Towards Conversational Data Annotation: Personalized Annotation Explanation Generation via Large Language Models

Literature Review, Datasets, Thesis Proposal

Mark Nagengast Porro

Contents I

Human-LLM collaboration for annotation

- CoAnnotating: Uncertainty-Guided Work Allocation between Human and LLMs for Data Annotation (Li et al., 2023)
  - Resource allocation, efficient collaboration
  - Estimation of LLM's annotation capability using uncertainty metrics such as entropy on an instance-level → fine-grained work-allocation decision
  - Model confidence as signal for model performance; Compute LLM uncertainty via (1) self-evaluation and (2) entropy
    - (1) LLMs can directly provide information about their uncertainty themselves (self-reported confidence)
    - (2) Black box LLMs' self-reported confidence not nec. reliable; Quantify the uncertainty assoc. w/ the class labels; the larger, the more uncertain
    - The top n instances are delegated to LLMs

Human-LLM collaboration for annotation

- Perspectives on LLMs for Relevance Judgment (Faggioli et al., 2023)
  - Spectrum of human-machine collaboration, task organization
- Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants (Bartolo et al., 2021)
  - Introduces Generative Annotation Assistants (GAAs), generator-in-the-loop models that provide real-time suggestions that annotators can either approve, modify, or reject entirely

Human-LLM collaboration for annotation

- Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels (Wang et al., 2024)
  - (1) LLMs generate labels and provide explanations,
  - (2) a verifier assesses the quality of LLM-generated labels, and
  - (3) human annotators re-annotate a subset of labels with lower verification scores
- MEGAnno+: A Human-LLM Collaborative Annotation System (Kim et al., 2024)
  - LLMs may fall short in understanding of complex, sociocultural, or domain-specific context $\rightarrow$ Human component deemed necessary
  - Provides workflow for human to utilize LLMs in text annotation

Intervention by LLM

- Calibration-Tuning: Teaching Large Language Models to Know What They Don't Know (Kapoor et al., 2024)
  - A model is well-calibrated when an outcome predicted w/ probab. p does occur p fraction of the time in reality.
  - "[I]t is desirable that LLMs be able to respond w/ a well-calibrated confidence, corresponding to a probability of correctness."
  - Calibration-tuning: An instruction tuning-inspired method for LLMs to output well-calibrated concept-level uncertainty estimates ($\rightarrow$ appl. to open-ended generation)

Intervention by LLM

- Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation (Kuhn et al., 2023)
  - Measuring uncertainty in natural language is challenging because of ‘semantic equivalence’ $\rightarrow$ Semantic entropy; an entropy which incorporates linguistic invariances
- Quantifying Uncertainty in Natural Language Explanations of Large Language Models (Tanneru et al., 2023)
  - Propose two novel metrics — Verbalized Uncertainty and Probing Uncertainty — to quantify the uncertainty of generated explanations

Intervention by LLM

- Language Models (Mostly) Know What They Know (Kadavath et al., 2022)
  - "[S]tudy whether language models can evaluate the validity of their own claims and predict which questions they will be able to answer correctly."
- Teaching models to express their uncertainty in words (Lin et al., 2022)
  - "[A] GPT-3 model can learn to express uncertainty about its own answers in natural language" using well-calibrated probabilities.
- The Calibration Gap between Model and Human Confidence in Large Language Models (Steyvers et al., 2024)
  - "[E]xplores the disparity between external human confidence in an LLM's responses and the internal confidence of the model"

Intervention by LLM

- Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness (Chen et al., 2023)
  - "[D]etecting bad and speculative answers from a pretrained Large Language Model by estimating a numeric confidence score for any output it generated."
- SaySelf: Teaching LLMs to Express Confidence with Self-Reflective Rationales (Xu et al., 2024)
  - Teach LLMs to express more fine-grained confidence estimates themselves
- To Believe or Not to Believe Your LLM (Yadkory et al., 2024)
  - Identifying when uncertainty in responses given a query is large
  - Epistemic uncertainties: Lack of knowledge about the ground truth
  - Aleatoric uncertainties: Irreducible randomnes (multiple possible answers)

Intervention by LLM

- Cycles of Thought: Measuring LLM Confidence through Stable Explanations (Becker et al., 2024)
  - "[M]easuring an LLM's uncertainty with respect to the distribution of generated explanations for an answer"
- Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs (Xiong et al., 2024)
- Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations (Huang et al., 2023)
  - "Study different ways to elicit the self-explanations [and] evaluate their faithfulness on a set of evaluation metrics"
- Can Unconfident LLM Annotations Be Used for Confident Conclusions? (Gligoric et al., 2024)
  - Combines LLM annotations and LLM confidence indicators to strategically select which human annotations should be collected

Finetuning on Explanation Generation

- Explain Yourself! Leveraging Language Models for Commonsense Reasoning
- Beyond Labels: Empowering Human Annotators with Natural Language Explanations through a Novel Active-Learning Architecture
- LEAN-LIFE: A Label-Efficient Annotation Framework Towards Learning from Explanation
- Chain-of-Thought Prompting Elicits Reasoning in Large Language Models
  - "[...] [H]ow generating a chain of thought – a series of intermediate reasoning steps – significantly improves the ability of [LLMs] to perform complex reasoning"

Personalized text generation

- Leveraging Similar Users for Personalized Language Modeling with Limited Data (Welch et al., 2022)
  - Personalizing language models not towards an individual novel user but rather a collection of known users with similar language patterns
- PersonalLLM: Tailoring LLMs to Individual Preferences
  - An alignment benchmark for "[...] adapting LLMs to provide maximal benefits for a particular user"
  - "[...] [A]ims to learn a unique users diverse preferences [...]"

Personalized text generation

- Adaptive Self-Supervised Learning Strategies for Dynamic On-Device LLM Personalization
  - "[...] [U]tilizes self-supervised learning techniques to personalize LLMs dynamically"
  - Collect interaction data on a user profiling layer, real-time fine-tuning w/ a neural adaptation layer $\rightarrow$ Continuous learning from user feedback
- A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More
- Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Datasets

Dataset	Task	Granularity	Collection	# Instances
MovieReviews [142]	sentiment classification	none	author	1,800
MovieReviews$_c$ [29]	sentiment classification	none	crowd	200‡◇
SST [113]	sentiment classification	none	crowd	11,855◇
WikiQA [136]	open-domain QA	sentence	crowd + authors	1,473
WikiAttack [22]	detecting personal attacks	none	students	1089◇
E-SNLI† [20]	natural language inference	none	crowd	~569K (1 or 3)
MultiRC [60]	reading comprehension QA	sentences	crowd	5,825
FEVER [118]	verifying claims from text	sentences	crowd	~136K‡
HotpotQA [137]	reading comprehension QA	sentences	crowd	112,779
Hanselowski et al. [47]	verifying claims from text	sentences	crowd	6,422 (varies)
NaturalQuestions [68]	reading comprehension QA	1 paragraph	crowd	n/a‡ (1 or 5)
CoQA [104]	conversational QA	none	crowd	~127K (1 or 3)
COS-E v1.0† [100]	commonsense QA	none	crowd	8,560
COS-E v1.11† [100]	commonsense QA	none	crowd	10,962
BoolQ$_c$ [29]	reading comprehension QA	none	crowd	199‡◇
EvidenceInference v1.0 [71]	evidence inference	none	experts	10,137
EvidenceInference v1.0$_c$ [29]	evidence inference	none	experts	125‡
EvidenceInference v2.0 [30]	evidence inference	none	experts	2,503
SciFact [123]	verifying claims from text	1-3 sentences	experts	995‡ (1-3)
Kutlu et al. [67]	webpage relevance ranking	2-3 sentences	crowd	700 (15)
SCAT [139]	document-level machine translation	none	experts	~14K
ECtHR [24]	alleged legal violation prediction	paragraphs	auto + expert	~11K
Hummingbird [48]	style classification	words	crowd	500
HateXplain [79]	hate-speech classification	phrases	crowd	20,148 (3)

Table 3: Overview of datasets with textual **highlights**. Values in parentheses indicate number of explanations collected per instance (if > 1). DeYoung et al. [29] collected or recollected annotations for prior datasets (marked with the subscript c). ◇ Collected > 1 explanation per instance but only release 1. † Also contains free-text explanations. ‡ A subset of the original dataset that is annotated. It is not reported what subset of NaturalQuestions has both a long and short answer.

Figure: Taken from "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing"

Datasets

Dataset	Task	Collection	# Instances
Jansen et al. [56]	science exam QA	authors	363
Ling et al. [76]	solving algebraic word problems	auto + crowd	~101K
Srivastava et al. [115]*	detecting phishing emails	crowd + authors	7 (30-35)
BABBLELABBLE [46]*	relation extraction	students + authors	200‡‡
E-SNLI [20]	natural language inference	crowd	~569K (1 or 3)
LIAR-PLUS [4]	verifying claims from text	auto	12,836
COS-E v1.0 [100]	commonsense QA	crowd	8,560
COS-E v1.11 [100]	commonsense QA	crowd	10,962
ECQA [2]	commonsense QA	crowd	10,962
SEN-MAKING [124]	commonsense validation	students + authors	2,021
CHANGEMYVIEW [10]	argument persuasiveness	crowd	37,718
WINOWHY [144]	pronoun coreference resolution	crowd	273 (5)
SBIC [111]	social bias inference	crowd	48,923 (1-3)
PUBHEALTH [64]	verifying claims from text	auto	11,832
Wang et al. [125]*	relation extraction	crowd + authors	373
Wang et al. [125]*	sentiment classification	crowd + authors	85
E-δ-NLI [18]	defeasible natural language inference	auto	92,298 (~8)
BDD-X†† [62]	vehicle control for self-driving cars	crowd	~26K
VQA-E†† [75]	visual QA	auto	~270K
VQA-X†† [94]	visual QA	crowd	28,180 (1 or 3)
ACT-X†† [94]	activity recognition	crowd	18,030 (3)
Ehsan et al. [34]††	playing arcade games	crowd	2000
VCR†† [143]	visual commonsense reasoning	crowd	~290K
E-SNLI-VE†† [32]	visual-textual entailment	crowd	11,335 (3)‡
ESPRIT†† [101]	reasoning about qualitative physics	crowd	2441 (2)
VLEP†† [72]	future event prediction	auto + crowd	28,726
EMU†† [27]	reasoning about manipulated images	crowd	48K

Table 4: Overview of EXNLP datasets with **free-text explanations** for textual and visual-textual tasks (marked with †† and placed in the lower part). Values in parentheses indicate number of explanations collected per instance (if > 1). ‡ A subset of the original dataset that is annotated. ‡‡ Subset publicly available. $*$ Authors semantically parse the collected explanations.

Figure: Taken from "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing"

Datasets

Dataset	Task	Explanation Type	Collection	# Instances
WorldTree V1 [57]	science exam QA	explanation graphs	authors	1,680
OpenBookQA [81]	open-book science QA	1 fact from WorldTree	crowd	5,957
Yang et al. [135]††	action recognition	lists of relations + attributes	crowd	853
WorldTree V2 [132]	science exam QA	explanation graphs	experts	5,100
QED [70]	reading comp. QA	inference rules	authors	8,991
QASC [61]	science exam QA	2-fact chain	authors + crowd	9,980
eQASC [58]	science exam QA	2-fact chain	auto + crowd	9,980 (∼10)
+ Perturbed	science exam QA	2-fact chain	auto + crowd	n/a‡
eOBQA [58]	open-book science QA	2-fact chain	auto + crowd	n/a‡
Ye et al. [138]*	SQuAD QA	semi-structured text	crowd + authors	164
Ye et al. [138]*	NaturalQuestions QA	semi-structured text	crowd + authors	109
R^{4}C [53]	reading comp. QA	chains of facts	crowd	4,588 (3)
StrategyQA [41]	implicit reasoning QA	reasoning steps w/ highlights	crowd	2,780 (3)
TriggerNER	named entity recognition	groups of highlighted tokens	crowd	∼7K (2)

Table 5: Overview of ExNLP datasets with **structured explanations** (§5). Values in parentheses indicate number of explanations collected per instance (if >1). †† Visual-textual dataset. ∗ Authors semantically parse the collected explanations. ‡ Subset of instances annotated with explanations is not reported. Total # of explanations is 855 for eQASC Perturbed and 998 for eOBQA.

Figure: Taken from "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing"

Datasets

	TASK								MODEL						EX. TYPE			METHOD					RESULTS			
Authors	DOC. CLS.	SENTIMENT ANALYSIS	HATE SPEECH	REL. EXTRACTION	MULTI CHOICE QA	NATURAL LANG. INFERENCE	EXTRACTIVE QA	OTHER	LOGREG	NAIVE BAYES	SVM	CNN	LSTM	PRETR. TRANSFORMER	HIGHLIGHT	FREE-TEXT	SEMI-STRUCTURED	REGULARIZATION	DATA AUGMENTATION	MTL	I→EX;EX→O	OTHER	OUT-OF-DOMAIN	EFFICIENCY	BIAS REDUCTION	MODEL EXPLANATION
Godbole et al. (2004)	■										■				■							■		▲		
Zaidan et al. (2007)		■									■				■							■		▲		
Zaidan and Eisner (2008)		■													■							■		▲		
Druck et al. (2009)								■							■							■		▲		
Small et al. (2011)		■									■				■							■		▲		
Settles (2011)	■														■							■		▲		
Kulesza et al. (2015)	■									■					■							■		▲		
Zhang et al. (2016)	■											■			■						■			▲		▲
Bao et al. (2018)		■												■	■							■	▲			
Zhong et al. (2019)		■											■		■			■						▲		•
Liu and Avci (2019)			■									■			■			■						▲	▲	
Strout et al. (2019)		■										■			■			■						▲		▲
Rieger et al. (2020)		■											■		■			■							▲	
Stacey et al. (2022)						■								■	■			■					▲	▲		
Carton et al. (2021)	■				■	■								■	■						■			▲		▲
Mathew et al. (2021)			■												■			■							▲	•
Antognini et al. (2021)								■						□	■					■				▲		▲
Pruthi et al. (2022)		■					■							■	■			■		■				▲		
Srivastava et al. (2017)	■								■	■						■						■		▲		
Hancock et al. (2018)				■					■							■			■					▲		
Wang et al. (2020)		■		■									■			■			■					▲		
Lee et al. (2020)		■		■				■					■		■	■			■					▲		
Ye et al. (2020)							■							■			■		■					▲		
Murty et al. (2020)				■										■		■						■		▲		
Yao et al. (2021)		■	■											■			■		■				▲	▲	▲	
Camburu et al. (2018)						■							■			■				■	■		•	•		▲
Rajani et al. (2019)					■									■		■					■		•	▲		▲
Kumar and Talukdar (2020)						■								■		■					■		▼	▲		▲
Zhao and Vydiswaran (2021)						■								■	■	■					■				•	

Figure: Taken from "A survey on improving NLP models with human explanations"

Datasets

Related work	Classification task	Granularity	Form		Value type	Collection aim: Improving ML	Task insight	Data quality	Gold explanation	Data generation	Annotator	Name (if available)
Zaidan et al. (2007)	Sentiment	Sn	E		C	✓					O	IMDB
Titov and McDonald (2008)	Sentiment	Se	E		C				✓		O	TripAdvisor*
Yano et al. (2010)	Bias	Sn	E		C		✓				Cw	
Abedin et al. (2011)	Aviation incident causes	Sn	E		C	✓				✓	O	ASRS
McAuley et al. (2012)	Sentiment	S	E		C	✓			✓		De	BeerAdvocate
Saleem et al. (2012)	Medical	Sn	E		C	✓					De	
Xia and Yetisgen-Yildiz (2012)	Medical	S		A	C			✓			De	
Tepper et al. (2013)	Medical	Sn	E		C	✓					De	CPIS/PNA
Marshall et al. (2015)	Bias	Sn	E		C				✓		De	RoB
McDonnell et al. (2016)	Webpage relevance	Se	E	A	C			✓			Cw	
Bao et al. (2018)	Sentiment	Sn	E		C	✓					O	BeerAdvocate*
Carton et al. (2018)	Personal attacks	Sn	E		C				✓		O	
Chhatwal et al. (2018)	Legal	Sn	E	A	C	✓					De	
Kaushik et al. (2019)	Sentiment	Sn	E		C	✓					Cw	IMDB*
Ramírez et al. (2019)	Topic	Sn	E	A	N		✓				Cw	SLR

Figure: Taken from "Human-annotated rationales and explainable text classification: a survey"

Datasets

Ramírez et al. (2019)	Topic	Sn	E	A	C		✓				Cw	Amazon
Wang et al. (2020)	Sentiment	Se		A	C					✓	Cw	SemEval-2014*
Hasanain et al. (2020)	Topic	Se	E	A	C	✓	✓		✓		De	ArTest
Kanchinadam et al. (2020)	Sentiment	Sn	E		C	✓					Cw	IMDB*
Kartal and Kutlu (2020)	Check-worthy claims	Sn		A	C		✓				O	TrClaim-19
Kreiss et al. (2020)	Guilt	Sn	E		C	✓	✓				Cw	SuspectGuilt
Kutlu et al. (2020)	Webpage relevance	Se	E	A	C			✓			Cw	
Sap et al. (2020)	Abusive content	Se		A	C	✓			✓		Cw	SBIC
Sen et al. (2020)	Sentiment	Sn	E		C				✓		Cw	Yelp-HAT
Arous et al. (2021)	Topic	Sn	E		C	✓			✓		Cw	Wiki-Tech
Chalkidis et al. (2021)	Legal	P	E		C				✓		De	ECtHR
Hayati et al. (2021)	Style	W	E		C				✓		Cw	Hummingbird
Jayaram and Allaway (2021)	Stance detection	W	E		C	✓					Cw	VAST*
Mohseni et al. (2021)	Sentiment	Sn	E		C				✓		Cw	IMDB*
Mohseni et al. (2021)	Topic	Sn	E		C				✓		Cw	20News*
Mathew et al. (2021)	Hate speech	Sn	E		N	✓			✓		Cw	HateXplain
Malik et al. (2021)	Legal	Se	E		C				✓		De	ILDC
Sharma et al. (2020)	Empathy expression	Sn	E		C				✓		Cw	EMH
Vidgen et al. (2021)	Abusive content	Sn	E		C				✓		De	CAD
El Zini et al. (2022)	Sentiment	W	E		C				✓		O	RottenTomatoes*
Chiang and Lee (2022)	Sentiment	Sn	E		C				✓		Cw	IMDB*

Figure: Taken from "Human-annotated rationales and explainable text classification: a survey"

Datasets

Related work	Classification task	Granularity	Form		Value type	Collection aim: Improving ML	Collection aim: Task insight	Collection aim: Data quality	Collection aim: Gold explanation	Collection aim: Data generation	Annotator	Name (if available)
Guzman et al. (2022)	Forced labor indicators	Sn	E		C	✓					De	RaFoLa
Jørgensen et al. (2022)	Sentiment	W	E		C				✓		O	SST*
Lu et al. (2022)	Sentiment	Sn	E		C	✓					Cw	IMDB*
Sullivan et al. (2022)	Sentiment	Sn	E		C		✓				Cw	IMDB*
Wang et al. (2022)	Topic	Sn	E		C	✓					O	AIvsCR
Jakobsen et al. (2023)	Sentiment	W	E		C	✓					Cw	DynaSent*
Jakobsen et al. (2023)	Sentiment	W	E		C	✓					Cw	SST*

Granularity is abbreviated as Paragraphs, Sentences, Snippets, and Words. Form is abbreviated as Extractive and Abstractive. Values types are abbreviated as Categorical and Numerical. The annotator type is abbreviated as Crowd worker, Domain expert, and Other. When available, the name of the dataset is provided. The * symbol is used when human-annotated rationales are added to an already existing dataset.

Figure: Taken from "Human-annotated rationales and explainable text classification: a survey"

Motivation, Problem and Research Questions

- Human annotation on subjective tasks such as hate speech detection can be quite difficult
- Have an LLM aid the annotator during the annotation process by occasionally providing possible explanations for a label it finds most fitting
- Such an intervention should be done when a model sees a label contrary to the one given by the human to be more fitting or also fitting
- LLM may thus be able to help human annotator on more nuanced, complex instances to arrive at the most plausible label
- The LLM should be aligned with some annotator's views (ideally an expert; not necessarily the same one)
- Extensive use of explanations furthers better understanding of disagreements

Motivation, Problem and Research Questions

- Problems
  - Alignment/Personalization of an LLM towards the views/values of a human annotator
  - Number of models to use (e.g. a personalized and a general one)
  - How to personalize a model, towards whom, and if it should be done at all
  - Which model(s) to use
  - When the model(s) should intervene
  - How to evaluate the annotation accuracy in the conversational case
- Research questions
  - Does the support of an LLM in the human annotation process lead to an increase in accuracy and also better understanding/trustworthiness of the provided labeling?
  - In what way should an LLM be personalized, i.e. aligned with a human, to best aid them in their annotation task?

Thesis Goals and Tasks to Tackle Each Goal

- Arrive at an LLM that is aligned with the values/views of a human annotator, i.e. a personalized LLM, able to provide appropriate explanations for why the labeling of some data instance is correct or not
- Have this LLM be able to judge whether or not it should intervene to rectify a label or provide a different view that supports a different choice of label more aligned with this personalization
- Ultimately improve the annotation accuracy of a human on various tasks/datasets aided by such an LLM

Outline

1. Gather related work
2. Choose tasks/datasets, collect own human-annotated explanations
3. Decide on model/architecture
4. Model personalization
5. Finetuning of model(s) on explanation generation
6. Explanation learning; personalization of model(s)
7. Annotation
8. Evaluation
9. Shaping and writing the thesis

Time Schedule

1. 2 wks.
2. 1 wk.
3. 1 wk.
4. 1 mo.
5. 1 mo.
6. 1 mo.
7. 2 wks.
8. 2 wks.
9. 2 mo.