Combating Hallucination via Structured Attribution and Edits

PI: Swabha Swayamdipta [swabhas@usc.edu] **Title:** Assistant Professor, USC Viterbi CS
Cash Funding Requested: $70,000 USD **AWS Promotional Credits Needed**: $50,000 USD
Amazon Contact: [TODO:]

Abstract

As large language models (LLMs) continue to dominate generative AI, **the challenge of hallucination in generated content needs to be combated**. Grounding generated content to supporting evidence in reliable sources, and editing it accordingly is slowly surfacing as a practical post-hoc solution. At the heart of attribution lies the alignment between generated content and retrieved supporting statements. However, the discovery of this alignment is often treated as a simple binary classification task—an ambiguous and often unreliable task—introducing foundational issues to combating LLM hallucination.

In this proposal, we propose a new post-hoc attribution and revision approach for generated content with linguistic structural annotations for higher quality alignment between generations and sources of evidence. Specifically, we propose to use propositional (denoting who did what to whom relationships) **structure-infused generations and evidential documents to automatically align common informational elements that need validation.** To this end, we will **utilize the generative power of LLMs for producing propositionally structured generations (i.e. text along with their propositional structure) at scale.** Our proposal will introduce algorithms to discover alignments between both structural annotations for robust attribution. Further, we will introduce algorithms to produce revisions for the generated content leveraging the same propositional structure, if they do not align with the evidential support. Overall, our proposal not only addresses the problem of hallucinations, but also serves as a step towards building the next paradigm of LLMs capable of generating structured information.

Keywords: Hallucination; Controlled Generation; Alignment; Linguistic Structure

1 Introduction

The very thing that makes LLMs powerful—a simple next token prediction objective trained on extreme-scale data—also promotes generalizing to content not seen before, some of which is just incorrect, non-factual, or misleading (Maynez et al., 2020). Hallucinations, or generations unsupported by factual information in the training data (Menick et al., 2022), are indeed emerging as one of the biggest challenge for generative AI, hurting user trust (Amodei et al., 2016). Factually incorrect or unattributable information provided by automated assistants might result in major legal damage (Weiser and Schweber, 2023), or even lead to misinformation (McGuffie and Newhouse, 2020). While modifying LLM training to generate text controlling for hallucinations (Gao et al., 2023b) might seem as an attractive solution, in practice this becomes infeasible given limited accessibility of LLMs, not to mention resource constraints.

Our proposal outlines a practical alternative: **attribute generations of existing LLMs to supporting evidence in retrieved documents** from trusted sources on the web (Rashkin et al., 2021; Bohnet et al., 2023; Gao et al., 2023a). Such attribution reports may result in both an increase in user trust as well as helps in the assessment of correctness of the generation, while allowing the underlying LLM to generate different kinds of text. Moreover, we propose to use these attribution reports to **further edit the generations to remove factual inconsistencies** with retrieved sources of knowledge. Thus, we will be combating hallucinations due to language generation through a model-agnostic, post-hoc approach which is robust to LLM training data issues. Moreover, our approach to combat hallucinations has implications for building reliable generative models, such as question answering agents, conversational assistants and summarization systems.

At the heart of our approach lies propositional semantic alignment, i.e. alignment between predicate-argument structures between a pair of sentences; see Figure 1 for an example. The particular flavor of propositional (predicate-argument) semantics we consider is frame semantics Ruppenhofer et al. (2016), where events and situations (predicates) are represented by semantic frames and their participants as frame-elements, or arguments to the frame. This structure offers a natural, fine-

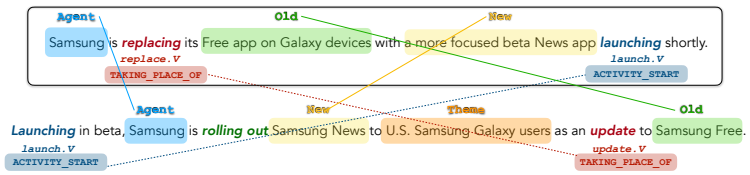

Figure 1: Our proposed alignment between two sentences using structural annotations. Alignments are with respect to the predicates 'replacing' and 'update' in the two sentences, respectively; both correspond to the `Taking_Place_Of` frame. Further alignment is shown between the named arguments (frame-elements) under the frame, such as `Agent`, `Old` and `New`.

grained alignment in terms of predicates and their named arguments. Compared to prior approaches that rely on simple binary classification based alignment for fact checking (Gao et al., 2023a), our approach offers robust solutions to alignment.

Our proposed research encompasses three distinct research thrusts, each characterized by its ambitious scope (§2). In each thrust, we endeavor to advance the state of the art while addressing a range of scientific challenges. In particular, our research will advance the understanding and capabilities of LLMs and expand their abilities to provide a strong foundation to combat an issue at their own core: hallucinations. Thrust 1 focuses on building few-shot models for predicting the propositional structure underlying natural language (§2.1). Building such models entails pushing the frontiers of the literature in two diverse research areas and their intersection: enabling LLM capabilities towards a core task in natural language processing and structure generation. Thrust 2 seeks to build on the structured representations produced in Thrust 1 by finding alignment between them to create an attribution report (§2.2). This will enable verifying the factuality of generations, and testing whether they contain hallucinations, in addition to citing evidence in real sources where there is a factual basis; this has direct impacts on user trust by justifying what part of the generation is verifiable and how. Finally, Thrust 3 will leverage the same structures for revising the generation wherever there is a lack of attribution (§2.3). This will indirectly enable LLMs to be grounded in factual evidence and be a step towards handling factual inconsistencies and even misinformation. Our proposed research has the potential to substantially improve the performance and utility of generative AI.

2 Methods

Our overall research plan for combating hallucinations can be broken down into three thrusts, described below. Each introduces novel technical approaches to solving the sub-problems:

1. GPT-3 based frame-semantic structure prediction (§2.1);
2. Attribution via Frame-Semantic Structural Overlaps (§2.2); and
3. Revisions via Frame-Semantic Structural Manipulations (§2.3).

The input to the system is a query, x, the generated answer, y and a set of evidence statements E obtained through information retrieval (e.g., from the web). The output is a revised answer y' along with an attribution report A, which contains evidence snippets $e_1, \ldots e_n \in E$ that support y' and their respective alignments, $a_1, \ldots a_n$, where each a_i is a tuple contain a span from y', the frame of alignment, f and the frame-element, l corresponding to the span. This specification makes our framework compatible with general LLMs, as well as retrieval language models, such as REALM (Guu et al., 2020).

2.1 Thrust 1: Generating Propositional Structure-Enhanced Text with LLMs

In order to leverage frame-semantic structural information towards robust content overlap for text attribution, we need access to reliable frame-semantic structures for arbitrarily complex sentences. We aim to produce accurate propositional frame-semantic structures for sentences in the generations as well as in the evidential documents which will allow us to process attributions at scale. However, existing state-of-the-art frame-semantic role labelers treat the task of frame-semantic role labeling as a pipeline of three tasks: target identification, frame identification and frame-semantic role labeling. For instance, PI Swayamdipta's prior work proposes neural models for the frame-semantic

prediction, which while being accurate also treat the structure as a graph, therefore needing quadratic time inference algorithms (Swayamdipta et al., 2018, 2017; Pancholy et al., 2021). Other end-to-end neural approaches also rely on graph-based methods, significantly reducing their efficiency (Lin et al., 2021). However, for practical and scalable attribution, we need methods that can produce frame-semantic structures efficiently (e.g. in linear time).

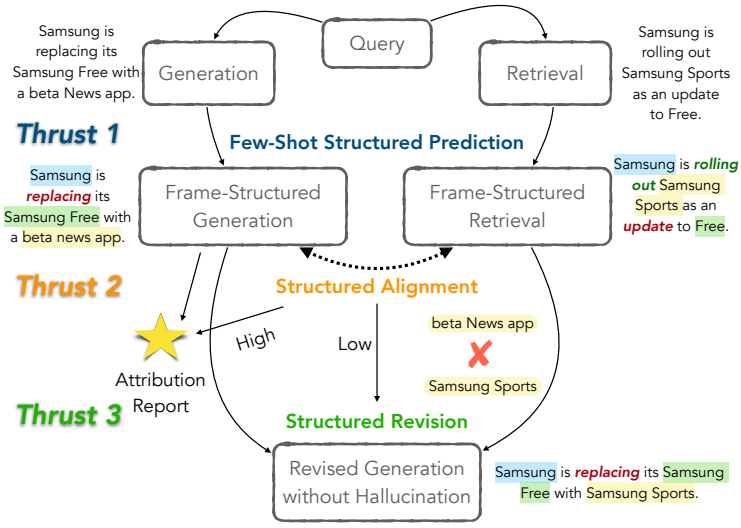

Figure 2: Illustration of research plan overview.

Motivated by this, we propose a new approach—few-shot structured prompting—that will predict as a sequence the frame-semantic analysis of a given sentence based on a few exemplar prompts or task demonstrations. We will use an autoregressive LLM for the task, which will result in a linear time method to produce the frame-semantic analysis, a significant speedup compared to existing approaches. Not only will this provide us the basis for combating hallucinations through attribution and revision, but also serves as a core task under NLP that demonstrates the capabilities of existing LLMs in understanding natural language. We believe our task is a concrete test of NLU as mapping tokens of text into propositional structures indicates an ability to process symbols at a level of abstraction offered by that flavor of semantic formalism.

The task of generating a sentence enhanced with its propositional frame-semantic structure needs to be done in two phases. While it is possible to pose the task as prediction of all possible frame-semantic analyses in the sentence simultaneously, that makes the task unduly complex because of all overlapping arguments within the sentence; see Figure 1. As a first phase, we will perform few-shot target and frame identification, which corresponds to predicting all potential predicates in the sentence, along with the semantic frames they evoke. As the second phase, for each predicted propositional predicate, we will detect the spans in the sentence corresponding to its arguments, or frame elements. This makes frame-semantic role labeling essentially a span prediction task. Under this phase, we will experiment with two formats: (i) representing the span structure with XML tags, or (ii) labeling each token in the span with a BIO label, following Blevins et al. (2023). Under (ii), given a span of m tokens labeled L (corresponding to the name of the frame-element of interest), we will label the first token as the beginning of the span with 'B-L', the remaining $m-1$ tokens are labeled as inside the span with 'I-L', and tokens not included in the span are labeled as outside the span or 'O'.

In either phase, we propose to use a lightweight prompt format with limited natural language guidance, in the form of instructions, about the task provided to the LLM. The model will be given k (sentence, tagged sentence) pairs as the task demonstration and the example sentence to be labeled, where k is determined based on the length of allowed context under the LLM. The examples in the demonstration may be derived from existing lexicographic annotations in FrameNet; we will use randomly chosen input-structured output pairs under settings (i) and (ii).

Overall, we expect performance to be stable across the varied prompt formats and choice of delimiter and other minor formatting differences, but special tokens in the instruction might be key.

2.2 Thrust 2: Attribution via Frame-Semantic Structural Overlaps

The next challenge in our system is the generation of an attribution report, A, given the generated answer, y and a set of evidence statements E obtained through information retrieval (e.g., from the web). A must contain alignment between snippets in the generation y and the evidence statements $e_1, \ldots e_n \in E$, $a_1, \ldots a_n$, where each a_i is a tuple contain a span from y', the frame of alignment, f and the frame-element, l corresponding to the span. As seen in Figure 1, propositional frame semantics naturally offers a way to consider whether or not two sentences refer to the same events, simply via matching the labeled elements (frame types, frame element types and spans) between two sentences. Moreover, FrameNet has a rich hierarchy of frames where similar frames are connected to common ancestors (e.g. frames AESTHETICS and POPULARITY are both descendants of the frame DESIRABILITY)—information we will leverage for matching.

We will consider a sliding window of sentences in the evidence documents E for alignment. For efficiency, the first pass of alignment will only consider first sequence-to-sequence phase from Thrust 1 (§2.1): predicting all potential predicates in the sentence, along with the semantic frames they evoke. If we notice an overlap in the structured elements in this phases, we will proceed to the second phase in Thrust 1: predicting the spans that correspond to different arguments or roles (FEs) under the same predicate / frame pair. Alignment of predicates does not guarantee alignment of FEs, given some instantiations might be refering to different events of the same type. In this phase, we will consider a simple classification or semantic similarity matching of the aligned FE spans to determine if they correspond to identical elements.

The other requirement from this thrust is identification of propositional arguments which do not align, despite the alignment of predicates and frames. Such spans potentially correspond to hallucinations since they cannot be supported by any evidence in the retrieved documents, E, and are prime candidates for replacement for revising generations in Thrust 3 (§2.3).

Importantly, our approach for attribution is more robust compared to coreference resolution which simply identifies all entities in a document without specifying how the entities may be related. For instance, a system relying on coreference for matching may find perfect alignment between the sentences "`Barack Obama` was born in `Hawaii`" and "`Barack Obama` was not born in `Hawaii`", since the key entities are identical, though joined via a contradictory relationship. Our approach finds alignments in terms of predicates and their corresponding frames, and would recognize the negation as an important semantic feature in under propositional frame semantics.

Prior work on attribution or fact checking has treated the agreement as a simple classification problem (Gao et al., 2023a; Honovich et al., 2022). Many have considered the task of natural language inference (NLI): given a premise and a hypothesis, infer whether the hypothesis entails, contradicts or is neutral to the premise; in this case, the premise corresponds to the sentence from the retrieved evidence, E. However, NLI is prone to many issues arising due to underspecification and ambiguity, as shown by the PI's prior work (Gururangan et al., 2018; Liu et al., 2023).

Prior work on structural agreement has been fewer and far in between. Brook Weiss et al. (2021) propose QA-Align, where they consider QA-SRL structures (He et al., 2015) for overlaps between sentences to compress information for multi-document summarization. However their primary focus is on verbal predicates, whereas we target all predicates: nominal, verbal, prepositional and so on. Moreover, in QA-SRL the questions can be arbitrarily vague (as they were primarily designed for ease of human annotation) introducing challenges in alignment.

2.3 Thrust 3: Revisions via Frame-Semantic Structural Manipulations

LLMs struggle to produce content that adheres to evidence (Dziri et al., 2022). Even retrieval-augmented models (Guu et al., 2020; Lewis et al., 2020) designed to retrieve relevant documents and generate responses to a query conditioning on the same, are known to struggle with attribution. In particular, the issues include generating additional (irrelevant or incorrect) information outside the retrieved documents (Dziri et al., 2022), ignoring (Krishna et al., 2021) or contradicting (Longpre et al., 2021) the document contents. Prior work on attribution has primarily been influenced by fact

checking (Thorne et al., 2018; Schuster et al., 2021; Thorne and Vlachos, 2021) where workflows are designed to attribute and / or correct unattributed claims made by humans.

As our final challenge, we consider editing a generation y to produce a revision y' that preserves all the attributed facts and removes unattributable snippets of information. Crucially, we will execute a model for revision if and only if a disagreement is detected via a misalignment from the methods in Thrust 2 (§2.2). Prior work on this task use few-shot prompting and chain-of-thought prompting to seek edits (Gao et al., 2023a), where the model is asked to first identify a particular span in the generation that needs to be edited before generating the revision. While this helps reduce the editor's deviation from the original generation, it also preserves unverifiable content due to the lack of robust alignment, as well as known issues with chain-of-thought prompting (Turpin et al., 2023).

In contrast, our revision is based on removing and replacing spans corresponding to frame elements which are misaligned. Figure 2 illustrates the idea of replacement: the structured misalignment of frame element spans, `beta News app` and `Samsung Sports` results in replacing that span in the final generation with its retrieved counterpart. For FEs detected in the generation which cannot be aligned to other FEs in the evidence, we will simply remove the resulting FE span from the generation. Following a removal, we will use few-shot prompting to achieve grammar error correction, as an additional check to ensure fluency. Note that replacement does not necessitate any fluency errors because of the nature of semantic arguments: similar arguments fill similar phrase types.

3 Expected Results

We hypothesize that using propositional semantics for matching components will result in more reliable attribution and revisions to remove hallucination. For Thrust 1, we will test the ability of LLMs towards NLU by evaluating the validity of the generated frame-semantic analyses on a labeled F1 score for the frame-structural annotations produced by LLMs in zero or few-shot settings. Our baselines for comparison will be existing state-of-the-art frame-semantic parsers, such as Lin et al. (2021). We expect our method to be more efficient in producing frame-semantic structural annotations, while not sacrificing accuracy due to the capabilities of LLMs.

For Thrusts 2 and 3, we will follow the evaluation setup from Gao et al. (2023a) using the Attributable to Identified Sources (AIS) metric from Rashkin et al. (2021). Specifically, given an evidence set A and a revised generation y', AIS assigns a binary judgment to validate whether "E supports y'". Crucially, AIS is not only important for attribution but also preservation: how much the revised text y' preserves other aspects of the original generation y. A system either receives full credit (1.0) if all content in y' can be attributed to E, and no credit (0.0) otherwise. We will report the micro-average of AIS across all sentences in a document, as well as the macro-average for the entire document. Our baseline for comparison will be the RARR model from Gao et al. (2023a) on the benchmark released by the authors. We expect our approach to perform better than baselines as well as be more robust to failure modes in natural language inference models, which form the basis of prior work on attribution and revision.

Months 1-3	Thrust 1: Generating Propositional Structure-Enhanced Text with LLMs
Months 4	Evaluation of Thrust 1 and Paper Writing
Months 5-7	Thrust 2: Attribution via Frame-Semantic Structural Overlaps
Months 8-10	Thrust 3: Revisions via Frame-Semantic Structural Manipulations
Months 11-12	Evaluation of Thrusts 2 and 3 and Paper Writing

Table 1: Milestones and Timeline of Proposed Research. We propose to publish papers on each research thrust in appropriate high-tier venues.

Our timeline of research is described under Table 1; we seek to produce two publications from this work, which we hope to present at top-tier NLP/ML conferences. We will release all generated data and code repositories publicly.

4 Funds Needed

The cash funding will be used to support one Ph.D. student for one year, and partial summer salary for

the PI, as outlined in Table 2. The AWS Promotional Credits ($50,000) will be used for dataset collection and model training. Specifically, the AWS products that we plan to use include: Sagemaker Ground Truth with Mechanical Turk for human validation and verification of the generated data; Sagemaker Studio Notebooks for model development, collaboration and sanity checks; Sagemaker Training and Sagemaker Neo for large-scale model training as well as hyperparameter tuning; and EC2, EBS, and S3 for additional compute and storage for our large-scale experiments. Most of the development will be on AWS PyTorch or Apache MXNet, available through AWS.

Item	**Amount**	**Explanation**
Graduate Research Assistant (GRA) salary	$41,000	50% effort, 12 months
Tuition remission for GRA	$14,907	6 units/ year / GRA at $2,424 / unit
PI salary	$10,557	For [TODO: *??*] summer months
Fringe benefits	$3,536	Fringe rate 33.5% $\times$ $10,557
Total Cost to Amazon	$70,000	(Indirect costs of $10,500 not included)

Table 2: Budget justification for cash funding.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, et al. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9879–9894, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations.

Suchin Gururangan, **Swabha Swayamdipta**, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

ZhiChao Lin, Yueheng Sun, and Meishan Zhang. 2021. A graph-based neural model for end-to-end frame semantic parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3864–3874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, **Swabha Swayamdipta**, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

Ayush Pancholy, Miriam R L Petruck, and **Swabha Swayamdipta**. 2021. Sister help: Data augmentation for frame-semantic role labeling. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 78–84, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. 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 704–718, Online. Association for Computational Linguistics.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI: Berkeley.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. 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 3298–3309, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

Benjamin Weiser and Nate Schweber. 2023. The ChatGPT Lawyer Explains Himself. `https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html`. Accessed: 2023-08-17.

CV of PI Swayamdipta

Professional Experience

- University of Southern California: Assistant Professor of Computer Science and Gabilan Assistant Professor (2022-Present)
- Allen Institute for Artificial Intelligence: Young Investigator (2019–2022)
- Columbia University: Research Assistant (2012-2013)
- Oracle Server Technologies: Member Technical Staff (2010-2011)

Education

- Carnegie Mellon University: Ph.D. in Language and Information Technologies (2019)

- Columbia University: M.S. in Computer Science (2012)
- National Institute of Technology, Calicut: B.Tech. in Computer Science and Engineering (2010)

Awards

- Intel 2023 Rising Star Award from Intel Labs
- ACL 2022 Young Rising Star for Invited Spotlight Presentation
- Young Investigator Research Award: Allen Institute for AI
- Outstanding Paper Award: ICML 2022
  - Ethayarajh, K., Choi, Y., & Swayamdipta, S. (2022). Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. *Proceedings of the 39th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 162:5988-6008.
- WiSE Gabilan Assistant Professorship at USC
- Outstanding Paper Award: NeurIPS 2021
  - Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., & Harchaoui, Z. (2021). Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34, 4816-4828.
- Honorable Mention for Best Paper: ACL 2020
  - Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342-8360).
- EECS Rising Star, UIUC: 2019

Publications

Over 30 peer-reviewed publications in journals and conferences (2012-present; the most relevant are cited in references above, and a full list is available online); Google Scholar estimates an h-index of 23 and 5,562 citations (as of Sep 13, 2023).

Mentoring

- Current: 3 Ph.D. advisees; 2 MS advisees; 5 undergraduate advisees.
- Past: 17 interns, undergraduate, M.S., and Ph.D. student collaborators while a postdoctoral young investigator at the Allen Institute for AI.

Service

- Member of DEI Committee / USC Thomas Lord Department of Computer Science 2023-Present;
- Area Chair for ICLR 2023, ACL 2023, EMNLP 2021-2023, NAACL 2021, EACL 2021;
- Standing reviewer for ACL-ARR (2021-present); Standing reviewer for Computational Linguistics (2020-present);
- Standing reviewer for Transactions of the ACL (2021-present);
- Workshop Organizer for Deep Learning for Low Resource NLP (2019; 2022);
- Reviewer for ACL (2015-2020);
- Reviewer for NAACL (2015-2020);
- Reviewer for EMNLP (2015-2020);
- Reviewer for NeurIPS (2018-2020);
- Reviewer for ICML (2015, 2019, 2020, 2022, 2023);
- Reviewer for EACL (2017);
- Reviewer for AAAI (2017-2020);
- Reviewer for CoNLL (2017-2018, 2020)