Paper Lantern

Candidate Research Brief

Publication Analysis & Interview Strategy

Alyssa Hwang

Research Role

CONTENTS

This brief surfaces interview-relevant signals from publication records. Use as a starting point for discussion, not a definitive verdict. All assessments are derived from publicly available research outputs.

Generated 2026-01-08

EXECUTIVE SNAPSHOT

RECOMMENDED ACTION

PRIORITY INTERVIEW

CONFIDENCE

Very High

WHY NOW

The market is shifting from 'Training Large Models' to 'Trusting and Using Models.' Alyssa sits exactly at this intersection. She builds the benchmarks that prove models fail (RAID, FanOutQA) and the user interfaces that help humans verify model outputs (Attribution Gradients, Ivie). She is an ideal hire for a Product AI or Trust & Safety team.

Publication Stats

Total Publications
14

Active Years
2017–2025

Key Venues
ACL, CHI, EMNLP

Primary Type
Both Industry and Academia

Publication Trajectory

- **2017–2020**
  Columbia University
- **2021–2025**
  University of Pennsylvania
- **2024 (approx)**
  AWS AI Labs (Internship)

Key Research Areas

- LLM Evaluation & Red Teaming
- Human-Computer Interaction (HCI) for AI
- Retrieval Augmented Generation (RAG)
- Synthetic Data Engineering

Authorship Pattern

> Highly collaborative. Frequent First Author on qualitative/experimental papers (NewsQs, Grounded Intuition, Rewriting the Script). Strong Second Author contributions on heavy engineering benchmarks (RAID, FanOutQA, Ivie), indicating she is likely the 'Architect' or 'Lead Evaluator' in team settings.

Key Co-authors

Name	Co-authored Papers	Citations	Co-author Impact
Chris Callison-Burch	6	35,099	Exceptional
Andrew Head	4	2,188	High
Kathleen McKeown	3	34,811	Exceptional
Liam Dugan	3	2,691	Very high

Verified Technical Audit

LLM Red Teaming & Adversarial Evaluation

She can scientifically break your models. She proved that simple tweaks (like repetition penalties) allow AI text to bypass detectors. Essential for Trust & Safety.

Evidence: RAID (2025)

RAG System Engineering & Verification

She understands RAG failure modes ('Context Forgetting') and has built UI solutions to force models to cite their sources correctly.

Evidence: Attribution Gradients (2025) & FanOutQA (2024)

Synthetic Data Generation

Experience generating 21k+ high-quality training examples using 'Control Codes' and NLI filtering. Critical for fine-tuning models when data is scarce.

Evidence: NewsQs (2024)

Capability Scorecard

Engineering Maturity 7 ± 1 / 10

Built working end-to-end systems: VS Code Extensions (Ivie), React frontends (Attribution Gradients), and large-scale data pipelines (NewsQs). Not limited to Jupyter notebooks.

Research Autonomy 8 ± 1 / 10

First author on internship project at AWS (NewsQs) and multiple lab papers (Grounded Intuition), demonstrating ability to drive ambiguity to publication.

Innovation Style 7 ± 2 / 10

Innovates via 'Methodology' and 'System Design' rather than new model architectures. She invents ways to measure and interact with AI, rather than the AI itself.

Domain Breadth 9 ± 1 / 10

Exceptional range: has published on Voice (Rewriting the Script), Vision (Grounded Intuition), Code (Ivie), and Text (RAID).

COMPETITIVE ADVANTAGE

The 'Vibe Check' Quantified.

Most engineers rely on automated metrics (BLEU/ROUGE). Alyssa specializes in 'Qualitative Rigor'—creating systems (Grounded Theory, Human-in-the-loop UIs) that catch the errors automated metrics miss. She brings a Human-Factor rigor that pure ML engineers lack.

POTENTIAL CONCERN

Is she a 'System Architect' or a 'Contributor'?

Mitigation Strategy:
In recent massive benchmarks (RAID) or frameworks (Kani), she is 2nd/3rd author. However, her First Author work (NewsQs, Grounded Intuition) proves she can lead. The interview must probe which specific modules of the larger systems she owned.

INTERVIEW STRATEGY

Interview Questions

Question 1: RAG Verification & UX

> **Context:**
> RAG (Chat with your Data) is a hot topic. A common failure is the model citing a document that doesn't actually contain the answer. This tests if she understands the *root cause* of these errors.

Ask This:
"In your 'Attribution Gradients' paper (2025), you built a system to verify RAG citations. What was the most common technical reason citations failed to support the claim (e.g., retrieval error vs. synthesis error), and how did your UI design help users catch that?"

LOOK FOR

- Hallucination of 2nd degree citations
- Synthesis errors (combining two true facts into a lie)
- UI linking specific text spans to PDF highlights

WARNING SIGNS

- Blaming the user
- Inability to distinguish between retrieval failure (bad search) and generation failure (bad summary)

Question 2: Adversarial Robustness

> **Context:**
> This tests deep technical understanding of how LLMs generate text. 'Repetition penalty' stops the model from saying the same thing twice. She needs to explain why this specific tweak fools safety systems.

Ask This:
"In the RAID benchmark (2025), you showed that changing 'repetition penalties' broke AI detectors. Why does such a simple parameter change make AI text look 'human' to a classifier?"

LOOK FOR

- Decoding strategies
- Distribution shift
- Detectors relying on perplexity/burstiness patterns that are disrupted by penalties

WARNING SIGNS

- Vague answers about 'model confusion'
- Failure to mention 'decoding strategies' or 'sampling'

Question 3: Synthetic Data Quality

> **Context:**
> We need to know if she relies on old methods or understands modern trade-offs. Control Codes are 'hard constraints.' Prompting is 'soft.' A good answer balances cost, reliability, and model capability.

Ask This:
"For NewsQs, you used 'Control Codes' to guide question generation. If you were building this pipeline today with GPT-4, would you still need Control Codes, or is prompt engineering enough? What's the trade-off?"

LOOK FOR

- Reliability/Determinism
- Cost (Fine-tuning T5 is cheaper than running GPT-4)
- NLI Filtering is still necessary regardless of the model

WARNING SIGNS

- Blind faith that GPT-4 solves everything
- Ignoring the cost/latency aspect of large models

Avoid These Topics

Topic 1. QuakerBot (2021)

This is older work with 14 authors. It's better to focus on her recent, high-impact work from 2024/2025.

Topic 2. Basic Sentiment Analysis

Her 2019 papers on idioms are linguistically interesting but technically obsolete compared to her modern GenAI work.

PAPER 1

Jul 2025

RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors

Liam Dugan, **Alyssa Hwang,** Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, Chris Callison-Burch

Paper Summary: In AI Content Moderation, the fragility of current detectors limits trust. This paper solves it by introducing RAID, a massive stress-test benchmark (6M+ samples) incorporating adversarial attacks and varied decoding strategies, achieving definitive proof that detectors fail against simple evasion techniques.

Authorship Explanation: The candidate is the second author in a large academic collaboration (8 authors). The first author (Dugan) and senior author (Callison-Burch) are from the same institution (UPenn). Second authorship typically indicates a primary contributor role, likely driving significant portions of the data generation pipeline, prompt engineering, or experimental execution alongside the lead.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●●●●	5
Clarity	●●●●○	4
Relevance	●●●●○	4

Research Signature

DOMAIN	AI Detection NLP Evaluation Adversarial ML
TECHNIQUES	Adversarial Attacks Repetition Penalty Analysis Zero-Shot Prompting
OBJECTIVES	Benchmarking Robustness Detecting Machine-Generated Text Red-Teaming

Recruiter Questions

Question 1: Assess Understanding of Model Inference

> **Context:** The paper found that changing how a model picks words (decoding strategy) breaks detectors. Specifically, 'repetition penalties' made text harder to detect. A good candidate should understand how inference parameters affect output quality.

Ask This: "Your paper highlights that 'repetition penalties' significantly hurt detector accuracy. Can you explain why penalizing repetition makes AI text look more human to these classifiers?"

Question 2: Verify Data Pipeline Experience

> **Context:** Generating 6 million samples is an engineering challenge. We want to know if they handled the technical execution or just analyzed the CSVs.

Ask This: "As the second author on RAID, what was your specific role in the data generation pipeline? How did you handle the scale of generating 6 million samples across different APIs and local models?"

PAPER 2

Jul 2023

Rewriting the Script: Adapting Text Instructions for Voice Interaction

Alyssa Hwang, Natasha Oza, Chris Callison-Burch, Andrew Head

Paper Summary: In Voice Assistants, reading text verbatim overwhelms users. This paper solves it by identifying 9 friction points and proposing 8 NLP-driven text transformations (e.g., splitting, summarizing), achieving a blueprint for audio-native instruction delivery.

Authorship Explanation: The candidate is the first author and lead researcher. She designed the observational study, conducted the qualitative analysis of user breakdowns, and developed the 'Rewrite the Script' framework connecting HCI needs to specific NLP tasks. The senior authors (Callison-Burch, Head) provided advisory support in NLP and HCI respectively.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●●●○	4
Clarity	●●●●●	5
Relevance	●●●●○	4

Research Signature

DOMAIN	Voice User Interfaces (VUI) Human-Computer Interaction Task-Oriented Dialogue
TECHNIQUES	Thematic Analysis Observational Study Text Simplification Procedural Text Summarization
OBJECTIVES	Reduce Cognitive Load Improve Instruction Following Adapt Text for Audio

Recruiter Questions

Question 1: Verify Product Sense for AI

> **Context:** The candidate found that Voice Assistants fail because they just read text from the web. A good answer should explain why 'direct translation' from text to audio is bad.

Ask This: "In your DIS '23 paper, you argued that voice assistants shouldn't just read text aloud. Can you explain the 'Time Insensitivity' challenge you found, and how that applies to modern real-time voice agents like GPT-4o?"

PAPER 3

Jun 2023

Large Language Models as Sous Chefs: Revising Recipes with GPT-3

Alyssa Hwang, Bryan Li, Zhaoyi Hou, Dan Roth

Paper Summary: In **Instruction Following**, **Complexity and Ambiguity** confuse users. This paper solves it by **Grounding LLM rewrites in ingredient lists**, achieving **62.5% user preference over original text**.

Authorship Explanation: The candidate is the first listed author among three equal contributors (Joint Lead). The paper explicitly notes this work was conducted as a final project for a Machine Reasoning course (CIS 7000-007) at UPenn. The candidate likely shared the workload of prompt engineering, MTurk interface design, and data analysis equally with the other two student leads.

Quality Scorecard

Novelty	●●○○○	2
Rigor	●●●○○	3
Clarity	●●●●○	4
Relevance	●●●○○	3

Research Signature

DOMAIN	Natural Language Generation
TECHNIQUES	Human-Computer Interaction Prompt Engineering In-Context Learning Chain-of-Thought (Implicit) Human Evaluation (MTurk)
OBJECTIVES	Text Simplification Hallucination Reduction Instruction Following

Recruiter Questions

Question 1: Verify Evaluation Rigor

> **Context:** Evaluating AI text is hard because humans get tired reading long documents. This candidate created a method to check text one sentence at a time. We want to know if they understand why this matters.

Ask This: “In your recipe paper, you mentioned that side-by-side comparison was too strenuous for annotators. Can you explain how your 'step-by-step' evaluation method improved the quality of your data?”

Question 2: Check Hallucination Mitigation

> **Context:** LLMs lie (hallucinate). The candidate fixed this by giving the model a list of 'ingredients' it had to stick to. This is similar to how we give models data to prevent lying in business apps.

Ask This: “You found that including the ingredient list in the prompt reduced hallucinations. How would you apply this 'grounding' concept to a business chatbot that answers questions based on PDF manuals?”

PAPER 4 Nov 2023

Grounded Intuition of GPT-Vision's Abilities with Scientific Images

Alyssa Hwang, Andrew Head, Chris Callison-Burch

Paper Summary: In **Multimodal Evaluation**, **reliance on aggregate metrics** masks critical safety failures. This paper solves it by **adapting social science Grounded Theory for rigorous model auditing**, achieving **a precise taxonomy of GPT-Vision's spatial and OCR limitations**.

Authorship Explanation: The candidate is the first author and lead researcher. The paper analyzes images from a previous study also authored by the candidate (Hwang et al., 2023), indicating she generated the dataset, performed the qualitative coding (Grounded Theory), and wrote the manuscript under the supervision of the subsequent authors.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●●●○	4
Clarity	●●●●●	5
Relevance	●●●●○	4

Research Signature

DOMAIN	Multimodal AI Computer Vision HCI Model Evaluation
TECHNIQUES	Grounded Theory Thematic Analysis Prompt Engineering Qualitative Auditing
OBJECTIVES	Alt Text Generation Hallucination Detection Spatial Reasoning Evaluation

Recruiter Questions

Question 1: Verify Evaluation Rigor

> **Context:** Engineers often rely on automated scores (like accuracy numbers) that miss the nuance of why a model fails. This candidate created a manual process to find those failures. We want to know if they can scale this.

Ask This: "You adapted 'Grounded Theory' to evaluate GPT-Vision. How would you scale this qualitative rigor when evaluating a model on 10,000 images instead of 20? How do you automate the 'intuition'?"

Question 2: Check Multimodal Intuition

> **Context:** The paper mentions the model struggles with spatial relationships (left vs right). This is a common problem in AI agents.

Ask This: "Your paper noted GPT-Vision struggles with spatial boundaries and specific OCR tasks. In a modern agentic workflow (e.g., a bot reading a screen), how would you architect a safeguard against these specific hallucinations?"

PAPER 5 Aug 2019

Confirming the Non-Compositionality of Idioms for Sentiment Analysis

Alyssa Hwang, Christopher Hidey

Paper Summary: In Sentiment Analysis, treating idioms (e.g., 'break a leg') as individual words causes classification errors. This paper statistically proves that an idiom's sentiment is unrelated to its component words, validating the need for specialized tokenization or handling of Multiword Expressions.

Authorship Explanation: The candidate is the first author, indicating they led the experimentation, data analysis, and writing. The second author appears to be a collaborator or mentor within the same department.

Quality Scorecard

Novelty	●●○○○	2
Rigor	●●●○○	3
Clarity	●●●●○	4
Relevance	●●●○○	3

Research Signature

DOMAIN	Natural Language Processing Sentiment Analysis Lexical Semantics
TECHNIQUES	Spearman Correlation Dictionary of Affect in Language (DAL) WordNet Integration
OBJECTIVES	Hypothesis Testing Multiword Expression (MWE) Analysis

Recruiter Questions

Question 1: Verify Understanding of NLP Fundamentals

> **Context:** Models often fail when words change meaning when put together (like 'hot dog'). This candidate studied that exact problem. Ask them why this matters for modern AI.

Ask This: “You analyzed how idioms break sentiment models. How does this problem manifest in modern Large Language Models, and does 'Attention' solve it completely?”

PAPER 6

Dec 2020

Towards Augmenting Lexical Resources for Slang and African American English

Alyssa Hwang, William R. Frey, Kathleen McKeown

Paper Summary: In Social Media NLP, the lack of dictionaries for slang and dialects (AAE) limits analysis. This paper solves it by clustering word embeddings to infer meanings of unknown words from context, achieving high-quality semantic groupings verified by human experts.

Authorship Explanation: The candidate is the first author, indicating they led the experimentation, writing, and analysis. The second author is from the School of Social Work (providing domain expertise on AAE/Sociolinguistics), and the last author is a senior professor (McKeown) providing supervision.

Quality Scorecard

Novelty	●●○○○	2
Rigor	●●●●○	4
Clarity	●●●●●	5
Relevance	●●●○○	3

Research Signature

DOMAIN	Natural Language Processing Computational Sociolinguistics Social Media Analysis
TECHNIQUES	Unsupervised Clustering K-Means Word Embeddings (GloVe) Human Evaluation
OBJECTIVES	Lexical Resource Augmentation Slang Detection Dialect Representation (AAE)

Recruiter Questions

Question 1: Handling Noisy Data

> **Context:** The candidate worked with slang and dialects that don't appear in standard dictionaries. This requires creativity in how to process data that looks 'broken' to standard tools. A good answer involves not just throwing data away, but finding ways to learn from it.

Ask This: "You worked with African American English and slang, which often breaks standard NLP tools. How did you handle the high rate of 'unknown' words, and how would you apply that to cleaning messy user data in our product?"

Question 2: Metric Design

> **Context:** They invented a new score because the standard ones were wrong. We want to see if they can define success metrics when the 'textbook' metrics fail.

Ask This: "You mentioned that Precision and Recall against WordNet were misleading for your task. Can you walk me through how you designed the 'Cluster Split Score' and why you trusted it over the standard metrics?"

PAPER 7

Jun 2024

FanOutQA: A Multi-Hop, Multi-Document Question Answering Benchmark for Large Language Models

Andrew Zhu, **Alyssa Hwang,** Liam Dugan, Chris Callison-Burch

Paper Summary: In **Enterprise Search/RAG**, models fail to answer **"fan-out" queries** (e.g., "List the CEOs of the top 5 banks and their tenures"). This paper creates a benchmark proving LLMs suffer from **context forgetting** during these multi-step lookups.

Authorship Explanation: The candidate is the second author. The paper involved a massive data collection effort recruiting 379 university students to generate complex queries. As second author, the candidate likely played a major role in managing the annotation pipeline, quality filtering (Appendix C), and executing the benchmark experiments alongside the lead author.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●●●○	4
Clarity	●●●●○	4
Relevance	●●●●○	4

Research Signature

DOMAIN	Natural Language Processing Information Retrieval Benchmarking
TECHNIQUES	Retrieval-Augmented Generation (RAG) Chain-of-Thought Decomposition Long-Context Evaluation
OBJECTIVES	Multi-Hop Reasoning Hallucination Detection Context Window Analysis

Recruiter Questions

Question 1: Verify Data Strategy

> **Context:** The candidate helped manage a huge data collection project involving hundreds of students. We need to know if they can handle messy data at scale.

Ask This: “This paper involved coordinating 379 annotators. How did you programmatically filter out low-quality or 'lazy' submissions from such a large group?”

Question 2: Assess RAG Knowledge

> **Context:** The paper shows that giving models *more* text sometimes makes them *worse* because they get distracted. This is a key problem in AI today.

Ask This: “You found that open-book performance was often worse than closed-book. Why does 'context forgetting' happen in RAG, and how would you fix it in a production system?”

PAPER 8

May 2024

Ivie: Lightweight Anchored Explanations of Just-Generated Code

Litao Yan, **Alyssa Hwang,** Zhiyuan Wu, Andrew Head

Paper Summary: In **AI-Assisted Programming**, **Blind Trust in Generated Code** limits reliability. This paper solves it by **Injecting Inline, Anchored Explanations**, achieving **Higher Comprehension and Lower Distraction than Chatbots**.

Authorship Explanation: The candidate is the second author. In HCI academic papers, this typically indicates a major contributor to the system implementation (VS Code extension) and the execution of the user study (32 participants), supporting the lead author (Yan). The last author is the faculty advisor.

Quality Scorecard

Novelty	●●●●○	4
Rigor	●●●●●	5
Clarity	●●●●●	5
Relevance	●●●●●	5

Research Signature

DOMAIN	Human-AI Interaction Developer Tools Program Comprehension
TECHNIQUES	LLM Prompt Engineering In-Situ Visualization Eye Tracking Analysis
OBJECTIVES	Reduce Cognitive Load Explainable AI (XAI) Code Verification

Recruiter Questions

Question 1: Verify UX Intuition for AI

> **Context:** Most companies just add a 'Chatbot' to their product and call it AI. This paper proves that Chatbots can be distracting. A good candidate understands when to use a Chatbot vs. when to use a subtle overlay.

Ask This: “Your paper 'Ivie' suggests that Chatbots aren't always the best way to interact with AI. If we were building an AI tool for [Company's Domain], how would you decide between a Chat interface versus the inline overlays you designed?”

PAPER 9

Jun 2019

AMPERSAND: Argument Mining for PERSuAsive oNline Discussions

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, **Alyssa Hwang**

Paper Summary: In Online Discourse, identifying how arguments support or attack each other is difficult due to data scarcity. This paper solves it by using 'distant supervision' (training on Reddit quotes and 'IMHO' tags) to adapt BERT, achieving state-of-the-art relation prediction.

Authorship Explanation: The candidate is the last author, listed after two senior Principal Investigators (Muresan and McKeown). In Computer Science, the last author is usually the senior advisor, but since the known advisors are in positions 3 and 4, the candidate is likely a supporting researcher (e.g., undergraduate or junior graduate student) who contributed to specific components like data annotation or the candidate selection module.

Quality Scorecard

Novelty	●●●●○	4
Rigor	●●●●○	4
Clarity	●●●●●	5
Relevance	●●●●○	4

Research Signature

DOMAIN	Argument Mining NLP Computational Social Science
TECHNIQUES	Transfer Learning Distant Supervision BERT Fine-tuning Rhetorical Structure Theory (RST)
OBJECTIVES	Relation Extraction Stance Detection Dialogue Analysis

Recruiter Questions

Question 1: Assess Data Intuition

> **Context:** The paper uses noisy data (Reddit comments with 'IMHO') to teach the model about arguments before showing it the clean data. This is a smart way to get more data for free.

Ask This: "This paper used 'distant supervision' from Reddit to improve performance. Can you explain the trade-offs you faced when using noisy, automatically generated labels versus high-quality human annotations?"

Question 2: Verify Collaboration/Role

> **Context:** Since the candidate is the 5th author, we need to understand what specific part of this complex system they owned.

Ask This: "You were part of a 5-person team on the AMPERSAND paper. Did you focus more on the data annotation expansion, the BERT fine-tuning pipeline, or the discourse feature integration?"

PAPER 10

Sep 2017

Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum

Christopher Hidey, Elena Musi, **Alyssa Hwang,** Smaranda Muresan, Kathleen McKeown

Paper Summary: In Argument Mining, detecting persuasion is limited by a lack of granular data. This paper creates a labeled dataset of Reddit 'Change My View' threads annotated with rhetorical strategies (Logos/Pathos), revealing that factual arguments (Logos) are statistically most effective for persuasion.

Authorship Explanation: The candidate is the third author in a five-person academic collaboration. The first author (Hidey) appears to be the lead CS researcher, while the second author (Musi) provides linguistic domain expertise. The candidate likely supported the execution of the crowdsourcing pipeline (Amazon Mechanical Turk) and statistical data analysis.

Quality Scorecard

Novelty	●●○○○	2
Rigor	●●●○○	3
Clarity	●●●●○	4
Relevance	●●●○○	3

Research Signature

DOMAIN	Argument Mining Computational Social Science NLP
TECHNIQUES	Crowdsourcing (Amazon Mechanical Turk) Statistical Correlation Analysis Annotation Schema Design
OBJECTIVES	Persuasion Detection Dataset Creation Discourse Analysis

Recruiter Questions

Question 1: Assess Data Engineering Capability

> **Context:** The candidate worked on collecting data from humans. This is messy. We want to know if they understand how to ensure high-quality data when the task is subjective.

Ask This: “This paper involved annotating subjective concepts like 'Pathos' and 'Logos.' How did you ensure the human annotators understood the task, and how did you handle disagreements between them?”

PAPER 11

Jun 2024

NewsQs: Multi-Source Question Generation for the Inquiring Mind

Alyssa Hwang, Kalpit Dixit, Miguel Ballesteros, Yassine Benajiba, Vittorio Castelli, Markus Dreyer, Mohit Bansal, Kathleen McKeown

Paper Summary: In **NLP Data Engineering**, training models to answer open-ended questions from multiple news sources is limited by **incomplete datasets**. This paper solves it by **synthetically generating questions** using a T5 model guided by keyword 'control codes' and filtering outputs with an NLI model. The outcome is **NewsQs**, a high-quality dataset of 21,000 examples.

Authorship Explanation: The candidate is the first author and an intern from the University of Pennsylvania. The work was conducted at AWS AI Labs under the supervision of the second author (Corresponding Author). This indicates the candidate led the experimentation, data generation, and writing, while the industry team provided the problem scope and resources.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●●●○	4
Clarity	●●●●○	4
Relevance	●●●●○	4

Research Signature

DOMAIN	Natural Language Processing Question Generation Multi-Document Summarization
TECHNIQUES	Synthetic Data Generation Control Codes (Prompting) T5 Fine-tuning NLI Filtering
OBJECTIVES	Dataset Creation Data Augmentation Quality Filtering

Recruiter Questions

Question 1: Verify Data Engineering Intuition

> **Context:** The candidate created a dataset by using AI to write questions for existing answers. We want to know how they ensured the AI didn't make things up.

Ask This: “You used a model to generate questions for your dataset. How did you prevent the model from hallucinating or asking irrelevant questions, and how did you validate the final quality at scale?”

Question 2: Assess Practicality of Control Codes

> **Context:** They used 'control codes' (keywords) to guide the AI. Ask why this was better than just letting the AI run freely.

Ask This: “In your NewsQs paper, you found that 'Control Codes' improved acceptability. Can you explain why unguided generation failed, and how you selected which keywords to use as controls?”

PAPER 12 Sep 2023

Kani: A Lightweight and Highly Hackable Framework for Building Language Model Applications

Andrew Zhu, Liam Dugan, **Alyssa Hwang,** Chris Callison-Burch

Paper Summary: In **LLM Application Development**, **rigid, opinionated frameworks** limit developer control and debugging. This paper solves it by **introducing Kani, a lightweight, hackable SDK**, achieving **robust function calling and transparent state management for complex agents**.

Authorship Explanation: The first two authors (Zhu and Dugan) are explicitly designated as equal contributors (joint lead). The candidate is the third author, positioned between the leads and the senior PI (Callison-Burch). This indicates a supporting researcher role, likely contributing to specific features, documentation, or evaluation of the framework, rather than being the primary architect.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●○○○	2
Clarity	●●●●●	5
Relevance	●●●●○	4

Research Signature

DOMAIN	LLM Orchestration Software Engineering Natural Language Processing
TECHNIQUES	Function Calling (Tool Use) State Management Asynchronous Programming
OBJECTIVES	Framework Design Developer Experience Reproducibility

Recruiter Questions

Question 1: Assess Engineering Pragmatism

> **Context:** This paper argues that popular tools like LangChain are too complex and 'opinionated' (they force you to work a certain way). A good engineer knows when to use a framework and when to build their own.

Ask This: “Your paper critiques 'opinionated' frameworks. Can you describe a specific instance where a framework's abstraction hid a bug from you, and how Kani's design prevents that?”

PAPER 13

Feb 2025

JumpStarter: Getting Started on Personal Goals with Adaptive Personal Context Curation

Sitong Wang, Xuanming Zhang, Jenny Ma, **Alyssa Hwang,** Zhou Yu, Lydia B. Chilton

Paper Summary: In **GenAI Productivity**, **Context Window Limits** prevent LLMs from effectively planning complex personal projects. This paper solves it by **Hierarchical Task Decomposition & Adaptive Context Selection**, achieving **Higher Quality Action Plans & Reduced Mental Load**.

Authorship Explanation: The candidate is the 4th author out of 6. The first two authors are joint leads from Columbia University. The candidate is the sole author from the University of Pennsylvania, indicating a cross-institutional collaboration. This position typically signifies a supporting role in implementation, experiment execution, or specific module development, rather than project leadership.

Quality Scorecard

Novelty	●●●○○	3
Rigor	●●●●○	4
Clarity	●●●●○	4
Relevance	●●●●○	4

Research Signature

DOMAIN	Human-Computer Interaction Generative AI Personal Productivity
TECHNIQUES	Chain-of-Thought Prompting Hierarchical Task Decomposition Context Curation Flask/Python Web Framework
OBJECTIVES	Reduce Cognitive Load Automate Planning Action Initiation

Recruiter Questions

Question 1: Verify Technical Contribution

> **Context:** The candidate was a middle author on a complex system paper. We need to know if they built the backend (Python/Flask) or just ran the user studies.

Ask This: “This paper involves both a Flask web backend and a user study. Which specific components of the JumpStarter system codebase were you responsible for implementing?”

Question 2: Assess Understanding of LLM Limitations

> **Context:** The paper deals with 'context curation' because LLMs can't remember everything perfectly or get confused by too much info. A good candidate understands why we can't just dump all data into the chat.

Ask This: “The paper mentions 'Context Dumping' vs 'Context Selection'. Why is filtering context necessary even with modern large-context LLMs like GPT-4?”

PAPER 14 Sep 2025

Attribution Gradients: Incrementally Unfolding Citations for Critical Examination of Attributed AI Answers

Hita Kambhamettu, **Alyssa Hwang,** Philippe Laban, Andrew Head

Paper Summary: In RAG systems, users struggle to verify if citations actually support AI-generated claims. This paper solves it by decomposing answers into atomic claims and linking them directly to highlighted PDF excerpts (Attribution Gradients), resulting in higher-quality user verification and deeper source engagement.

Authorship Explanation: The candidate is the second author in a four-person collaboration involving UPenn and Microsoft Research. The first author (Kambhamettu) likely led the writing and primary execution. As second author, the candidate likely played a major role in the system implementation (React/Python pipeline) or the execution of the 20-person user study.

Quality Scorecard

Novelty	●●●●○	4
Rigor	●●●●○	4
Clarity	●●●●●	5
Relevance	●●●●●	5

Research Signature

DOMAIN	Human-Computer Interaction (HCI) Retrieval-Augmented Generation (RAG) Natural Language Processing
TECHNIQUES	Attribution Gradients Citation Analysis PDF Parsing (PaperMage) React/Frontend Engineering
OBJECTIVES	Hallucination Mitigation Source Verification Sensemaking

Recruiter Questions

Question 1: Assess Full-Stack/Prototype Capability

> **Context:** The candidate helped build a system that connects AI text to PDF highlights. We need to know if they can build end-to-end prototypes.

Ask This: “This paper describes a complex system involving PDF parsing, LLM claim extraction, and a React frontend. Which parts of this pipeline did you personally implement?”

Question 2: Verify Understanding of RAG Limitations

> **Context:** RAG systems often cite the wrong papers. This candidate worked on fixing that. A good answer acknowledges that citations are often 'hallucinated' or irrelevant.

Ask This: “Your study mentions that citations often fail to support claims. In your view, is this primarily a retrieval failure or a generation failure, and how does your UI mitigate that?”

PAPER 15

Jun 2021

QuakerBot: A Household Dialog System Powered by Large Language Models

Artemis Panagopoulou, Manni Arora, ... **Alyssa Hwang,** Chris Callison-Burch, Mark Yatskar

Paper Summary: In Task-Oriented Dialog, purely neural models lack reliability and safety. This paper solves it by engineering a hybrid architecture that routes requests between Large Language Models (for flexibility) and rule-based state managers (for control), achieving a production-ready bot for Alexa users.

Authorship Explanation: The candidate is the 10th author on a large 14-person team participating in the Amazon Alexa Prize TaskBot Challenge. This positioning typically indicates a role as a core contributor responsible for a specific module (e.g., data annotation pipelines, specific responders, or testing frameworks) rather than the lead architect or principal investigator.

Quality Scorecard

Novelty	●●○○○	2
Rigor	●●●○○	3
Clarity	●●●●○	4
Relevance	●●●●○	4

Research Signature

DOMAIN	Conversational AI Task-Oriented Dialog Safety Engineering
TECHNIQUES	Few-Shot Prompting RAG (Retrieval Augmented Generation) Hybrid Neuro-Symbolic Architecture Intent Detection
OBJECTIVES	Household Task Assistance Harm Mitigation System Robustness

Recruiter Questions

Question 1: Verify Specific Contribution

> **Context:** This was a massive team effort. We need to know exactly what piece of the robot this candidate built. Did they build the brain, the safety filter, or just label data?

Ask This: "In the QuakerBot project, there were 14 authors. Which specific module (e.g., Harm Classifier, Retrieval, State Manager) did you own end-to-end, and what was the hardest engineering constraint you faced?"

Question 2: Assess Production Reality

> **Context:** This bot actually talked to real people on Alexa. Ask about what happened when real users tried to break it.

Ask This: "Since this bot was deployed to real Alexa users, can you share an example of a 'failure mode' where the LLM did something unexpected, and how you patched that behavior?"