ChatGPT vs. Crowdsourcing vs. Experts: Annotating Open-Domain Conversations with Speech Functions

Lidiia Ostyakova[1,2]*, Veronika Smilga[1]*, Kseniia Petukhova[1]*, Maria Molchanova[1], Daniel Kornev[1]

[1]Moscow Institute of Physics and Technology, DeepPavlov Lab
[2]HSE University

*These authors contributed equally to this work

Discourse & Pragmatics:
Theoretical Approaches for Dialogue Analysis

Theory of Speech Acts
(Searle, 1969)

Speaker's intentions are embedded in the EDU.

DAMSL
Dialog Act Markup in Several Layers
James Allen and Mark Core (1997)

SWBD-DAMSL
Switchboard Shallow-Discourse-Function Annotation
Dan Jurafsky, Liz Shriberg, and Debra Biasca (1997)

DiAML (ISO standard)
Dialog Act Markup Language
Harry Bunt, Michael Kipp, and Volha Petukhova (2009)

MIDAS
A Dialog Act Annotation Scheme for Open-Domain Human-Machine Spoken Conversations
Dian Yu, Zhou Yu (2019)

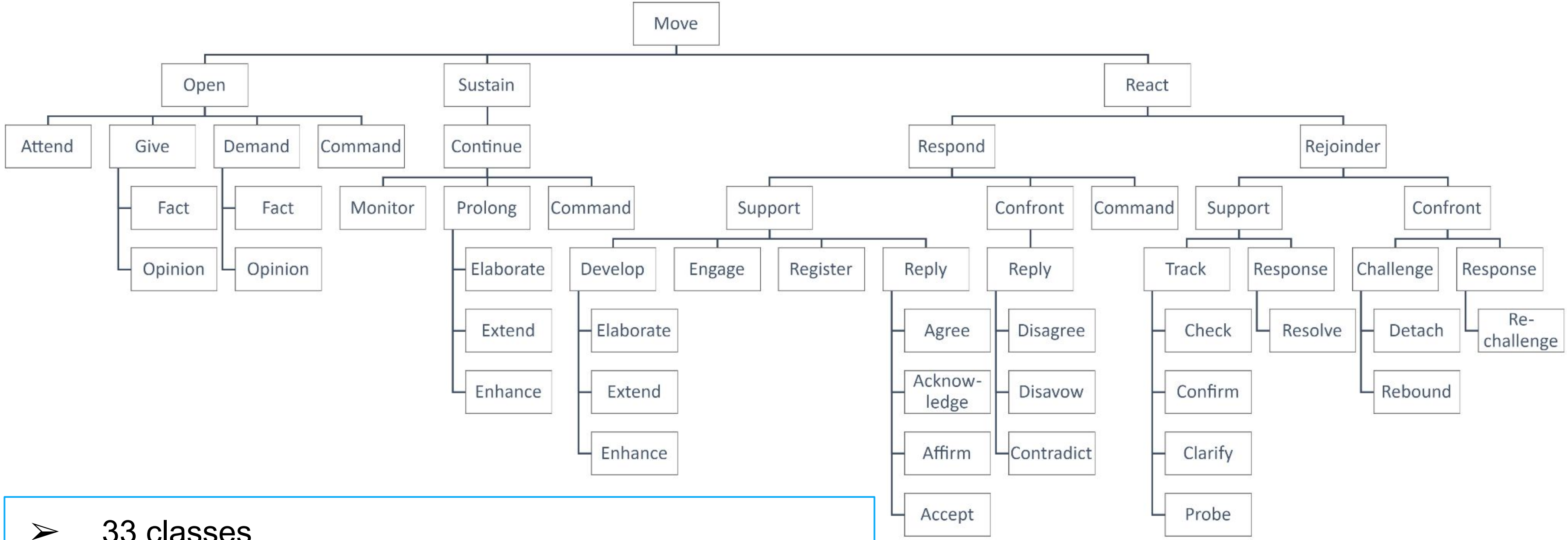
Rhetorical Structure Theory
(Mann, W. C., Thompson, S. A, 1978)

Relations between EDUs have to be defined and then characterized with a pragmatic class.

SDRT
Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure
Alex Lascarides, Nicholas Asher (2016)

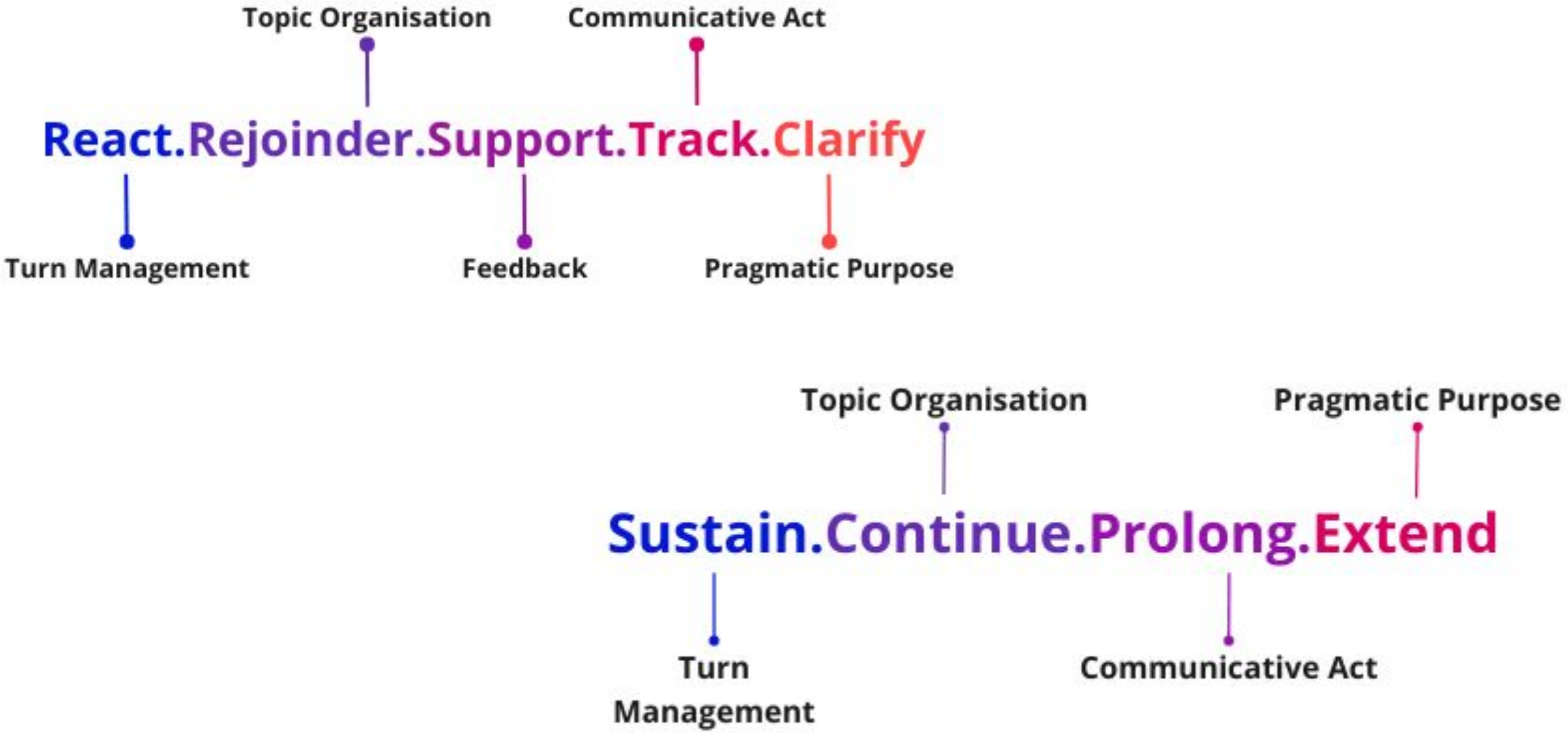
DDA (Dependency Dialogue Acts) Jon Z. Cai, Brendan King, Margaret Perkoff, et al. (2023)

Speech Function Taxonomy



- ➢ 33 classes
- ➢ designed for analyzing casual conversations
- ➢ a hierarchical taxonomy including several layers of annotation
- ➢ a topic-oriented taxonomy
- ➢ 5 functional dimensions

Speech Function Theory: Functional Dimensions



Topic Organisation

Communicative Act

React.Rejoinder.Support.Track.Clarify

Turn Management

Feedback

Pragmatic Purpose

Topic Organisation

Pragmatic Purpose

Sustain.Continue.Prolong.Extend

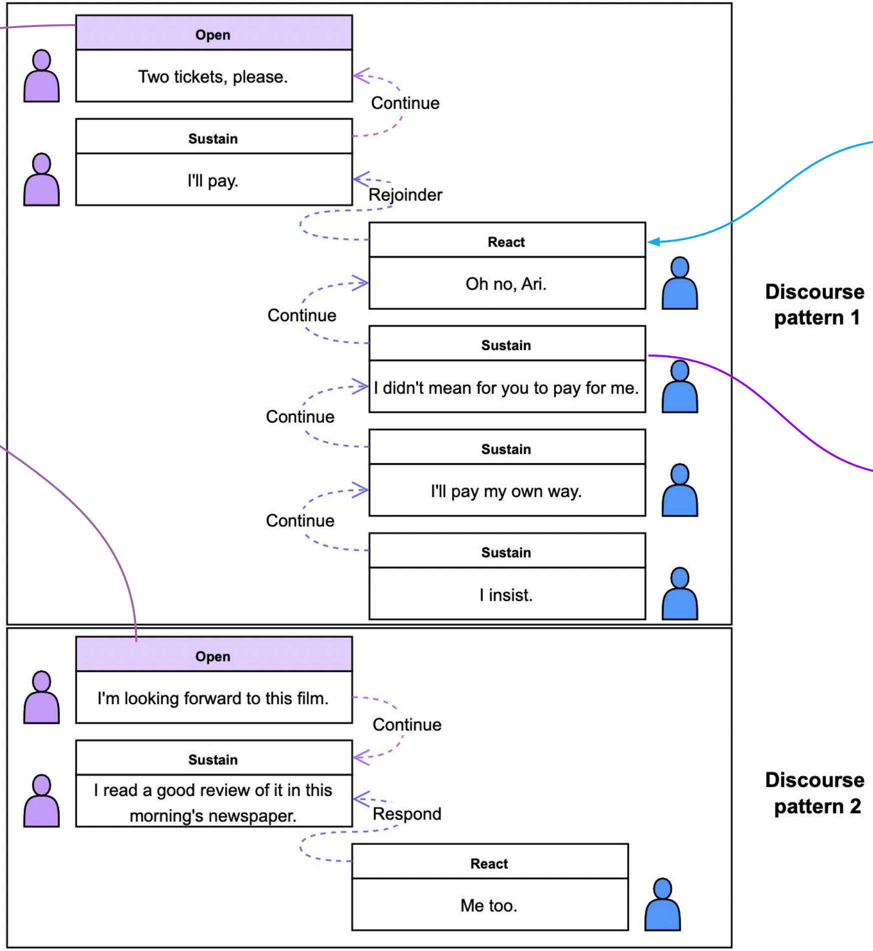
Turn Management

Communicative Act

Motivation

- There is not enough data annotated using a multi-layer scheme.
- Discourse annotation with Large Language Models has not been researched enough.
- There are no strategies for prompting LLMs to perform complex discourse annotation.

Open moves define a new topic or a start of a dialogue



Discourse pattern 1

Open

Two tickets, please.

Continue

Sustain

I'll pay.

Rejoinder

React

Oh no, Ari.

Continue

Sustain

I didn't mean for you to pay for me.

Continue

Sustain

I'll pay my own way.

Continue

Sustain

I insist.

Discourse pattern 2

Open

I'm looking forward to this film.

Continue

Sustain

I read a good review of it in this morning's newspaper.

Respond

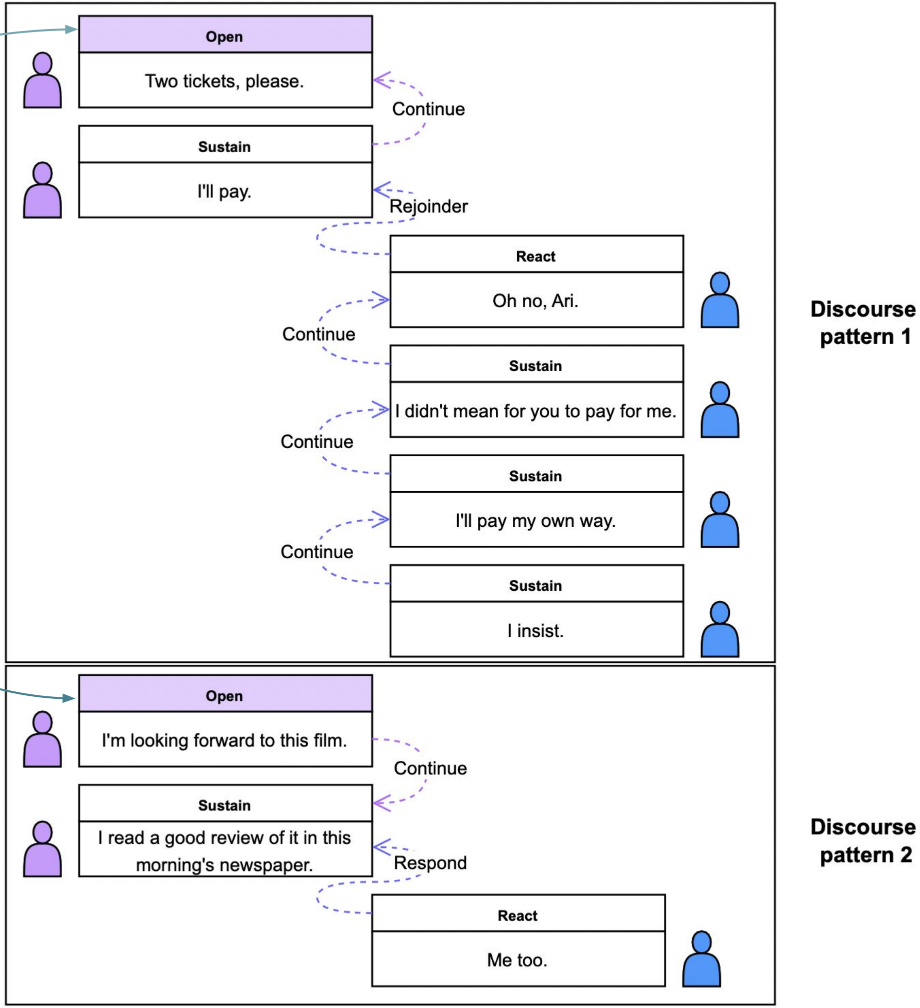
React

Me too.

React moves denote reactions to the previous utterances of other speaker

Sustain moves portray a topic development provided by the same speaker

Open moves define **discourse patterns** within a dialogue



Open

Two tickets, please.

Continue

Sustain

I'll pay.

Rejoinder

React

Oh no, Ari.

Continue

Sustain

I didn't mean for you to pay for me.

Continue

Sustain

I'll pay my own way.

Continue

Sustain

I insist.

Discourse pattern 1

Open

I'm looking forward to this film.

Continue

Sustain

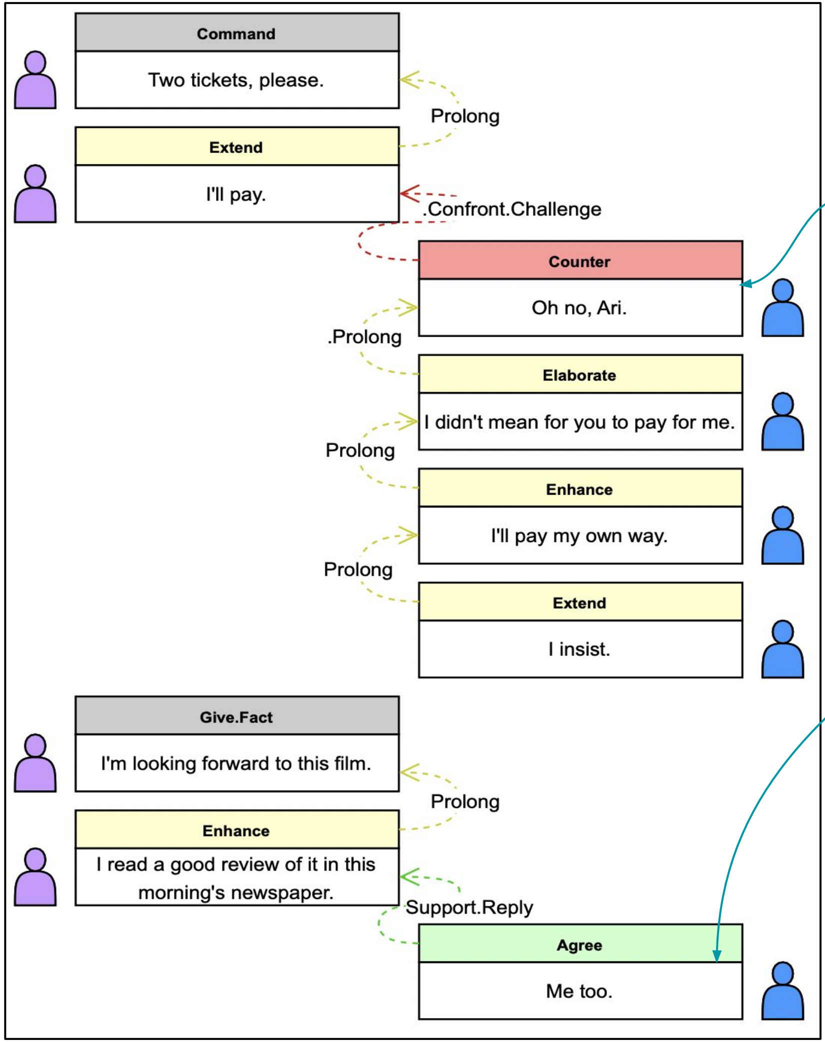
I read a good review of it in this morning's newspaper.

Respond

React

Me too.

Discourse pattern 2



Command

Two tickets, please.

Prolong

Extend

I'll pay.

.Confront.Challenge

Counter

Oh no, Ari.

.Prolong

Elaborate

I didn't mean for you to pay for me.

Prolong

Enhance

I'll pay my own way.

Prolong

Extend

I insist.

Give.Fact

I'm looking forward to this film.

Prolong

Enhance

I read a good review of it in this morning's newspaper.

Support.Reply

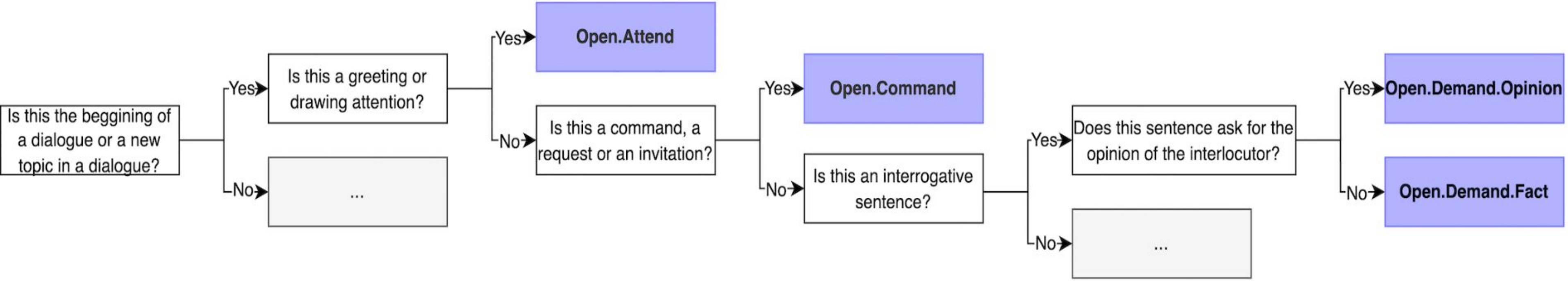
Agree

Me too.

Confront and **Support** moves define negative or positive speaker's *feedback* on the interlocutor's previous utterances

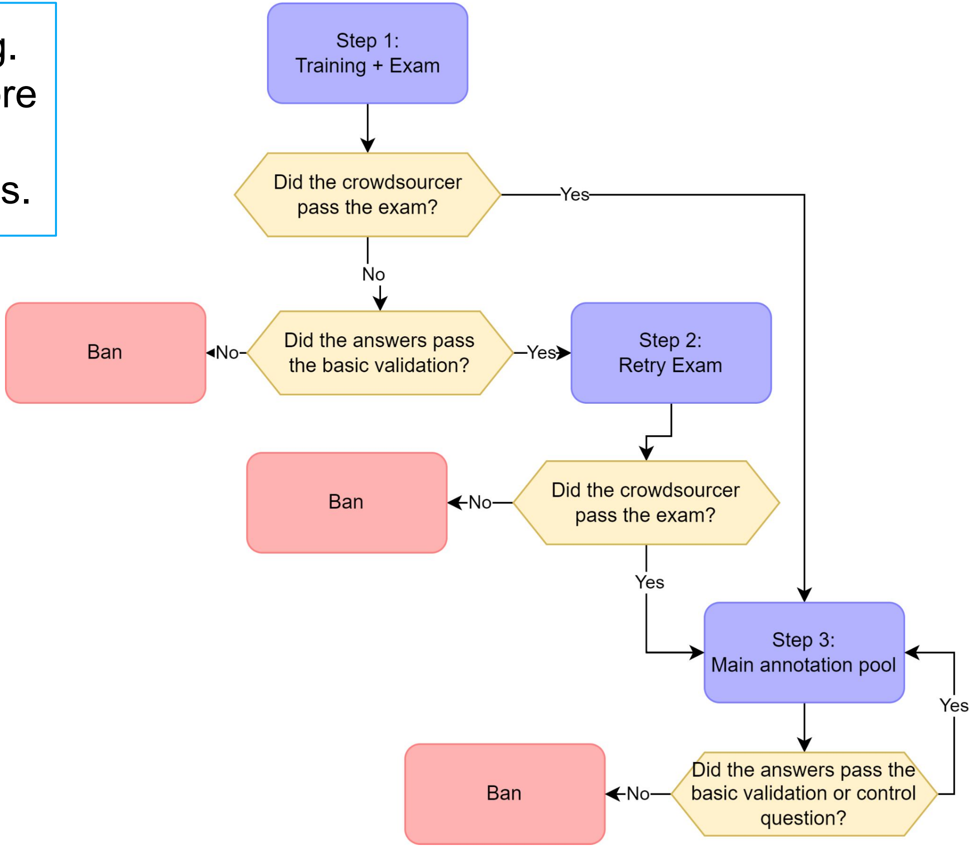
Design of Guidelines for Annotation

- ➢ Annotators answer simple questions about a previous utterance and a current one.
- ➢ A number of questions varies from a particular utterance and its communicative function in the dialogue.
- ➢ All questions are provided with the examples.
- ➢ Gold standard (64 dialogs from DailyDialog) was annotated using these guidelines.


Is this the beggining of a dialogue or a new topic in a dialogue?
Yes
Is this a greeting or drawing attention?
Yes
Open.Attend
No
Is this a command, a request or an invitation?
Yes
Open.Command
No
Is this an interrogative sentence?
Yes
Does this sentence ask for the opinion of the interlocutor?
Yes
Open.Demand.Opinion
No
Open.Demand.Fact
No
...
No
...

Crowdsourcing: Annotation Process

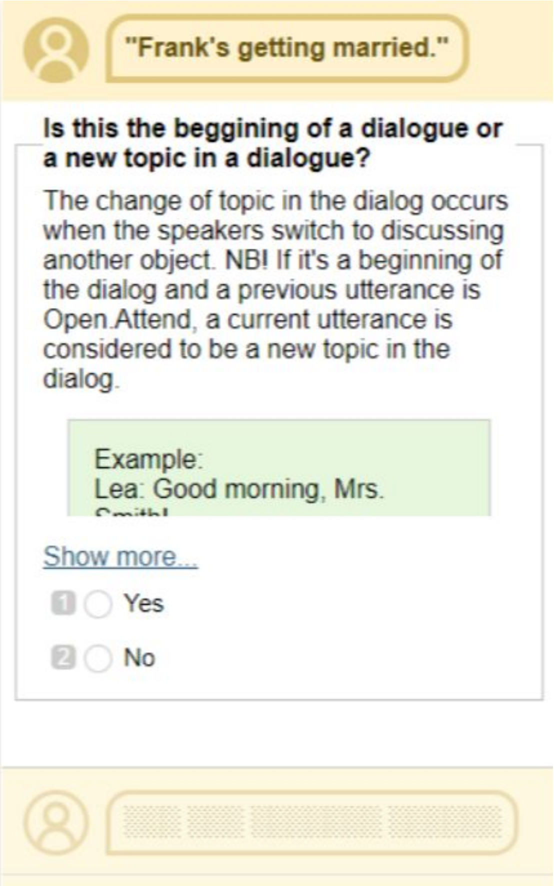
- ➢ Toloka platform was used for crowdsourcing.
- ➢ All the crowdsourcers had an exam before annotation.
- ➢ All questions are provided with the examples.



Step 1:
Training + Exam
Did the crowdsourcer pass the exam?
Yes
No
Ban
No
Did the answers pass the basic validation?
Yes
Step 2:
Retry Exam
Ban
No
Did the crowdsourcer pass the exam?
Yes
Step 3:
Main annotation pool
Yes
Ban
No
Did the answers pass the basic validation or control question?

Crowdsourcing

- ➢ *<u>The key criterion</u>* for recruitment was the successful completion of the test task assessing the annotators' labeling quality.
- ➢ Access to the test task was granted to those who previously passed *<u>the English language proficiency test</u>* on the Toloka platform.
- ➢ The largest number of annotators originated from Brazil and Egypt.



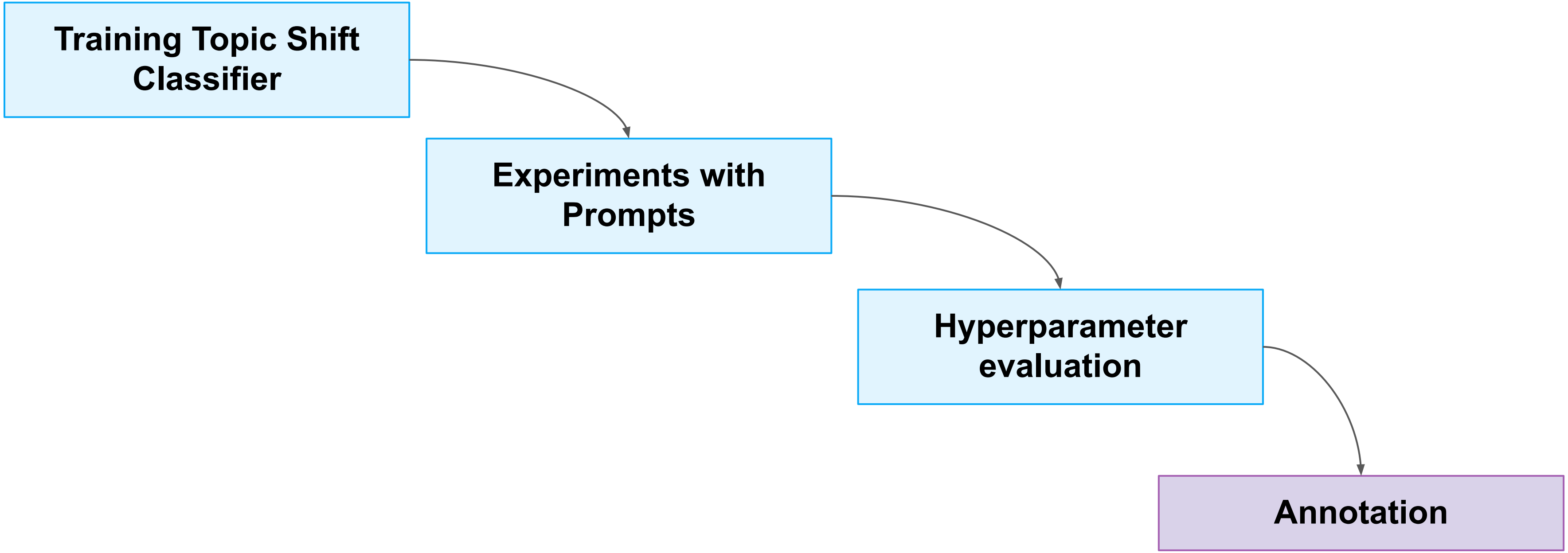
"Frank's getting married."

Is this the beggining of a dialogue or a new topic in a dialogue?

The change of topic in the dialog occurs when the speakers switch to discussing another object. NB! If it's a beginning of the dialog and a previous utterance is Open.Attend, a current utterance is considered to be a new topic in the dialog.

Example:
Lea: Good morning, Mrs.

ChatGPT Annotation: Pipeline



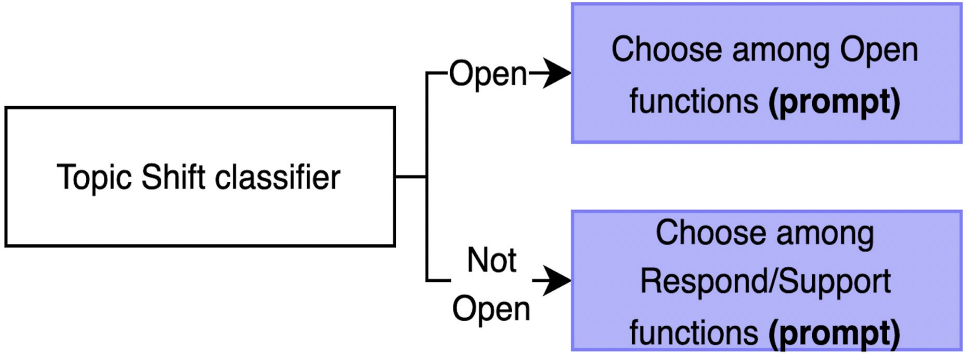
Training Topic Shift Classifier

Experiments with Prompts

Hyperparameter evaluation

Annotation

ChatGPT Annotation: Direct Annotation



Topic Shift classifier

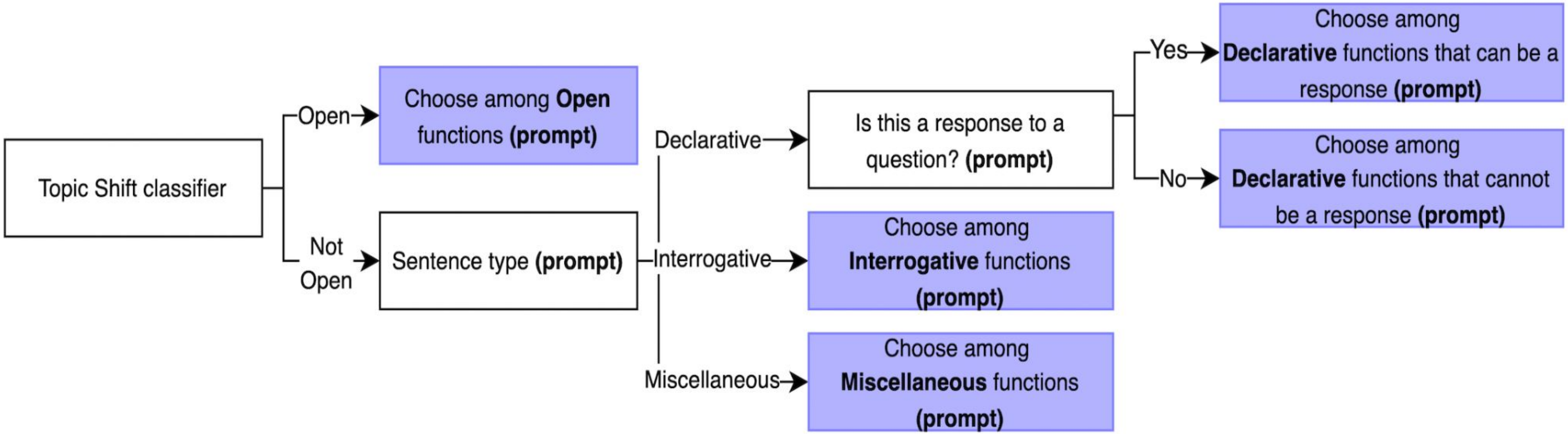
Open

Not Open

Choose among Open functions **(prompt)**

Choose among Respond/Support functions **(prompt)**

ChatGPT Annotation: Step-by-Step scheme



Topic Shift classifier

Open

Choose among **Open** functions **(prompt)**

Not Open

Sentence type **(prompt)**

Declarative

Is this a response to a question? **(prompt)**

Yes

Choose among **Declarative** functions that can be a response **(prompt)**

No

Choose among **Declarative** functions that cannot be a response **(prompt)**

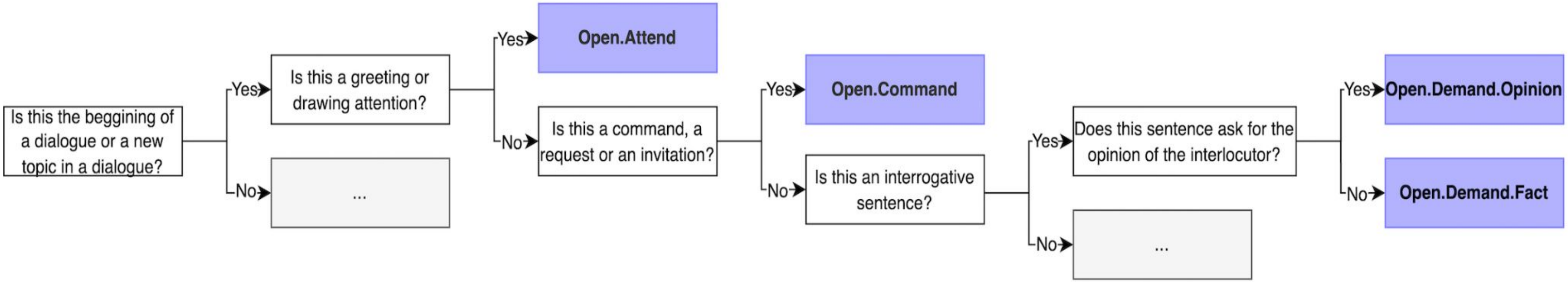
Interrogative

Choose among **Interrogative** functions **(prompt)**

Miscellaneous

Choose among **Miscellaneous** functions **(prompt)**

ChatGPT Annotation: Tree-like Scheme



Is this the beggining of a dialogue or a new topic in a dialogue?

Yes → Is this a greeting or drawing attention?

No → ...

Is this a greeting or drawing attention? Yes → Open.Attend

Is this a greeting or drawing attention? No → Is this a command, a request or an invitation?

Is this a command, a request or an invitation? Yes → Open.Command

Is this a command, a request or an invitation? No → Is this an interrogative sentence?

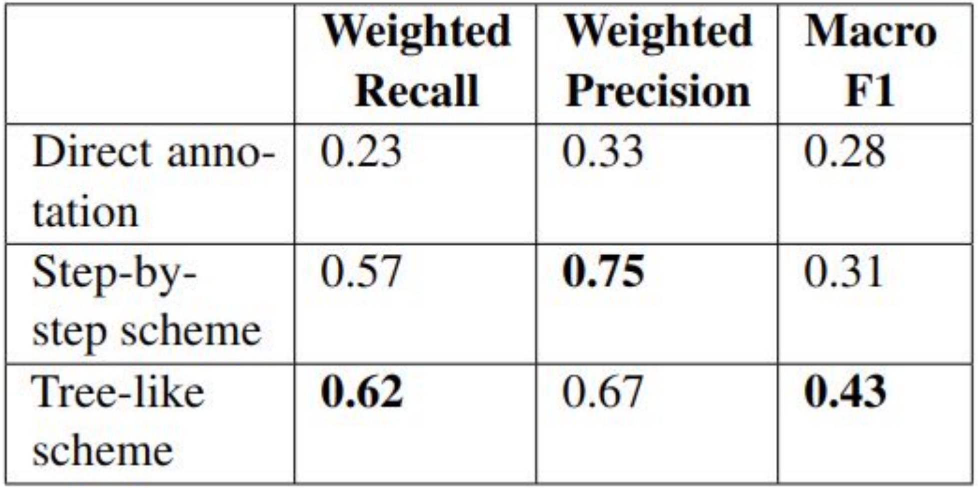
Is this an interrogative sentence? Yes → Does this sentence ask for the opinion of the interlocutor?

Is this an interrogative sentence? No → ...

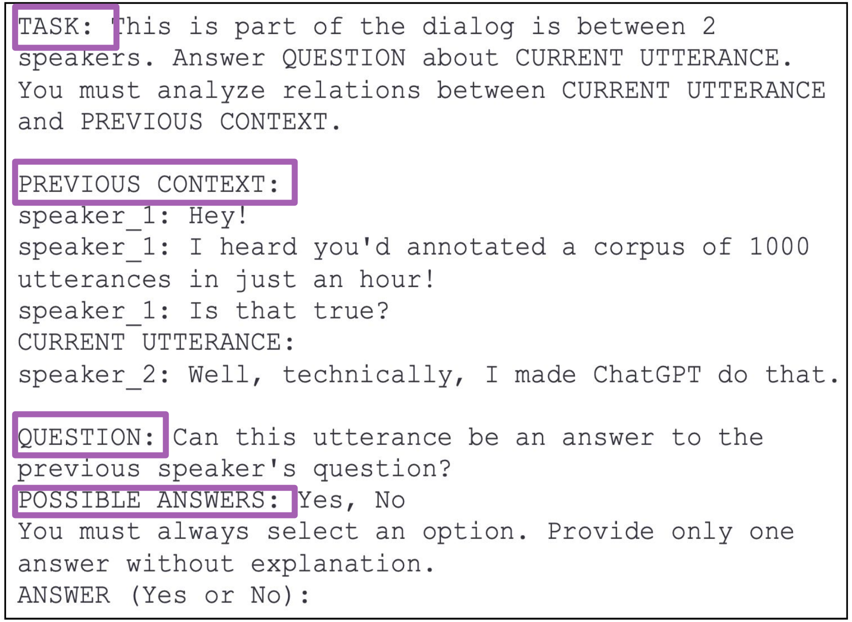
Does this sentence ask for the opinion of the interlocutor? Yes → Open.Demand.Opinion

Does this sentence ask for the opinion of the interlocutor? No → Open.Demand.Fact

ChatGPT Annotation

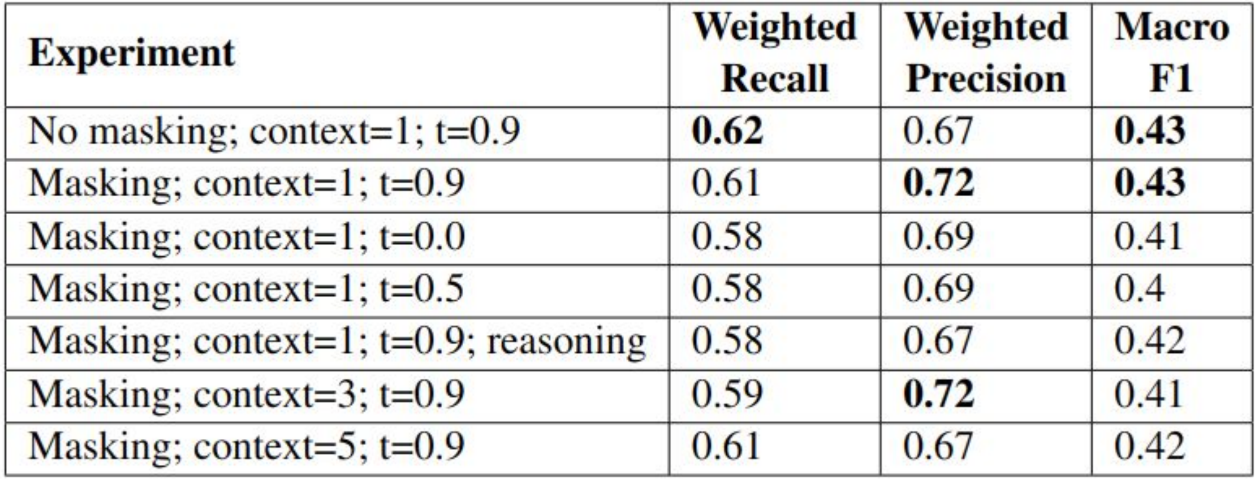


	Weighted Recall	Weighted Precision	Macro F1
Direct annotation	0.23	0.33	0.28
Step-by-step scheme	0.57	**0.75**	0.31
Tree-like scheme	**0.62**	0.67	**0.43**



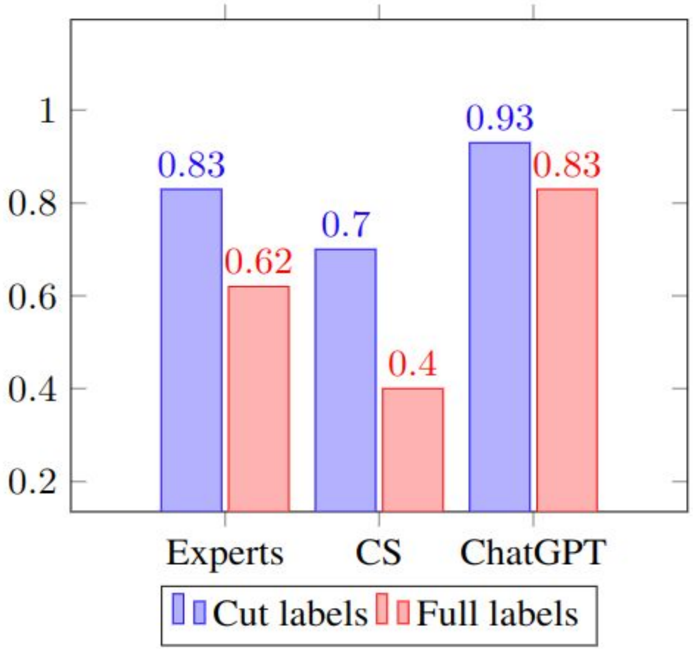
```
TASK: This is part of the dialog is between 2
speakers. Answer QUESTION about CURRENT UTTERANCE.
You must analyze relations between CURRENT UTTERANCE
and PREVIOUS CONTEXT.

PREVIOUS CONTEXT:
speaker_1: Hey!
speaker_1: I heard you'd annotated a corpus of 1000
utterances in just an hour!
speaker_1: Is that true?
CURRENT UTTERANCE:
speaker_2: Well, technically, I made ChatGPT do that.

QUESTION: Can this utterance be an answer to the
previous speaker's question?
POSSIBLE ANSWERS: Yes, No
You must always select an option. Provide only one
answer without explanation.
ANSWER (Yes or No):
```

ChatGPT Annotation: Results



Experiment	**Weighted Recall**	**Weighted Precision**	**Macro F1**
No masking; context=1; t=0.9	**0.62**	0.67	**0.43**
Masking; context=1; t=0.9	0.61	**0.72**	**0.43**
Masking; context=1; t=0.0	0.58	0.69	0.41
Masking; context=1; t=0.5	0.58	0.69	0.4
Masking; context=1; t=0.9; reasoning	0.58	0.67	0.42
Masking; context=3; t=0.9	0.59	**0.72**	0.41
Masking; context=5; t=0.9	0.61	0.67	0.42

ChatGPT vs. Crowdsourcing vs. Experts: Inter-annotator Agreement

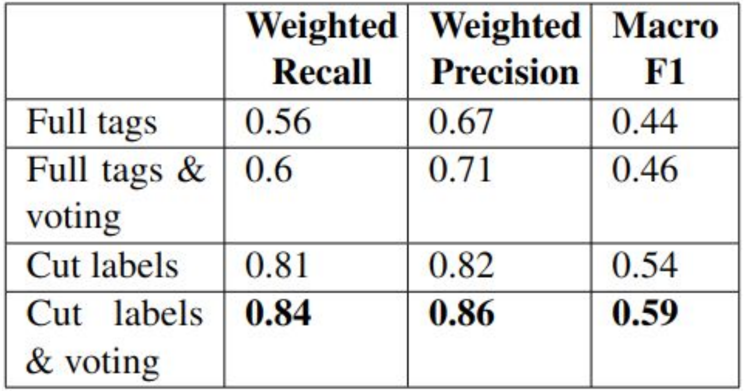


Conclusions

- ➢ Inter-annotator agreement between crowdsourcers for full labels is quite low.
- ➢ It is impossible to control the annotation quality to a full extend while crowdsourcing.
- ➢ ChatGPT performance is quite stable.

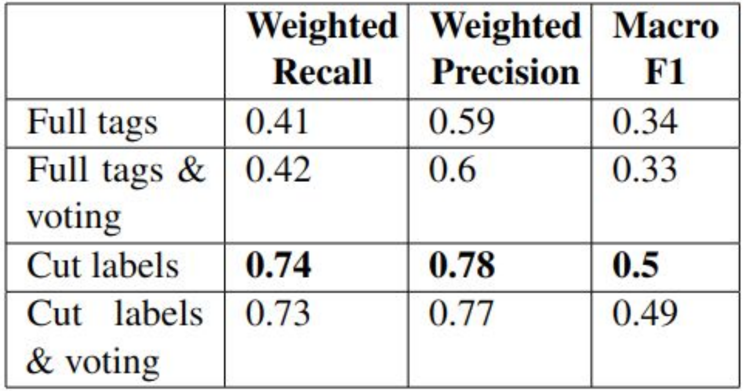
Cut label: Sustain.Continue.Prolong
Full label: Sustain.Continue.Prolong.Extend

ChatGPT vs. Crowdsourcing vs. Experts



	Weighted Recall	Weighted Precision	Macro F1
Full tags	0.56	0.67	0.44
Full tags & voting	0.6	0.71	0.46
Cut labels	0.81	0.82	0.54
Cut labels & voting	**0.84**	**0.86**	**0.59**

(a) Crowdsourcers



	Weighted Recall	Weighted Precision	Macro F1
Full tags	0.41	0.59	0.34
Full tags & voting	0.42	0.6	0.33
Cut labels	**0.74**	**0.78**	**0.5**
Cut labels & voting	0.73	0.77	0.49

(b) ChatGPT

Conclusions & Future Work

- Experiments with ChatGPT have demonstrated the potential of using LLMs for linguistic annotation with accuracy that is close to crowdsourcing workers' performance on some dialogs.
- Experts are needed for developing guidelines (prompts) and the validation of the annotation.
- Possible areas for the *future work* are:
  - trying out other instruction-based models;
  - conducting a more comprehensive selection of hyperparameters;
  - adding criticism steps to the current pipeline, enabling self-reflection and self-correction.

Lidiia Ostyakova lostaaa15@gmail.com

Thank you for attention!