tasks.

With these considerations, we designed a new upper-level undergraduate course "**Computational Discourse and Natural Language Generation**", cross-listed between the Linguistics and Computer Science departments. This course was offered at the University of Texas at Austin (UT Austin) in Fall 2025. The course development team spans researchers with diverse background and training: we have three professors with expertise in discourse processing, psycholinguistics, and NLG and its evaluation; a postdoc that has led the development of state-of-the-art language models (Lambert et al., 2025); and a computational linguistics PhD student working on discourse and pragmatics.

We collaboratively designed the course content with the following philosophy: **(1)** A deep integration of discourse and NLG enables us to connect linguistically-motivated frameworks or theories (e.g., local and global coherence, discourse structure, entity tracking, questions under discussion) with the evaluation, analyses, and improvement of LLMs. This will further teach students how to come up with hypotheses and design controlled experiments for them. **(2)** Some class sessions are reserved for mini workshops, helping students to further ground abstract concepts in data. **(3)** It is important to foster an explorer's mindset and to engage students in critical thinking, cutting-edge research, and open-ended problems. **(4)** Lastly, course content should be accessible but also adequately challenging for students across distinct backgrounds. By the end of the course, students will have a richer understanding of how discourse-level linguistic insights can inform computational models, and how advances in NLP can, in turn, provide new insights into the nature of language and communication more broadly.

This paper presents our inspirations (Section 2), design principles (Section 3), course content (Section 4), and assignments/projects (Section 5). We also engaged with a third-party education evaluation (TACC Education Services) to study student reception in the second half of the semester, and these results are discussed in Section 6. Overall, we are optimistic about the timeliness and relevance of a course like this that tightly integrates theoretical and empirical insights, and we conclude with challenges for future iterations to address (Section 7).

2 Motivation and Inspirations

To the best of our knowledge, prior or existing courses of a similar kind have treated discourse processing and NLG largely as separate areas.

On the discourse side, we take inspirations from prior graduate courses, such as *Computational Discourse Modeling* offered by Amir Zeldes at Georgetown University, and *Computational Models of Discourse* offered by Alexis Palmer at The University of Colorado Boulder. They both emphasize theoretical and computational approaches to discourse structure and its applications through a graduate-level, project-based seminar. We also studied individual lectures in NLP courses about discourse, such as Julia Hockenmaier's lectures on discourse coherence and centering theory at the University of Illinois Urbana-Champaign.

In the context of NLG, we consulted the latest openly available courses. We were particularly inspired by Ehud Reiter's lectures on NLG,[3] Silvia Casola's lecture on NLG evaluation at LMU Munich,[4] Greg Durrett's NLP course offered at UT Austin (Durrett et al., 2021), Yejin Choi's NLP course offered at the University of Washington,[5] and Song Han's *TinyML and Efficient Deep Learning Computing* offered at MIT.[6]

While these courses equipped students with deep, focused expertise in their respective areas, they also highlighted the need for an integrated treatment that unites discourse and generative models, motivating the design of the current course, "**Computational Discourse and Natural Language Generation**".

3 Design Principles

Our curriculum is guided by a set of design principles intended to support students in developing a

[3] https://www.abdn.ac.uk/registry/courses/postgraduate/2024-2025/computing_science/cs551h

[4] https://slvcsl.github.io

[5] https://safe-fernleaf-26d.notion.site/Winter-24-CSE-447-517-Natural-Language-Processing-4142333a001143d2be5ecff1a535c4ab

[6] https://hanlab.mit.edu/courses/2024-fall-65940

well-rounded and hands-on understanding of discourse processing, modeling, and NLG. These principles highlight the interplay between conceptual frameworks, state-of-the-art techniques, and analytical methodologies, with an inclusive pedagogy that takes into account the diversity in students' academic backgrounds and training.

Threading Discourse within NLG. A central principle of the course is to foreground discourse phenomena across various aspects of NLG. Rather than treating discourse as a stand-alone module, we embed core concepts such as coherence (Grosz et al., 1995), discourse reference (Karttunen, 1969), anaphora (Beaver, 2004), questions under discussion (Roberts, 1996), and discourse structure (Mann and Thompson, 1988; Webber et al., 2003) into modeling and decoding techniques for NLG, as well as the evaluation of LLM outputs. This 'threading' approach allows students to see discourse as foundational building blocks for coherent and meaningful long-form outputs from language models, which further helps students think about how information is selected, structured, and realized to account for discourse-level phenomena, instead of just sentence-level correctness.

Mini Workshops for Concept Grounding. We believe it is important for students to critically investigate different viewpoints and engage in contemporary debates on LLMs' capabilities and limitations. This was achieved through two means: (1) A series of in-class group workshops where students apply analytical frameworks and perform annotations on real data. (2) Open-ended, reasoning-based questions with boilerplate code in homework assignments, allowing students to think about computational models, algorithms, and their consequences. Hands-on work also creates natural opportunities to conduct error analysis and qualitative analyses that one would include in a research paper, for example.

Exposure to Research. We believe in engaging undergraduate students in active research, not only through individual research experiences, but also *in the classroom*. The fast-paced nature of today's NLP landscape makes it especially important to do so, and we believe that exposing students to recent work that is core to discourse processing, yet not necessarily labeled so, is important in helping them recognize the interplay between theory and practice. These topics are either incorporated into course material organically, or delivered as invited talks from speakers throughout the semester. This also helps students to develop an appreciation for rigorous dataset and experimental design in research.

Exploration, Not Just Problem Solving. All assignments are intentionally open-ended, encouraging students to define the scope and direction of their work within a given instruction. This is our response to the use of Generative AI in education: the focus of assignments for upper-level undergraduate classes should incorporate high-level exploration and project design, rather than problem solving only, which is common in traditional problem sets or exams. Our structure mirrors real industry and research practice: students must select datasets, justify methodological choices, and articulate the discourse-level phenomenon they aim to improve or analyze. Such design allows students at different experience levels to contribute meaningfully while also fostering creativity and critical thinking.

Accessibility for Diverse Backgrounds. Recognizing that students may come from linguistics, computer science, information science, and other related fields, all content is designed to be accessible without extensive assumptions about prior technical knowledge. Linguistic concepts are introduced with computational relevance, highly technical material such as reinforcement learning is delivered at a level appropriate for the class goal, and programming tasks are scaffolded to accommodate varying levels of experience. We also demo common libraries for LLM inference and evaluation in class. This inclusive design supports interdisciplinary engagement and positions language-centric perspectives as equal partners to technical ones.

Collaborative Design of Content. A key feature of the course is that it is developed collaboratively by instructors and researchers from multiple institutions and disciplinary backgrounds in linguistics, cognitive science, and computer science. This distributed design process brings together diverse pedagogical traditions and research perspectives, enriching the course content and ensuring that it reflects the multifaceted nature of the 'threading' approach mentioned above. Cross-institutional collaboration also allows the curriculum to remain adaptable and responsive to evolving research, teaching practices, and student needs.

4 Course Content

Below, we provide an overview of the various aspects covered in this course, organized into themes. Note that this does *not* reflect the order of content delivery; as discussed in Section 3, we took a threaded approach across all themes. The delivery of the content was mostly built into the lectures with required and optional readings.

4.1 Content on LLMs and Generation

Language Modeling and NLG. The curriculum starts with an overview of autoregressive language modeling. The goal of this component is to solidify first principles behind language models, and help students understand how they work without a deep dive into specific techniques that improve training. We start with the next word prediction task, moving into transformer-based decoders and training LLM at scale (pre-training, instruction-tuning, and post-training). Decoding algorithms (including greedy, nucleus sampling, Holtzman et al. 2020, min-p, Minh et al. 2025) as well as the basics of long-context models (Touvron et al., 2023; Su et al., 2024) are also covered.

After this, we provide an overview of NLG by introducing students to NLG tasks from conditional generation such as summarization and translation, to more open-ended ones such as story generation, emphasizing how task open-endedness affects both modeling and evaluation.

Finally, to help students gain a deeper understanding of the inner representations of language models and their multimodal counterparts, we provided them a high-level overview of mechanistic interpretability, covering example papers that used causal mediation, activation patching (Golovanevsky et al., 2025b), early decoding, and activation steering (Golovanevsky et al., 2025a).

RL Overview and Applications. Post-training describes different training processes applied to LMs to make them more usable, better instruction followers, and more aligned with human values (Li et al., 2025b). Common post-training methods employ reinforcement learning (RL), such as RLHF (Ouyang et al., 2022) and RLVR (Lambert, 2011). Post-training and RL can also be beneficial in training an LLM to be better at discourse-related tasks. Stiennon et al. (2020), for example, use RL with human feedback to improve summarization, and similarly, RL can be used for factuality. Specifically, we give students an introduction on modern post-training techniques and RL for NLP, and then discuss recent research on using RL for discourse-related tasks such as factuality (Roit et al., 2023), question generation (Pyatkin et al., 2023), and summarization (Stiennon et al., 2020).

Demos. The course also included two class sessions devoted to coding demos, focusing on practical implementation and widespread machine learning and LM libraries. We designed the demos around giving students the opportunity to practice with the libraries we introduced, and demonstrate how to use documentation on their own. In the second week of the course we covered basic functionality of `pytorch` (Paszke et al., 2019) and `transformers` (Wolf et al., 2020), with a focus on loading LMs, tokenizing input text, and generating text. In addition, the minimal pair evaluation component was delivered in tandem with a demo involving `minicons` (Misra, 2022), a python library that facilitates straightforward, efficient, and fast computation of language model behavioral scores like log-probabilities, and is popularly used to perform minimal pair evaluation.

4.2 Content on Discourse

Discourse Structure. The course covers discourse structure as a key component for analyzing and generating coherent (long-form) text beyond individual sentences, focusing on major linguistically-motivated discourse frameworks or theories such as Rhetorical Structure Theory (RST, Mann and Thompson 1988), Penn Discourse Treebank (PDTB, Webber et al. 2019), and Questions Under Discussion (QUD, Roberts 1996). Through these frameworks, students learn how to characterize relationships between discourse units: through taxonomies (e.g., ELABORATION, CONTRAST, CAUSAL, and TEMPORAL), or through question–answer relations in the case of QUD (Wu et al., 2023a). They also learn how these frameworks differ in terms of using tree-like structures (RST and QUD) or being lexically grounded (PDTB); and how QUDs interact with reader expectations (Kehler and Rohde, 2017; Westera et al., 2020; Wu et al., 2024). Going beyond, we steer students into thinking about the challenges that lie in modeling these theories and ways to address them. For instance, discourse relations can be represented as QA pairs, helping to scale data collection (Pyatkin et al., 2020; Ko et al., 2022).

We then connect these representations to NLG applications, including: how discourse structure can inform content selection in summarization (Li et al., 2016; Xu et al., 2020; Liu and Zeldes, 2023; Trienes et al., 2025) and planning in simplification (Wu et al., 2023b; Trienes et al., 2024; Liu et al., 2025). Overall, this module emphasizes how explicitly incorporating discourse structure bridges linguistic theories and NLG tasks to enable more controlled-generation.

Entity-based Coherence. Entities introduced throughout the discourse play a central role in coherence. We introduce the classic Centering Theory (Grosz et al., 1995) which tracks the focus of attention, as well as the more flexible Centering in Optimality Theory (Beaver, 2004). Students work through examples for both accounts in a workshop style, as described in Section 3. These concepts can be directly used to analyze and assess the coherence of a document. We introduce approaches inspired by Centering Theory, such as Entity Grid (Barzilay and Lapata, 2008) and DiscoScore (Zhao et al., 2023). Conversely, generation can inform discourse: e.g., summarization informs entity salience (Lin and Zeldes, 2025).

Entity Tracking. We discuss Entity Tracking, a broad class of problems in discourse understanding that involves representing entities and their states as the discourse unfolds (Karttunen, 1969; Kamp et al., 2010, *i.a.*). Entity tracking has seen a recent reemergence in popularity, primarily as an analytical lens using which researchers can conclude about LLMs' internal representations of the 'world' (Li et al., 2021, 2023; Kim and Schuster, 2023). This content is introduced right after Centering and Entity Grid, to illustrate a rich, multi-facet view of the role of entities in discourse processing.

We specifically used material from Kim and Schuster (2023), a dataset containing textual descriptions of a scenario with entities in a given state followed by a series of state changes. This dataset provides an especially controlled set of stimuli that tease apart shallow processing—e.g., using lexical cues or heuristics that can allow the prediction of entity states without necessarily consulting with the discourse context—from *genuine* entity tracking. All in all, the proposed outcome of this component was to: 1) introduce entity tracking as an important discourse-sensitive component of long-context understanding; and 2) understand the relationship between everyday entity-state tracking and classical concepts in theories of coherence.

4.3 Evaluation

NLG Evaluation. We aim to equip students with both practical knowledge and a critical understanding of NLG evaluation, situating current evaluation practices in a broader perspective, highlighting progress since 2015 alongside persistent challenges such as robustness, replication, impact evaluation, and commercial pressures (Reiter, 2025). We discuss the evaluation of NLG systems, contrasting intrinsic evaluation (which directly measures properties of the generated text, such as fluency or likelihood via metrics like perplexity and entropy) with extrinsic evaluation (which measures downstream application performance). The course comparatively examines reference-based metrics (e.g., BLEU, Papineni et al. 2002, ROUGE, Lin 2004, BLEURT, Sellam et al. 2020, BERTScore, Zhang* et al. 2020) vs. reference-free ones such as factuality (Tang et al., 2024), LLM-as-a-judge (Bavaresco et al., 2025), and long-context evaluation (Liu et al., 2024). On a finer-grained level, we discuss how to evaluate LMs' grasp of lexical semantics, such as the discourse particle "just" (Sheffield et al., 2025).

We also discuss the central role of discourse frameworks in long-form generation tasks: the evaluation of book summaries (Chang et al., 2024), creative writing (Chakrabarty et al., 2024), discourse diversity (Namuduri et al., 2025), narrative understanding (Ahuja et al., 2025), multi-turn conversations (Laban et al., 2025), and AI-story detection (Pham et al., 2025).

Minimal Pair Evaluation. One of our earlier components in this course was to introduce students to the concept of 'minimal pair evaluation', a methodology that has been generally used in psycholinguistics-inspired analysis of models (Linzen et al., 2016; Warstadt et al., 2020; Misra et al., 2023). Here, LMs are analyzed as pairs of natural language stimuli that usually differ in a single feature. For instance, for the number-agreement phenomenon, we could compare models on *the keys to the cabinet **are** on the table* vs. *the keys to the cabinet **is** on the table*. Models are typically evaluated based on the probabilities they assign to surface forms. Once we understand their fundamental objective, which is predicting words in context, the concept of minimal pairs emerges as a natural extension for probing their capabilities. In the context of discourse, the concept most conducive to minimal pair evaluation is that of discourse connectives—words or multi-word expressions that are used to mark discourse relations between arguments given a piece of text. For example, (1) is more (discourse) coherent than (2),[7] and thus one might hypothesize it to be more likely under a competent LM's distribution.

(1) *The councilmen refused the demonstrators a permit **because** the **councilmen** feared violence.*

(2) *The councilmen refused the demonstrators a permit **because** the **demonstrators** feared violence.*

We specifically used prior psycholinguistics and LLM-evaluation work on discourse connectives (Drenhaus et al., 2014; Pandia et al., 2021; Beyer et al., 2021) as our basis for examples and assignments (see Assignment 3 in Section 5). Overall, this component introduced the students to methodology imported from the now well-established psycholinguistics toolkit to evaluate LMs (Futrell et al., 2019), and in particular adapt it to test for core components in discourse processing.

5 Assignments & Course Project

The course involved three assignments, designed to teach students to reason about broad areas that can eventually be incorporated into their course project: basic implementation and analyses of decoding algorithms; understanding the differences between base vs. instruction-tuned LMs; discourse connectives, relations, and parsing; local coherence; QUDs; and NLG evaluation. All assignments contain both programming and written portions.

Assignment 1. The first assignment focuses on getting students acclimatized to loading an LM from the `transformers` library, implementing decoding methods with them, and answering a range of analysis-based open-ended questions. We provided students with basic wrapper code to load an LM, tokenize its input text, and a general decoding function that took as its arguments the input tokens, a few arguments for decoding, and importantly a placeholder logits processing function which implemented the actual decoding algorithm. The students were then asked to implement 1) greedy decoding; 2) top-p sampling (Holtzman et al., 2020); and 3) min-p sampling (Minh et al., 2025).

Our open-ended questions involved asking students to relate top-p and min-p sampling configurations to greedy decoding, and to analyze the resulting outputs from the three algorithms, given a fixed prompt. In the context of story generation, students were asked to analyze both qualitatively and quantitatively, where they were free to implement their own functions for analysis, such as measuring surface form diversity metrics, etc. Lastly, we asked students to compare generations between a standard, next-word prediction LM vs. its instruction-tuned counterpart by implementing a perplexity calculation function and computing perplexity of each type of LM on the other's generations. Overall, the assignment uses story generation to investigate diversity, and gives students access to toolkits that could lay the groundwork for more complex discourse-sensitive analyses.

Assignment 2. The second assignment focuses on students running three prompting experiments: (1) implementing and comparing BLEU and LLM-as-a-judge (Kocmi and Federmann, 2023) for machine translation evaluation; (2) classifying discourse relations (a subset of the DISRPT 2025 shared task datasets, Braud et al. 2025, which cover a wide variety of languages and genres and are represented in a unified format across frameworks) using an LLM-as-a-judge paradigm; and (3) prompting LLMs to generate RST trees. These experiments (especially 2 and 3) are open-ended, providing students with the opportunity to practice prompt engineering using the various principles covered in lectures, as well as the opportunity to experiment with a variety of model sizes. Students were given basic boilerplate code to call models and set up basic formatting, and the data necessary to complete the assignment. Tasks 1 and 2 integrate non-English data, with Task 1 allowing students to bring in their own background by providing their own. Students were also asked to conduct an analysis given various aspects with regard to performance correlation with model size and choice of models, proposing ways to improve the performance, and the errors the model tends to make etc.

Assignment 3. The final assignment touches upon coherence, discourse connectives, and QUDs. For coherence, students were asked to use Centering Theory and Entity Grid to evaluate the coherence of book summaries generated from Wu et al. (2021). For the connectives component, we first demonstrated the concept and construction of minimal pairs (echoing our lecture material), and then presented a dataset of minimal pair sentences from Drenhaus et al. (2014), which studied humans' pro-

[7] Examples modified from the Winograd Schema Challenge (Levesque et al., 2012).

cessing of the discourse as modulated by simple changes in discourse connective (thereby adding minimality). We then asked students to load and evaluate an LM of their choice on the dataset, comparing the model's performance with chance-level accuracy, which they had to infer from the dataset, as one of the answers to a question. Finally, for QUDs, we asked students to engage with the research question in Wu et al. (2024): how do QUDs relate to the salience of potential questions? By exploring the generation of potential questions from news articles, judging their salience, and finding out whether they are QUDs, students gain an understanding of expectation-driven QUDs.

Course Project. In addition to the three assignments, this course has a final project, letting students showcase what they have learned in a project of their own design. The projects engaged in a variety of topics, e.g., studying new model architecture for generation tasks, unique designs for entity tracking, and investigating multi-lingual applications from a discourse lens. These projects are done in groups of 1-3, with several checkpoints during the course to check progress and provide feedback: (1) an initial brainstorming session, (2) a project pitch presentation, (3) a written proposal, (4) a final presentation, and (5) a final written report.

(1) While not a formal requirement, all groups met with either the instructor or the TA to discuss project ideas in the middle of the semester, in advance of the project pitch (2). This initial phase helped many groups handle project scope, and rule out ideas that were too vague/broad. Formalizing this as a course requirement is encouraged.

(2) Mid-semester, students pitched their projects in an in-class presentation of around six minutes, focusing on motivation, background, and proposed data and methodology. Some students also provided pilot results, though we did not explicitly require this. A portion of their presentation grade came from how well they were able to field questions from the audience, in addition to presentation flow and clarity.

(3) At the end of the semester, students submitted a two-page written proposal for their project. Again, students were tasked to focus on background, motivation, and proposed data and methodology—this was intended to provide them with groundwork for their final reports. This also provided students with another chance for more concrete feedback from the instructor and TA.

(4) Students presented their experimental results in the last week of class in a six-minute presentation. This presentation focused on presenting methodology and (at least some) results, and allowed the instructor, TA, and fellow students to provide additional feedback before the final write-up one week later. Students were also encouraged to explain any additional experiments they had yet to run for the final report.

(5) The final deliverable is a four-page written report. We encouraged students to follow the design of an ACL-style short paper, and encouraged them to use their proposal for background, motivation, and related work sections.

5.1 Logistics + Grading

All assignments were submitted online in the form of interactive python (e.g., jupyter) notebooks; this enabled students to write the intertwined code and analysis that many of our problems asked for.

Students were given at least two weeks from assignment release to submission. Since the course took place over 15 weeks, this allowed for pacing between assignments. The last assignment was due two-thirds of the way through the semester, allowing students to focus on the course project towards the end of the semester. In addition to office hours and in-class time for questions on homework assignments, students had access to Chatter message boards, monitored by the TA, where they could receive help from peers. We also encouraged the use of code assistants, provided students included explanatory comments. We found this to be quite reasonable, especially since the bulk of assignment questions did not focus on strict implementation, but were instead formulated as mini-experiments.

Given the open-ended design of the assignments, most grading is performed manually. There is no way around this, as most problems include high-level analysis questions, though a few have a rigid answer key. For high-level analysis questions, student grades were broken down into three main parts: *code* – did their code run without errors, *soundness* – did their code take a reasonable approach to answer the question, and *analysis* – did they provide conclusions and motivations for those conclusions using results from their code. These high-level questions either asked for specific points of analysis to be met or were open-ended, in which case they were graded relative to a baseline "good" score, with exceptional answers receiving additional points. Points were allocated primarily to analysis and then soundness, with code usually only being allocated a small portion of the points.

Additionally, since later assignments allow students to use large, closed-source models in addition to smaller, open-source ones, re-running all student code with all models they used would be infeasible in terms of both time and resources. Instead, for Assignments 2 and 3, student code was verified using a small model as a test-case. For both open-source and proprietary models, students reported prompts and outputs for grading.

6 Course Evaluation

To get insights into student reception and learning experience, we worked with the Texas Advanced Computing Center (TACC) Education Services[8] which conducted an anonymous survey among the students as formative evaluation of this course. Of the 26 students enrolled in this course at the time of the survey, 19 responded. The student body consisted of 24 undergraduate students and two PhD students.

Students enroll in a range of majors (UT Austin allows multiple majors): Computer Science, Linguistics, Mathematics, Statistics and Data Science, Informatics, Robotics, and other Liberal Arts majors. 75% of the students have prior knowledge of Machine Learning or AI; and 55% of the students have prior knowledge in Linguistics. All students have used LLMs prior to taking this course, 75% had some experience with deep learning software packages, and 70% have used cloud computing/data resources.

In terms of content difficulty, the majority of the students feel that the course is about the same or more challenging than their prior relevant coursework, with a 95% course satisfaction rate. To our delight, 80% of the students intend to apply what they learned in this course to their future courses, research, and/or career, and 75% said that this course made them want to explore new career paths or fields. In Figure 1, we show a topic-wise breakdown of perceived learning outcomes. Overwhelmingly, students found the assignments and projects valuable to their critical thinking skills and in applying theoretical concepts. In Figure 2, we show that the majority of students rated this course as having a moderate to significant impact on their academic or career plans.

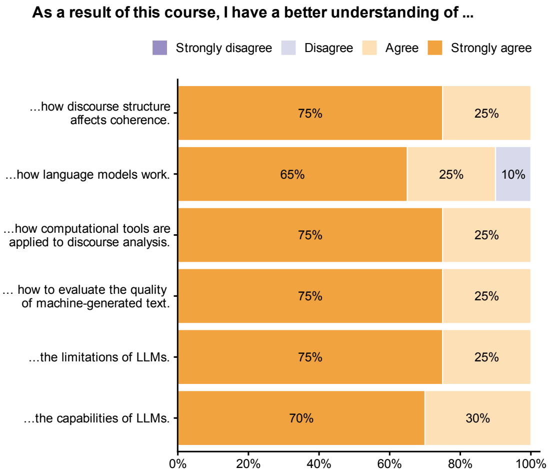

Figure 1: Perceived learning outcomes by topic.

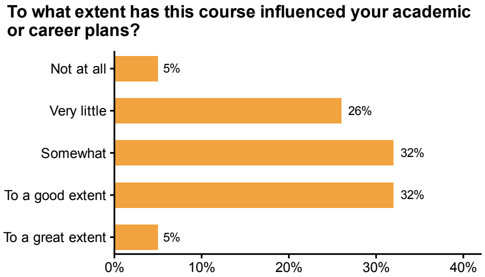

Figure 2: Academic/career impact.

7 Challenges and Future Directions

By designing and offering this class, engaging with students, and assessing what they have learned, we are confident that our main course objectives were achieved: to enable students appreciate the deep entanglement of discourse and long-form generation; to perform critical evaluation of theories and empirical findings; and to creatively analyze the capabilities and limitations of LLMs. We believe that these are important skills for the future workforce and our students recognized them as such.

One main challenge for this class is compute resources. For this first time teaching the course, we were unable to provide students compute resources beyond Colab Pro for Education.[9] While this was ultimately sufficient for most students, some assignments and projects would benefit from running models that are too large for this setup. Providing students with compute resources and/or API credits would be very helpful for this course.

Although most students used LLMs before taking the class, many were concerned about the time it took to run the experiments in the assignments, as

[8] https://tacc.utexas.edu/use-tacc/evaluation-services/

[9] https://colab.research.google.com/signup

well as the time it took for prompt-engineering. We believe this class is a great opportunity to expose them to this, but it is important to discuss these factors in the first lecture in future iterations.

In addition, our homework assignments were designed to be open-ended. This worked well and was positively received by students. However, such open-ended homework assignments did entail manual grading, which is difficult to scale up to a larger class. Innovative ways that combine the spirit of exploration in assignments and automatic grading are needed for scaling this to a larger class size.

Another challenge for this course was the varying levels of student experience with different course topics. Given its interdisciplinary nature, students were from several different majors (primarily CS and Linguistics) and so prior exposure to coding and formal linguistics varied. While office hours effectively supported individual students, additional recitation sessions may be needed as class size increases.

In terms of content, this first iteration still remains largely English-centric; we plan to expand our content to other languages in the future. Future iterations of this course could also include additional applications of discourse and NLG, such as implicit event-argument relations across sentence boundaries (Roit et al., 2024), the evaluation of long-context QA (Xu et al., 2022), long distance dependencies, and RLHF for improved coherence.

Acknowledgments

We thank Lisa Garbrecht, Stephanie Baker, and Yiwen Yang at the TACC Education Services for the formative evaluation of this course.

We thank our invited speakers: Kabir Ahuja, David Beaver, Yapei Chang, Stella Li, Jessica Lin, Chau Minh Pham, and William Rudman.

We also thank those who provided slides for various portions of this course: Anne Beyer, Antoine Bosselut, Silvia Casola, Tuhin Chakrabarty, Yejin Choi, Greg Durret, Tanya Goyal, Song Han, Julia Hockenmaier, Jaehun Jung, Najoung Kim, Philippe Laban, Nathan Lambert, Nelson F. Liu, Ramya Namuduri, Ehud Reiter, Paul Roit, Liyan Tang, and Amir Zeldes. We thank Yan Cong for providing stimuli for Assignment 3.

The development and evaluation of this course was partially supported by NSF CAREER grant IIS-2145479.

References

Kabir Ahuja, Melanie Sclar, and Yulia Tsvetkov. 2025. Finding Flawed Fictions: Evaluating Complex Reasoning in Language Models via Plot Hole Detection. In *Second Conference on Language Modeling*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation. *Preprint*, arXiv:2509.08825.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

David I Beaver. 2004. The optimization of discourse anaphora. *Linguistics and philosophy*, 27(1):3–56.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Johan J Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. 2024. Three reasons why AI doesn't model human language. *Nature*, 627(8004):489.

Chloé Braud, Yang Janet Liu, Philippe Muller, Amir Zeldes, and Chuyuan Li, editors. 2025. *Proceedings of the 4th Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2025)*. Association for Computational Linguistics, Suzhou, China.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or Artifice? Large Language Models and the False Promise of Creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The False Promise of ChatGPT. *The New York Times*.

Heiner Drenhaus, Vera Demberg, Judith Köhne, and Francesca Delogu. 2014. Incremental and predictive discourse processing based on causal and concessive discourse markers: ERP studies on German and English. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Greg Durrett, Jifan Chen, Shrey Desai, Tanya Goyal, Lucas Kabela, Yasumasa Onoe, and Jiacheng Xu. 2021. Contemporary NLP modeling in six comprehensive programming assignments. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 99–103, Online. Association for Computational Linguistics.

Jacob Eisenstein. 2019. *Introduction to Natural Language Processing*. MIT press.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, page 1–98.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. 2025a. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24848–24863, Suzhou, China. Association for Computational Linguistics.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. 2025b. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11462–11482, Albuquerque, New Mexico. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, and 3 others. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Sara Hooker. 2025. On the slow death of scaling. *Available at SSRN 5877662*.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 6, 2026.

Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2010. Discourse representation theory. In *Handbook of Philosophical Logic: Volume 15*, pages 125–394. Springer.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Lauri Karttunen. 1969. Discourse referents. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*, Sånga Säby, Sweden.

Andrew Kehler and Hannah Rohde. 2017. Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*, 54(3):219–238.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. Discourse comprehension: A question answering framework to represent sentence connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. LLMs Get Lost In Multi-Turn Conversation. *Preprint*, arXiv:2505.06120.

Mathias Lambert. 2011. Repérer les phrases évaluatives dans les articles de presse à partir d'indices et de stéréotypes d'écriture. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues (articles courts)*, pages 15–20, Montpellier, France. ATALA.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Kangming Li, Andre Niyongabo Rubungo, Xiangyun Lei, Daniel Persaud, Kamal Choudhary, Brian DeCost, Adji Bousso Dieng, and Jason Hattrick-Simpers. 2025a. Probing out-of-distribution generalization in machine learning for materials. *Communications Materials*, 6(1):9.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. 2025b. Prefpalette: Personalized preference modeling with latent attributes. In *Second Conference on Language Modeling*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jessica Lin and Amir Zeldes. 2025. GUM-SAGE: A novel dataset and approach for graded entity salience prediction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 438–455, Vienna, Austria. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025. Explanatory summarization with discourse-driven planning. *Transactions of the Association for Computational Linguistics*, 13:1146–1170.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Janet Liu and Amir Zeldes. 2023. GUMSum: Multi-genre data and evaluation for English abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9315–9327, Toronto, Canada. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Gary Marcus. 2025. "Scale Is All You Need" is dead.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *Preprint*, arXiv:2203.13112.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying discourse similarities in LLM-generated text. In *Second Conference on Language Modeling*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041.

Chau Minh Pham, Jenna Russell, Dzung Pham, and Mohit Iyyer. 2025. Frankentext: Stitching random text fragments into long-form narratives. *Preprint*, arXiv:2505.18128.

Steven T. Piantadosi. 2024. Modern language models refute chomsky's approach to language. In *From fieldwork to linguistic theory*, pages 353–414. Language Science Press.

Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *Preprint*, arXiv:2208.02957.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Ehud Reiter. 2025. NLG Evaluation 2025 vs 2015: much improved but needs to be better.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:1–69.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Paul Roit, Aviv Slobodkin, Eran Hirsch, Arie Cattan, Ayal Klein, Valentina Pyatkin, and Ido Dagan. 2024. Explicating the implicit: Argument detection beyond sentence boundaries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16394–16409, Bangkok, Thailand. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

William Sheffield, Kanishka Misra, Valentina Pyatkin, Ashwini Deo, Kyle Mahowald, and Junyi Jessy Li. 2025. Is it *JUST* semantics? a case study of discourse particle understanding in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21704–21715, Vienna, Austria. Association for Computational Linguistics.

Manfred Stede. 2012. *Discourse processing*, volume 15. Morgan & Claypool Publishers.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. InfoLossQA: Characterizing and recovering information loss in text simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294, Bangkok, Thailand. Association for Computational Linguistics.

Jan Trienes, Jörg Schlötterer, Junyi Jessy Li, and Christin Seifert. 2025. Behavioral analysis of information salience in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23428–23454, Vienna, Austria. Association for Computational Linguistics.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. *Preprint*, arXiv:2109.10862.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023a. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Yating Wu, Ritika Rajesh Mangla, Alex Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should I answer? salience prediction of inquisitive questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19969–19987, Miami, Florida, USA. Association for Computational Linguistics.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023b. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3556–3572, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.