POTATO 2.0: A Comprehensive Annotation Platform with Support for AI-in-the-Loop and Agentic Systems

David Jurgens
University of Michigan
jurgens@umich.edu

Michael Chen
University of Michigan
mecow@umich.edu

Lina Iyer
University of Michigan
linaiyer@umich.edu

Abstract

Annotated data remains essential for training and evaluating NLP systems. Large language models have broadened the kinds of data researchers need, including multimodal and agentic system data. Here, we introduce POTATO 2.0, a major update to our open source annotation platform designed for easy deployment, customization, and fully reproducible and shareable annotation designs. POTATO offers broad support for many types of annotations in NLP, including 39 different types of annotation tasks, support for text, audio, image, and video modalities, or mixtures thereof. POTATO 2.0 includes robust support for labeling agentic system outputs through reading common trace formats, or live interaction and annotation with agents in multiple settings, such as chatting, web-browsing, and coding. POTATO also includes multiple AI-assistance features to help annotators more easily label data. Finally, POTATO introduces a new agentic AI-in-the-loop workflow where a single human annotator collaborates with an LLM through iterative prompt refinement, uncertainty-driven instance selection, and progressive autonomy—enabling efficient dataset creation without a large annotation team.

1 Introduction

Annotated data is still central to NLP and Generative AI, yet the demands on annotation have grown in both scale and complexity (Halevy et al., 2009; Monarch, 2021). Beyond traditional text classification, researchers now require annotations for multimodal content (Russell et al., 2008; Kirillov et al., 2023), LLM evaluation via human preference judgments (Ouyang et al., 2022; Rafailov et al., 2023; Chiang et al., 2024), agentic tasks (Yehudai et al., 2025; Mohammadi et al., 2025), and subjective tasks where annotator disagreement is itself informative (Plank, 2022; Aroyo and Welty, 2015). While multiple tools have emerged to help researchers create this data, the annotation design itself remains difficult to create, reproduce, or rapidly iterate upon.

We introduce POTATO 2.0 built as an open-source annotation platform to support easy deployment and customization, while also making all annotation designs sharable and reusable. Since its initial release (Pei et al., 2022), annotation practice has changed substantially. LLMs now rival or exceed crowd workers on many annotation tasks (Gilardi et al., 2023; Ding et al., 2023; He et al., 2024a), and human–LLM collaborative and LLM-assisted workflows have become an active research area (Li et al., 2023; He et al., 2024b; Wang et al., 2024). Surveys report both the promise and the risks of LLM-generated labels and stress that per-task validation is still essential (Pangakis et al., 2023; Tan et al., 2024). Concurrently, the rise of data-centric AI (Zha et al., 2025) and the increasing complexity of multimodal NLP demand tools that support annotation across images, audio, and video alongside text. Researchers also need reliable ways to combine human and machine judgments, calibrate LLM confidence (Geng et al., 2024; Gligoric et al., 2025), and handle the subjectivity inherent in many labeling tasks (Uma et al., 2021; Hovy and Prabhumoye, 2021).

POTATO addresses needs through three major advances: (1) **Expanded annotation capabilities**: 39 annotation types (e.g., rating, comparison, dependencies, typed events, spans, bounding regions, code review, step-level process rewards, and agent-trajectory ratings) and support for multiple modalities including text, image, audio, video, agentic interactions, agentic coding, and dialogue modalities with an extensible registry-based architecture. (2) **Deep AI/LLM integration**: Multiple AI integrations allow administrators to enable specific types of assistance, such as intelligent hints, label suggestions, AI rationales, and option highlighting, with support for local and commercial model

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 374–386

providers.(3) **Agentic Annotation**: A novel multi-phase LLM-in-the-loop workflow enabling a single annotator to collaboratively label datasets with LLMs through prompt refinement, uncertainty estimation, and progressive autonomy.

POTATO is fully open-source and free to use in academic and commercial settings, licensed under the GNU GPL v3 license, and is deployable with minimal dependencies. Code and documentation are at https://github.com/davidjurgens/potato. POTATO uses a simple YAML template to configure its entire annotation scheme, which allows easy reproducibility; we have also released an ever-growing showcase of 360+ annotation designs from academic papers[1] covering multiple modalities and annotation types, which allows practitioners to easily label their own data using best practices. The software is designed to be extended and can serve as a testbed for research on AI integration and human–AI collaboration for data labeling (Kim et al., 2024; Zhang et al., 2023; Xiao et al., 2023; Qin et al., 2025; Xiong et al., 2025).

2 Architecture and Design

POTATO uses a Flask-based server architecture to support web-based annotation. Following, we detail the core architectural elements in the software.

2.1 Annotation Schema and Data Types

POTATO supports 39 annotation types, from the common radio buttons, checkboxes, and text spans to the more advanced types like best-worst scaling, event and role annotation, dependency parsing, code review with inline diff comments, step-level process-reward judgments, and full-trajectory agent ratings. POTATO uses integrated audio, image, and video displays to support other modality-specific annotation: image annotation (bounding boxes, polygons, freeform drawing, landmarks), audio annotation (waveform-based temporal segmentation), and video annotation (temporal segments, frame classification, keyframes). Custom displays are also provided for annotating specific types of text: html, markdown, dialog, or pairwise comparisons. Task designers can specify multiple annotation schema and input sources for a task, each of which can be combined with any compatible annotation schema. For example, a designer can use span annotation on dialogue transcripts or radio-button classification on images without custom code.

Administrators specify an annotation task using a YAML file that contains the data sources, the annotation schema defining what judgments annotators will make (e.g., radio buttons, checkboxes, sliders, spans), and optional details on how each instance should be displayed. The configuration also controls the annotator workflow, from optional consent forms, instructions, and training phases with gold-standard feedback through to post-study surveys drawn from a library of 55 validated instruments. Conditional display logic allows schema to appear or hide dynamically based on prior answers, enabling complex branching annotation flows without any code. The same file governs quality control settings such as adjudication queues, and custom task layouts that can fully restyle the annotation interface. POTATO has default settings for most options so administrators can design a task with only the minimal changes needed. For non-technical users, the https://www.potatoannotator.com/playground website provides an interactive GUI to design tasks and produce the task's YAML file.

2.2 Annotation Workflow

Annotators progress through up to eight configurable phases: login, consent, pre-study survey, instructions, training, annotation, post-study survey, and completion. This progression follows established annotation methodology (Hovy and Lavid, 2010; Artstein and Poesio, 2008): each phase can span multiple pages, and deployers can enable or skip phases via YAML configuration. The training phase supports practice items with feedback, preparing annotators before the main task, and can remove annotators from the study should they fail to correctly annotate data with known labels. The pre-study and post-study phases allow practitioners to collect more data about who their annotators are, recognizing that an individual's background may influence their annotation behavior (e.g., Davani et al., 2024); to support this practice, we have included 55 common questionnaires for a variety of constructs such as demographics, personality (e.g., Big-5; John and Srivastava, 1999), well-being, and attitudes.

During annotation, POTATO supports multiple options for which items annotators see next: random, least-annotated first, fixed order, active learning based, LLM confidence-based, and diversity-based options that use semantic embeddings to clus-

[1] https://github.com/davidjurgens/potato-showcase/

ter texts. Administrators can configure these latter three strategies, such as by specifying the model to use for active learning (from scikit-learn) and how often active learning is run.

2.3 Data Infrastructure and Crowdsourcing

POTATO supports 8 data source types: local files (JSON, CSV, TSV, JSONL), web-hosted files, Amazon S3, Google Drive, Dropbox, HuggingFace Datasets (e.g., Parquet), Google Sheets, and SQL databases. Tasks can read from these sources as data, e.g., annotating data stored on Dropbox, and then write the eventual output to a Parquet file for uploading to Huggingface. A directory watcher enables live ingestion of new files during annotation.

As a browser-based annotation platform, POTATO supports both local and crowd-based annotation. The library provides full lifecycle integration with Amazon Mechanical Turk and Prolific for crowdsourced annotation campaigns. However, not all data can be shared publicly, or practitioners may want to annotate with a custom group of individuals—or even test out the annotation task themselves. POTATO can be launched on any OS supported by python. The POTATO website contains extensive documentation for how to host the platform.

2.4 Quality Control and Adjudication

Especially in crowdsourcing settings, quality control is needed to address potential noise and disagreements. POTATO includes multiple features to help practitioners. When gold-standard data is available, POTATO can automatically insert attention checks at configurable intervals to verify annotator engagement, with an optional customizable warnings to annotators and blocking thresholds. POTATO includes fine-grained behavioral tracking during annotation, logging interface events (clicks, keypresses, focus changes, navigation) and per-instance response times; these statistics make it easier to spot annotators who speed through data or type content unnaturally quickly (e.g., copying an LLM response when copy/paste is disabled).

POTATO features an administrator dashboard that shows current progress and statistics on annotators, including each annotator's accuracy on gold-standard data, agreement with other annotators, and relative speed. Administrators can also see statistics on the data itself, identifying contentious items to review more carefully.

When disagreements do happen, POTATO provides a dedicated interface for resolving inter-annotator disagreements. The interface shows all annotators' judgments, as well as behavioral information (e.g., annotation time), and allows adjudicators to write optional notes about the decision and label the disagreement with a taxonomy for tracking causes. For some data types (e.g., rating scales), administrators can optionally also run MACE (Hovy et al., 2013) for a Bayesian estimation of annotator reliability and the true labels.

2.5 AI and LLM Integration

POTATO supports four modes of assistance during annotation that are aimed at supporting, rather than replacing, the human in annotation. **Intelligent hints**: Keyword highlighting based on LLM analysis, drawing annotator attention to relevant portions of the text, without suggesting which label they choose. For images, a configurable VLLM suggests bounding boxes they examine. **Label suggestions**: Pre-filled annotation predictions that annotators can accept, modify, or reject; the current design prompts the LLM to select the two most likely labels and thus still requires a human to pick, but narrows their attention to more probable labels, which is important in tasks with many possible labels. **AI rationales**: Per-label explanations generated by vision or text models and grounded in the annotation codebook, helping annotators understand why a particular label may apply. This assistance is intended for annotators working with complex, multi-page annotation instructions. **Option highlighting**: Confidence-weighted visual cues that visually emphasize the labels an LLM scores most likely, without prefilling them. Unlike label suggestions, highlighting preserves the full choice set, which is helpful when the annotator wants to scan quickly without losing the ability to inspect any label.

POTATO integrates with 12 LLM/VLLM endpoint types: OpenAI (text and vision), Anthropic (text and vision), Google Gemini, Ollama (text and vision), OpenRouter, HuggingFace, vLLM, YOLO for object detection, and any other networked model that exposes an OpenAI-compatible API. AI configurations can be defined in external files and shared across projects. An optional in-context learning (ICL) module constructs few-shot prompts from existing annotations, and an LLM-based active-learning module uses model confidence to prioritize uncertain instances. A shared

caching layer reduces redundant API calls across sessions and lowers cost.

3 Major Innovations

POTATO introduces two major features aimed at emerging annotation needs: (i) annotation for agentic systems and data and (ii) AI-in-the-loop annotation.

3.1 Agentic Annotation

As LLM-based agents are increasingly deployed for complex multi-step tasks—e.g., web browsing, code generation, tool use, and conversational assistance—researchers need tools for evaluating agent behavior at multiple levels of granularity (Yehudai et al., 2025; Mohammadi et al., 2025). POTATO introduces dedicated support for annotating these agentic traces through three complementary modes: *post-hoc trace evaluation*, *interactive agent testing*, and *coding agent annotation*.

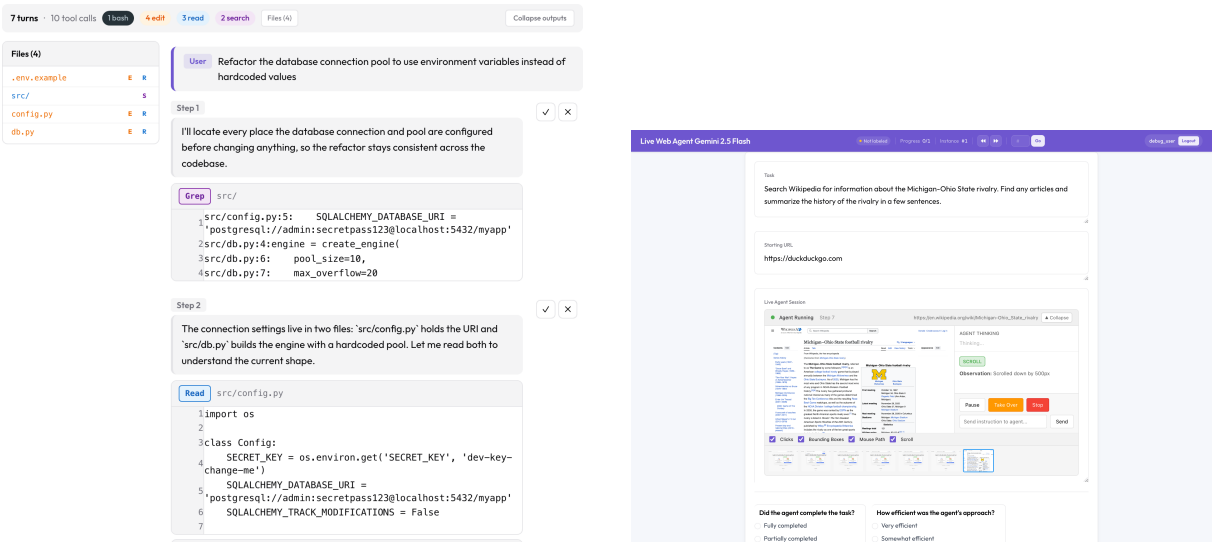
Agent Trace Evaluation Modern agentic systems frequently use a ReAct like framework (Yao et al., 2022) where a reasoning LLM uses an internal multistep think-action-observation sequence, often based on tool use. These systems often record the full sequence, so evaluators can assess overall task success as well as the quality and correctness of each individual step. POTATO can directly import trace data from agentic libraries and has a special display to render recorded agent trajectories as color-coded step cards, with each step classified as a *thought*, *action*, *observation*, or *system* event. Step types are either specified explicitly in the data or inferred from speaker names via pattern matching. The display includes a summary header, optional inline screenshots for GUI-based agents (e.g., web agents), and collapsible observation blocks for long outputs. POTATO accepts three input formats: speaker–text pairs, structured thought/action/observation dicts, and explicit step-type/content records, so traces from many agent frameworks can be ingested without preprocessing. POTATO ships with 15 trace converters covering the major agent ecosystems: `langchain` (LangSmith), `langfuse`, `react` (generic ReAct JSON), `webarena` (Zhou et al., 2024), `atif` (academic interchange format), OpenAI Agents SDK and Anthropic tool-use traces, OpenTelemetry (OTEL) spans, Model Context Protocol (MCP) sessions, multi-agent traces, and three coding-agent formats (`swe-bench`, `swe-agent`, `aider`) together with Claude Code session logs.

Combined with POTATO's existing annotation schemas, this enables multi-level evaluation: span-level annotations for marking hallucinations or incorrect facts within individual steps, per-turn ratings for assessing each agent action, and trajectory-level judgments for overall task success, safety, and efficiency. We provide example configurations for evaluating ReAct-style traces, RAG pipelines with citation faithfulness, visual GUI agents with screenshots, pairwise A/B agent comparison, and coding-agent trajectories with diff-and-test inspection.

Live Interactive Agent Testing POTATO also allows live interaction with some agents, so annotators can label behavior as it happens, probe specific scenarios, and stop early when an agent fails rather than running it to completion. In the first phase, annotators interact with a live agent according to that agent's design, e.g., sending messages, observing agent responses, and completing the task. A safety sandbox enforces configurable limits on maximum turns, session duration, and per-user rate limits. The interactive setting also supports multimodal agentic behaviors such as web-based browsing (Figure 1b), where users can instruct models and observe their progress, and even intervene and redirect if necessary. In the second phase, after the annotator finishes the conversation, the completed dialogue is persisted and rendered as a trace, with annotation forms enabled for evaluation. This sequencing ensures annotators experience the full interaction before rating.

The agent proxy architecture supports multiple backends: an OpenAI-compatible endpoint for hosted models, a generic HTTP proxy for custom agent APIs with configurable field mapping, and an echo proxy for testing. The session manager provides thread-safe multi-user isolation keyed by annotator and instance, with transparent session recovery across page refreshes. All agent configuration options are specified in the YAML.

Coding Agent Annotation Coding agents such as Claude Code and OpenCode are a fast-growing class of agentic systems whose trajectories include thoughts and tool calls as well as file edits, diffs, test runs, and shell state. POTATO adds dedicated support for evaluating them through three live backends—Claude Code, SWE-Agent, and Aider—each driven by a sandboxed workspace with check-



(a) Coding agent annotation

(b) Web-browsing agent annotation

Figure 1: Two examples of the agentic annotation interfaces in POTATO: (left) Annotating a live coding agent where file diffs from the sandboxed workspace appear alongside a process-reward form (here, shown as the checkbox on the right side) that captures step-level quality judgments coding-agent evaluation with diffs and step-level process reward, and (right) Annotating a web-browsing agent: each step's rendered screenshot is overlaid with SVG cues for clicks, hovers, and bounding boxes; an annotation form is attached to the active step in a filmstrip review. These complement the post-hoc trace and live-chat interfaces shown in Figure 2.

point and rollback (Figure 1a). Annotators can step through a recorded trajectory, branch the session at any step to explore counterfactual edits, or run a new task end-to-end and label the resulting trace. Two specialized schemas complement the interface: *code review*, which displays diffs with line-anchored inline comments in a GitHub-style layout, and *process reward*, which collects step-level quality judgments for use in process-reward modeling. We provide example configurations for evaluating Aider patches, SWE-bench submissions, and Claude Code sessions.

3.2 AI-in-the-Loop Annotation

Annotation generally requires multiple human annotators to achieve reliable labels (Artstein and Poesio, 2008; Paun et al., 2022), but recruiting, training, and paying annotators is expensive and time-consuming (Snow et al., 2008; Shmueli et al., 2021), especially for researchers creating an initial dataset. LLMs can generate labels at scale and in some cases outperform crowd workers (Gilardi et al., 2023; Ding et al., 2023), but their predictions are unreliable on ambiguous instances, susceptible to systematic biases (Hovy and Prabhumoye, 2021; Ziems et al., 2024), and require per-task validation (Pangakis et al., 2023; Tan et al., 2024). POTATO introduces a new agentic annotation model with AI-in-the-loop that aims to (1) continuously revise the annotation instructions for humans and LLMs to enable accurate, reproducible annotation behavior, and (2) select items to annotate that are most likely to inform the decision boundary for the guidelines. Recent work has explored uncertainty-guided allocation between humans and LLMs (Li et al., 2023; Gligoric et al., 2025), active learning with LLM annotators (Zhang et al., 2023; Xiao et al., 2023), and noise-robust training on LLM-generated labels (Yuan et al., 2024). We combine these threads into a single workflow: one human annotator works with an LLM through iterative prompt refinement, uncertainty-driven instance routing, and a gradual handoff to LLM autonomy under continued human oversight.

Workflow The agentic workflow begins with the annotator providing an initial task description, which the LLM synthesizes into instructions for a human to review and edit. Once approved, the LLM generates synthetic examples (informed by the real data) that are likely boundary cases; the human labels these to inform how the instructions should be refined by the LLM and then by the human again, optionally iterating on synthetic examples until satisfied. Once this initial phase completes, the human and LLM begin labeling real data in parallel. The human sees a subset of this data

sampled by multiple strategies: a random sample and one or more LLM-uncertainty based strategies. These latter strategies aim to identify cases where the LLM's decision boundary is not clear. POTATO includes multiple uncertainty-estimating methods, such as prompting the model for its own confidence score, using the log-probabilities of the predicted label (if available), and repeatedly sampling the label from the model to measure answer diversity; this framework is also extensible, allowing us to include new uncertainty methods in this active research area. When human-LLM disagreements are detected, the annotator reviews conflicts, and the system optionally triggers a prompt revision to improve guidance. Ultimately, when satisfied, the human can approve the prompt and label remaining data. The instructions can also be given to other human annotators to perform labeling to calculate inter-annotator agreement and consensus with the LLM's judgments. POTATO's workflow lets a single annotator refine the instructions iteratively until they are clear to both people and models.

4 Comparison with Existing Systems

Established Platforms The annotation tool ecosystem spans several categories that largely focus on a specific modality. INCEpTION (Klie et al., 2018; Eckart De Castilho et al., 2024) and BRAT (Stenetorp et al., 2012) support rich linguistic annotation, including relations and coreference; INCEpTION has recently added experimental LLM integration (Ollama, ChatGPT, Azure OpenAI) and remains the strongest tool for knowledge-base linking. Label Studio (Tkachenko et al., 2021) provides the broadest single-tool coverage of modalities with ML backend support available in both its open-source and enterprise editions, including LLM endpoints. Prodigy (Explosion, 2017) pioneered active learning in annotation and has added LLM support via spacy-llm, relation annotation, audio/video segmentation, IAA metrics, and a review/adjudication recipe. doccano (Nakayama et al., 2018) offers lightweight open-source text annotation including relation labeling. For images, CVAT (CVAT.ai Corporation, 2022) provides robust computer vision annotation with AI-assisted labeling (SAM, YOLO, HuggingFace model integration) but does not support text annotation. Foundational image annotation tools like LabelMe (Russell et al., 2008) and advances in segmentation models (Kirillov et al., 2023) have shaped the image annotation landscape now served by CVAT and Label Studio. Finally, ELAN (Wittenburg et al., 2006; Brugman and Russel, 2004) excels at time-aligned audio/video annotation for linguistics research.

Recent Systems A growing body of annotation tools has appeared at recent NLP venues. MEGAnno+ (Kim et al., 2024) is perhaps the most directly comparable to POTATO, offering a Jupyter-based system where LLM agents label data first and humans verify uncertain instances. Thresh (Heineman et al., 2023) shares POTATO's YAML-driven philosophy but focuses specifically on fine-grained text evaluation with a community hub for sharing frameworks. GATE Teamware 2 (Wilby et al., 2023) provides JSON-configurable document classification annotation with annotator training and quality screening. ALANNO (Jukić et al., 2023) focuses on active learning for annotation, and CodeAnno (Schneider et al., 2023) extends WebAnno (Yimam et al., 2013) for social science coding tasks.

Several recent tools address AI-integrated annotation directly. ITAKE (Song et al., 2024) combines LLM annotation with online machine learning for interactive knowledge extraction. DocSpiral (Sun et al., 2025) introduces a "human-in-the-spiral" approach to document annotation where iterative cycles progressively reduce human effort, reporting 41% annotation time reduction. CrowdAgent (Qin et al., 2025) orchestrates LLMs, small language models, and human experts in a multi-agent system for multimodal classification. Co-DETECT (Xiong et al., 2025) uses mixed-initiative human–LLM annotation for collaborative edge case discovery. For LLM evaluation specifically, ChatHF (Li et al., 2024) and BotEval (Cho et al., 2024) provide interactive annotation interfaces for chatbot evaluation and RLHF data collection. Fabricator (Golde et al., 2023) takes a different approach entirely, using teacher LLMs to generate labeled training data without human annotation. Argilla (Argilla, Inc., 2024) has gained significant traction as an open-source platform specifically designed for LLM alignment, preference annotation, and integration with the HuggingFace ecosystem. Specialized annotation tools have also been proposed for specific tasks: EventFull (Eirew et al., 2025) for temporal and causal relation annotation, FirstAID (Menini et al., 2025) for knowledge-driven dialogue data collection, and Commentator (Sheth et al., 2024) for code-mixed multilingual text.

Distinguishing Features of POTATO Against this

		Label Studio	CVAT	Prodigy	INCEpTION	BRAT	ELAN	doccano	POTATO 2.0
Annotation Types	Text classification (radio/checkbox)	✓		✓	✓			✓	✓
	Span annotation	✓		✓	✓	✓		✓	✓
	Relation/link annotation	✓		✓	✓	✓		✓	✓
	Likert / rating scales								✓
	Free text / textbox	✓		✓				✓	✓
	Image bbox / polygon	✓	✓	✓					✓
	Image segmentation masks	✓	✓						✓
	Audio segmentation	✓		✓			✓		✓
	Video temporal annotation	✓	✓	✓			✓		✓
	Agent workflow annotation								✓
	Coding agent annotation								✓
	Multirate matrices								✓
AI / LLM Integration	ML-assisted labeling	✓	✓	✓	✓				✓
	LLM endpoint support	✓		✓	✓†				✓
	Multiple LLM providers	✓		✓					✓
	AI rationales / explanations								✓
	Human–LLM collaboration								✓
Quality Control	Inter-annotator agreement	✓*		✓	✓				✓
	Adjudication interface	✓*		✓	✓				✓
	MACE competence estimation								✓
	Attention checks								✓
	Gold standard items	✓*							✓
	Behavioral tracking								✓
Research Workflow	Multi-phase progression								✓
	Pre/post-study surveys								✓
	Validated survey instruments (55)								✓
	Training phase with feedback								✓
Infra-structure	Active learning	✓*		✓	✓				✓
	Keyboard shortcuts	✓	✓	✓	✓	✓	✓	✓	✓
	Multiple data sources (>3)	✓							✓
	Config-driven setup	✓		✓					✓
	Open-source	✓	✓		✓	✓	✓	✓	✓
	Free	✓	✓		✓	✓	✓	✓	✓

Table 1: Feature comparison between POTATO 2.0 and widely used annotation platforms. *Enterprise/paid tier only. †Experimental. Blank cells indicate the feature is not available or not a core capability of the tool.

landscape, POTATO distinguishes itself in multiple ways, shown in Table 1, which we summarize across three key dimensions. First, *annotation breadth*: while tools like Label Studio and Prodigy have expanded their modality coverage, POTATO provides 15 annotation schemas across text, image, audio, and video unified through registry-based architecture, all configurable via YAML without code. Second, *AI integration depth*: POTATO supports 12 LLM endpoint types spanning major commercial and open-source providers, substantially more than the 1-3 providers available in other tools. Beyond label suggestions, POTATO uniquely offers AI rationales (per-label explanations) and option highlighting (confidence-weighted visual cues). Building on recent work in human–LLM collaboration (Kim et al., 2024; Qin et al., 2025; Xiong et al., 2025), POTATO adds the multi-step AI-in-the-loop workflow described in §3.2. Third, *research workflow support*: features like multi-phase progression (consent through post-study), MACE integration (Hovy et al., 2013), 55 validated survey instruments, behavioral tracking, and crowdsourcing integration (MTurk, Prolific) are absent from or only partially supported by other tools. This infrastructure supports research on annotation methodology itself: studies of annotator demographics (Pei and Jurgens, 2023; Sap et al., 2022), inter-annotator disagreement (Uma et al., 2021; Plank, 2022), and human–LLM collaboration can all build on POTATO's tracking and survey features.

5 Conclusion

POTATO 2.0 is a new annotation platform with 39 annotation types across text, image, audio, and video, support for labeling agentic interactions, deep AI/LLM integration, and human–LLM collaborative annotation (see examples in Appendix Figure 2). The software supports the full research-annotation life cycle (consent, training, labeling, quality control, and export) in a single YAML-configured, open-source tool. These YAML files allow easy and full replication of any annotation design, and we have released over 360 such designs from research papers to help practitioners get started. POTATO is freely available and under active development.

Ethics

As with any configurable annotation tool, POTATO's ethical implications depend substantially on how it is deployed. We discuss key considerations specific to the release.

AI Assistance and Bias POTATO includes multiple AI integrations that can potentially influence an annotator's judgments. These integrations create a risk of propagating LLM biases into annotated datasets; this risk is greatest for the AI-in-the-loop annotation once it begins labeling data independently from the task designer. We mitigate these through several design choices. First, for AI assistance, our designs do not directly tell annotators which specific label or output to choose. Instead, they provide additional information or a more narrow focus on fewer options to help direct an annotator's attention. As a result, annotators are still able to choose whichever option they prefer and cannot blindly accept an AI-suggested label. These suggestions or rationales may still be incorrect, so careful review is needed; POTATO supports this through adjudication.

For the AI-in-the-loop, our approach includes multiple places of human review: (i) reviewing prompt instructions as an interactive process to inspect mistakes, (ii) periodic spot checks on AI-labeled data to assess correctness, and (iii) a human-AI disagreement detection can trigger a return to human annotation. The system design and documentation still encourage a final step of soliciting independent human annotators to label data to assess the instruction clarity and baseline level of agreement. Nevertheless, deployers should be aware that LLM-generated labels may reflect biases present in the model's training data (Blodgett et al., 2020).

Fair Compensation and Annotator Welfare POTATO includes behavioral tracking and per-instance timing to support identifying potentially-adversarial users who speed through data labeling without reviewing the item (or who use LLMs to quickly label). In such cases, an administrator may remove these annotators from the study and potentially reject their work, leaving them without payment. However, there is a risk that some annotators may be faster than others and appear adversarial, when they are simply more proficient at that task. POTATO does not specify any direct guidance for what might be adversarial and instead requests that task deployers look at the distribution of times (or behaviors) directly to identify outliers.

POTATO includes 55 integrated survey instruments enable researchers to study the relationship between annotator backgrounds and labeling behavior (Sap et al., 2022). We encourage deployers to use these tools to promote equitable representation of different views in annotation and for post-hoc analysis of which backgrounds might not be sufficiently represented. While none of the included survey instruments collect personally identifiable information, these responses still need to be treated responsibly.

Acknowledgments

The POTATO team thanks all the folks that have contributed code and submitted issues, especially Aldo Costa whose issues and fixes made POTATO 2.0 significantly more robust. This work was supported in part by the National Science Foundation under Grant No. IIS-2143529 and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Argilla, Inc. 2024. Argilla: Open-source data curation platform for LLMs. https://github.com/argilla-io/argilla. Open-source tool for LLM alignment annotation.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Hennie Brugman and Albert Russel. 2004. Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I Jordan, Joseph E Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*.

Hyundong Cho, Thamme Gowda, Yuyang Huang, Zixun Lu, Tianli Tong, and Jonathan May. 2024. BotEval: Facilitating interactive human evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

CVAT.ai Corporation. 2022. CVAT: Computer vision annotation tool. https://github.com/cvat-ai/cvat. Open source software.

Aida Davani, Mark Diaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11173–11195.

Richard Eckart De Castilho, Jan-Christoph Klie, and Iryna Gurevych. 2024. Integrating INCEpTION into larger annotation processes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Alon Eirew, Eviatar Nachshoni, Aviv Slobodkin, and Ido Dagan. 2025. Eventfull: complete and consistent event relation annotation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 494–508.

Explosion. 2017. Prodigy. https://prodi.gy.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the 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.

Kristina Gligoric, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can unconfident LLM annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3514–3533.

Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. Fabricator: An open source toolkit for generating labeled training data with teacher LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–11.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024a. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Industry Track*, pages 165–190.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024b. If in a crowdsourced data annotation pipeline, a GPT-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

David Heineman, Yao Dou, and Wei Xu. 2023. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 339–352.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. 22:13–36.

Oliver P John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A Pervin and Oliver P John, editors, *Handbook of personality: Theory and research*, volume 2, pages 102–138. Guilford Press, New York, NY.

Josip Jukić, Fran Jelenić, Miroslav Bićanić, and Jan Šnajder. 2023. Alanno: An active learning annotation system for mortals. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 228–235.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo,

Piotr Dollar, and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Andrew Li, Zhenduo Wang, Ethan Mendes, Duong Minh Le, Wei Xu, and Alan Ritter. 2024. ChatHF: Collecting rich human feedback from real-time conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 270–279.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. pages 1487–1505.

Stefano Menini, Daniel Russo, Alessio Palmero Aprosio, and Marco Guerini. 2025. First-aid: the first annotation interface for grounded dialogues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 563–571.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6129–6139.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. *arXiv preprint arXiv:2306.00176*.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Maosheng Qin, Renyu Zhu, Mingxuan Xia, Zhen Zhu, Minmin Lin, Junbo Zhao, Lu Xu, Changjie Fan, Runze Wu, Haobo Wang, et al. 2025. Crowdagent: Multi-agent managed multi-source annotation system. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 925–942.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*.

Florian Schneider, Seid Muhie Yimam, Fynn Petersen-Frey, Gerret Von Nordheim, Chris Biemann, et al. 2023. Codeanno: Extending webanno with hierarchical document level annotation and automation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 11–17.

Rajvee Sheth, Shubh Nisar, Heenaben Prajapati, Himanshu Beniwal, and Mayank Singh. 2024. Commentator: A code-mixed multilingual text annotation framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of nlp crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Jiahe Song, Hongxin Ding, Zhiyuan Wang, Yongxin Xu, Yasha Wang, and Junfeng Zhao. 2024. ITAKE: Interactive unstructured text annotation and knowledge extraction system with LLMs and ModelOps. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 326–334.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Qiang Sun, Sirui Li, Tingting Bi, Du Q Huynh, Mark Reynolds, Yuanyi Luo, and Wei Liu. 2025. Docspiral: A platform for integrated assistive document annotation through human-in-the-spiral. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 267–274.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Maxim Tkachenko, Mikhail Malyuk, Nikita Shevchenko, Andrey Holmanyuk, and Nikolai Liubimov. 2021. Label studio: Data labeling software, 2020–present. Open source software available from https://github.com/HumanSignal/label-studio.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–21.

David Wilby, Twin Karmakharm, Ian Roberts, Xingyi Song, and Kalina Bontcheva. 2023. GATE teamware 2: An open-source tool for collaborative document classification annotation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 145–151.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535.

Chenfei Xiong, Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Lorena Calvo-Bartolomé, Alexander Hoyle, Zhijing Jin, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, Mennatallah El-Assady, and Elliott Ash. 2025. Co-DETECT: Collaborative discovery of edge cases in text classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 354–364, Suzhou, China. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

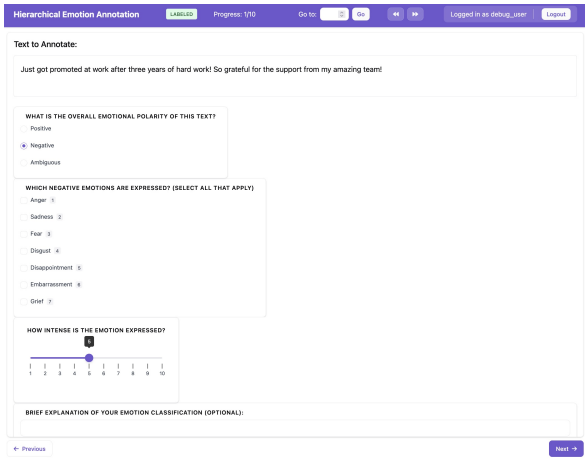
Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. Hide and seek in noise labels: Noise-robust collaborative active learning with LLMs-powered assistance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10977–11011.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*.

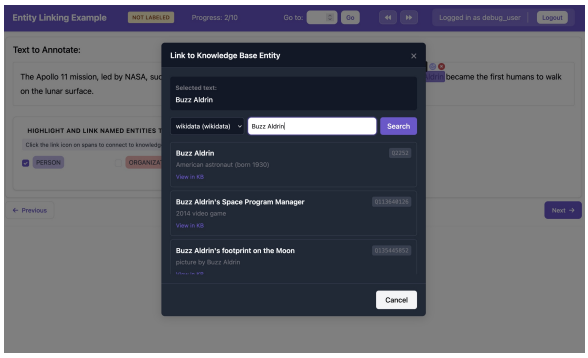
Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. pages 13088–13103.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.

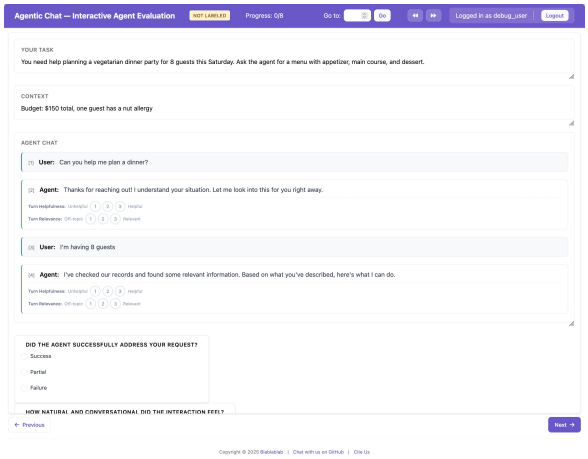
Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.



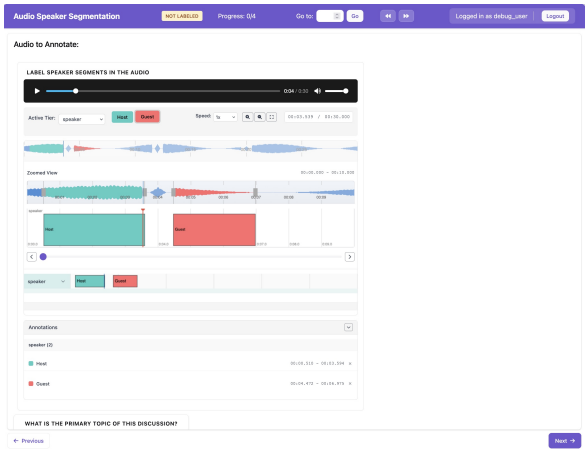
(a) A custom task featuring multiple annotation types for text input.



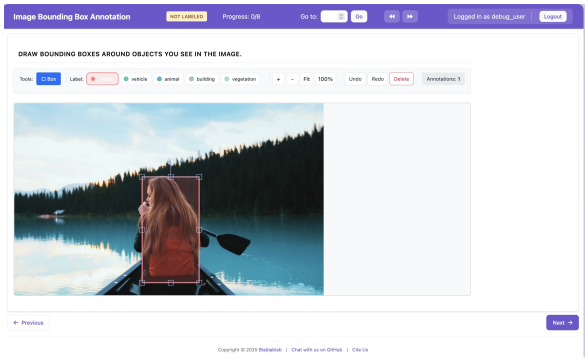
(b) An entity linking task where entities are directly linked to WikiData using this lookup functionality.



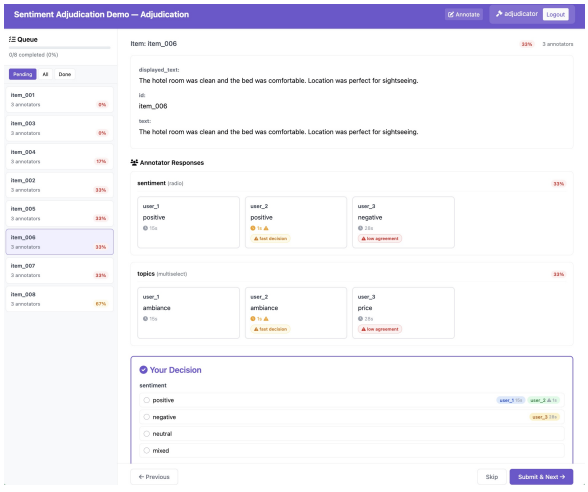
(c) The annotation interface for rating individual conversation turns with an agentic system after an annotator has ended the conversation.



(d) A sequence labeling guest for audio for identifying hosts and guests in an audio file.



(e) An object category bounding-box task for images.



(f) The adjudication interface for a sentiment analysis task.

Figure 2: Examples of POTATO tasks across different designs and modalities (a)–(e) and the adjudication interface for a sentiment analysis task.