

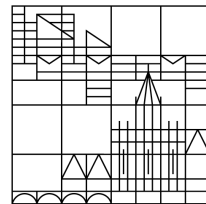
Argument Quality Prediction for Ranking Documents

Moritz Plenz, Raphael Buchmüller and Alexander Bondarenko
From Team Renji Abarai



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

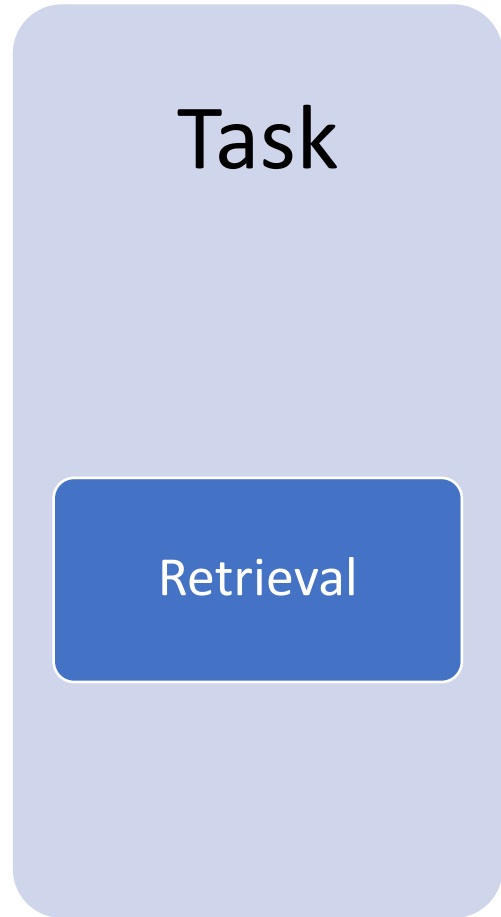
Universität
Konstanz



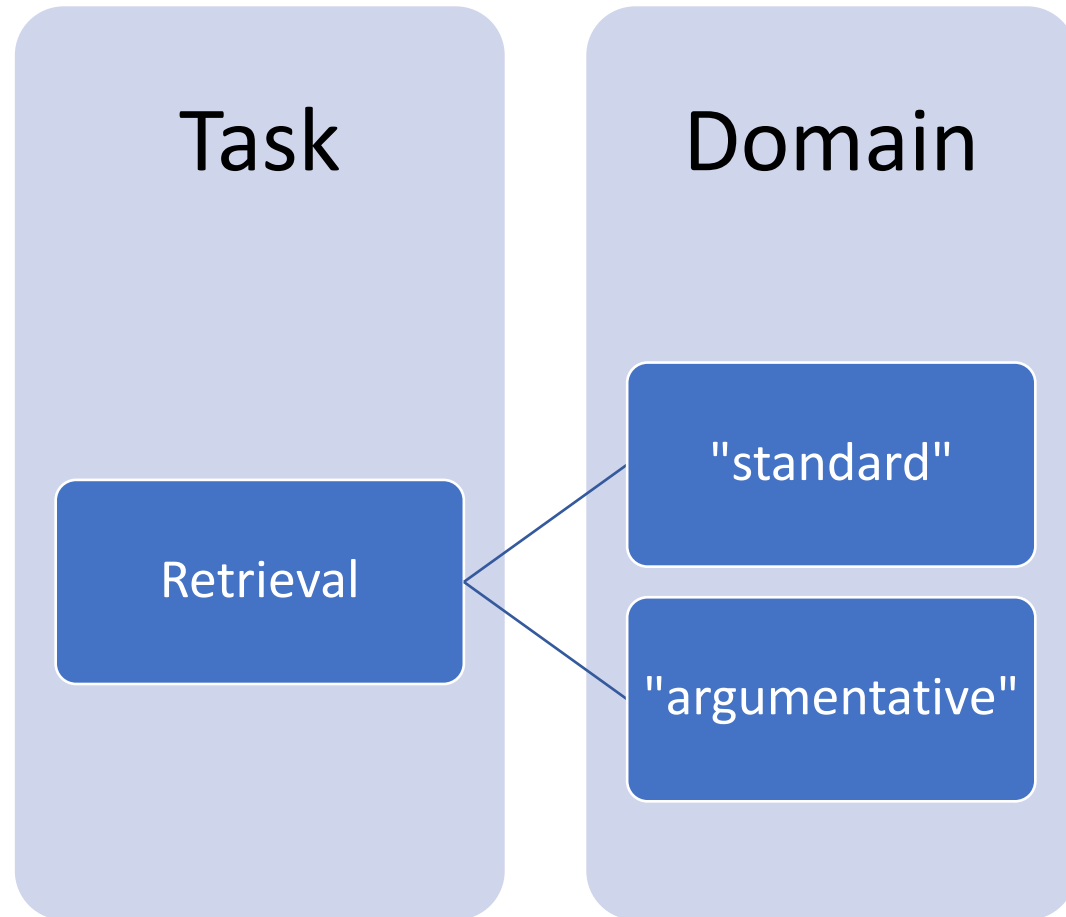
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA



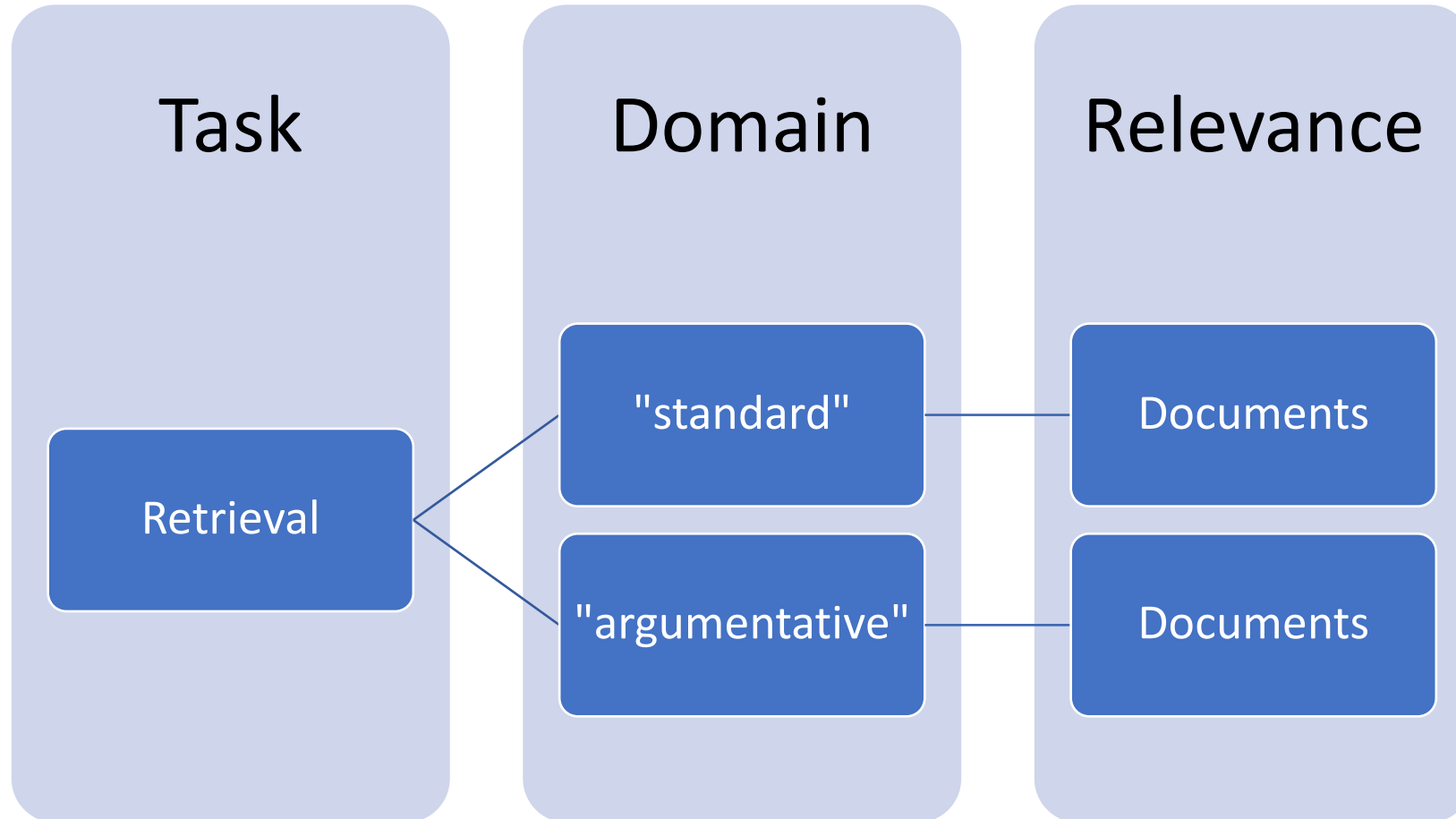
Motivation



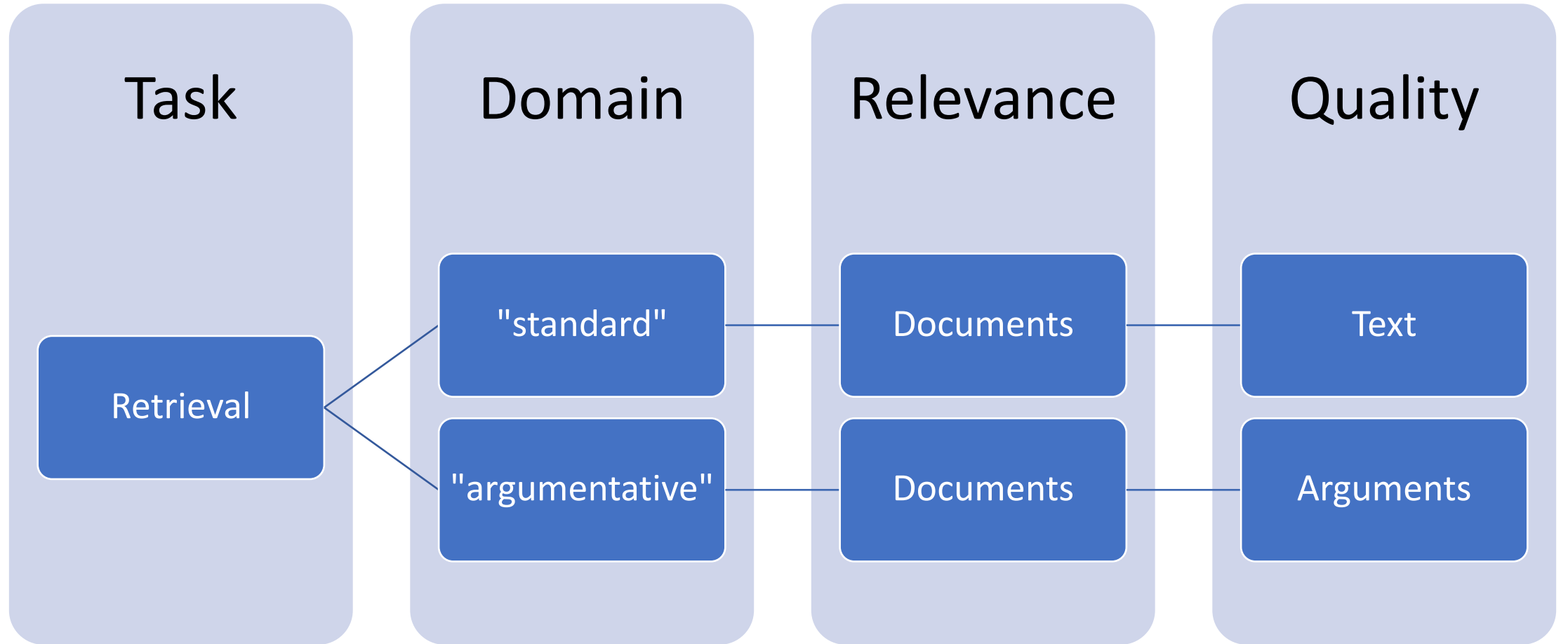
Motivation



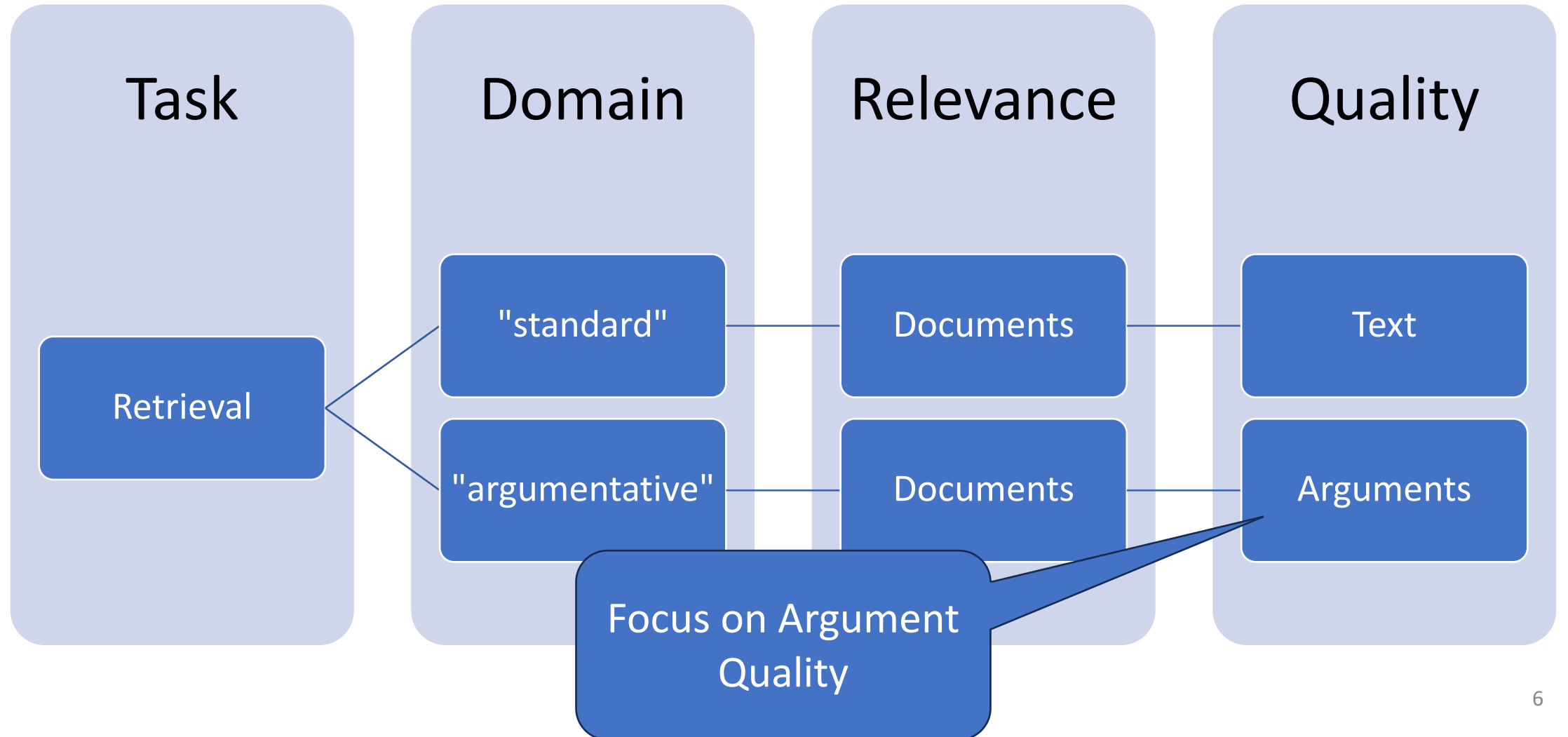
Motivation



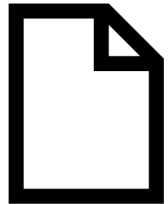
Motivation



Motivation



Initial Retrieval



Relevance: 3

Quality: 2

Stance: None



Relevance: 2

Quality: 3

Stance: Pro



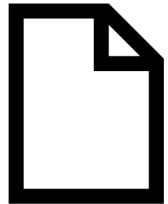
Relevance: 1

Quality: 1

Stance: Con



Initial Retrieval



Relevance: 3

Quality: 2

Stance: None



Relevance: 2

Quality: 3

Stance: Pro



Relevance: 1


Quality: 1

Stance: Con



Keep top-10
documents

Initial Retrieval

 **Relevance: 3**
Quality: 2
Stance: None

 **Relevance: 2**
Quality: 3
Stance: Pro

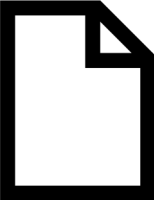


 **Relevance: 1**
Quality: 1
Stance: Con



Re-ranking by Quality


 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 3
Quality: 2
Stance: None



 Relevance: 1
Quality: 1
Stance: Con

Initial Retrieval

 **Relevance: 3**
Quality: 2
Stance: None

 **Relevance: 2**
Quality: 3
Stance: Pro

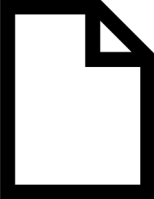


 **Relevance: 1**
Quality: 1
Stance: Con



Re-ranking by Quality

 Relevance: 2
Quality: 3
Stance: Pro


 Relevance: 3
Quality: 2
Stance: None



 Relevance: 1
Quality: 1
Stance: Con


Re-ranking by Quality also improves Relevance

Initial Retrieval

 **Relevance: 3**
Quality: 2
Stance: None

 **Relevance: 2**
Quality: 3
Stance: Pro

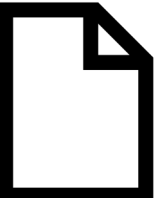


 **Relevance: 1**
Quality: 1
Stance: Con



Re-ranking by Quality

 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 3
Quality: 2
Stance: None



 Relevance: 1
Quality: 1
Stance: Con

Re-ranking by Stance & Quality

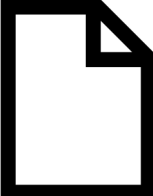
 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 1
Quality: 1
Stance: Con



 Relevance: 3
Quality: 2
Stance: None

Initial Retrieval

 **Relevance: 3**
Quality: 2
Stance: None

 **Relevance: 2**
Quality: 3
Stance: Pro

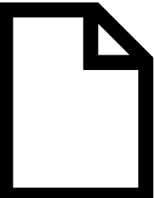


 **Relevance: 1**
Quality: 1
Stance: Con




Re-ranking by Quality

 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 3
Quality: 2
Stance: None



 Relevance: 1

None means no arguments, i.e. poor Quality

Re-ranking by Stance & Quality

 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 1
Quality: 1
Stance: Con



 Relevance: 3
Quality: 2
Stance: None

Initial Retrieval

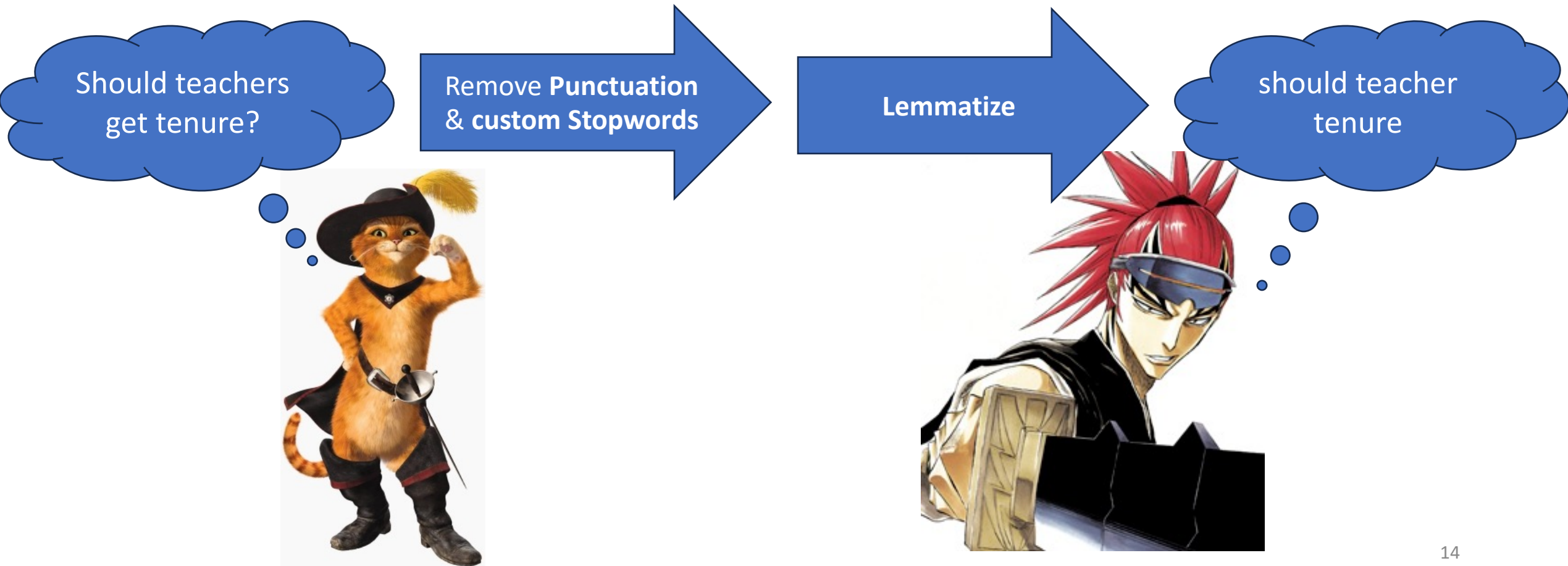
- BM25F-based search engine ChatNoir

Should teachers
get tenure?



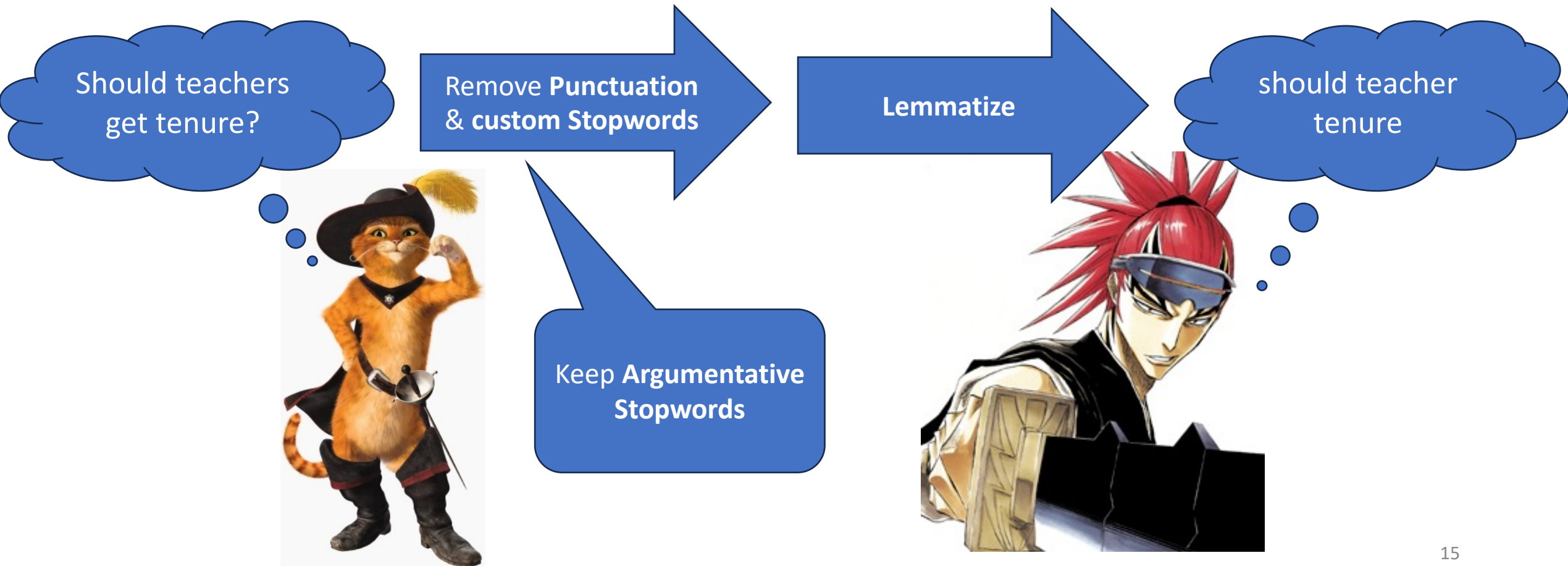
Initial Retrieval

- BM25F-based search engine ChatNoir



Initial Retrieval

- BM25F-based search engine ChatNoir



How to predict Argument Quality?



Manual Features

How to predict Argument Quality?



Manual Features

Neural Embeddings

How to predict Argument Quality?



Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length



Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length



- Punctuation
- Numerics
- External links
- Uppercase
- ...

Occurrences



Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length



- Punctuation
- Numerics
- External links
- Uppercase
- ...

Occurrences



- Academic
- Profanity
- Vocabulary richness
- ...

Word lists



Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length



- Punctuation
- Numerics
- External links
- Uppercase
- ...

Occurrences



- Academic
- Profanity
- Vocabulary richness
- ...

Word lists



- Number of arguments
- Subjectivity
- Sentiment
- Readability
- ...

Complex



Quality prediction: Manual Features

- Document
- Paragraphs
- Sentences
- Words

Length



- Punctuation
- Numerics
- External links
- Uppercase
- ...

Occurrences



- Academic
- Profanity
- Vocabulary richness
- ...

Word lists



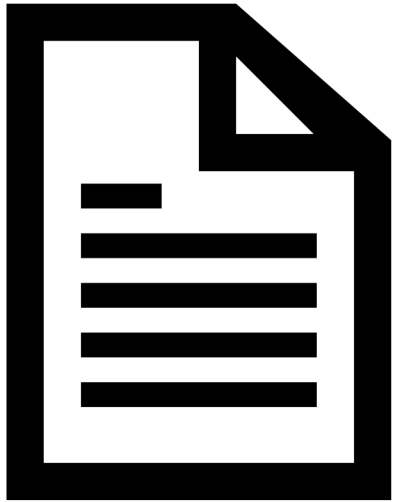
- Number of arguments
- Subjectivity
- Sentiment
- Readability
- ...

Complex



- Total of 32 features

Quality prediction: Automatic Features

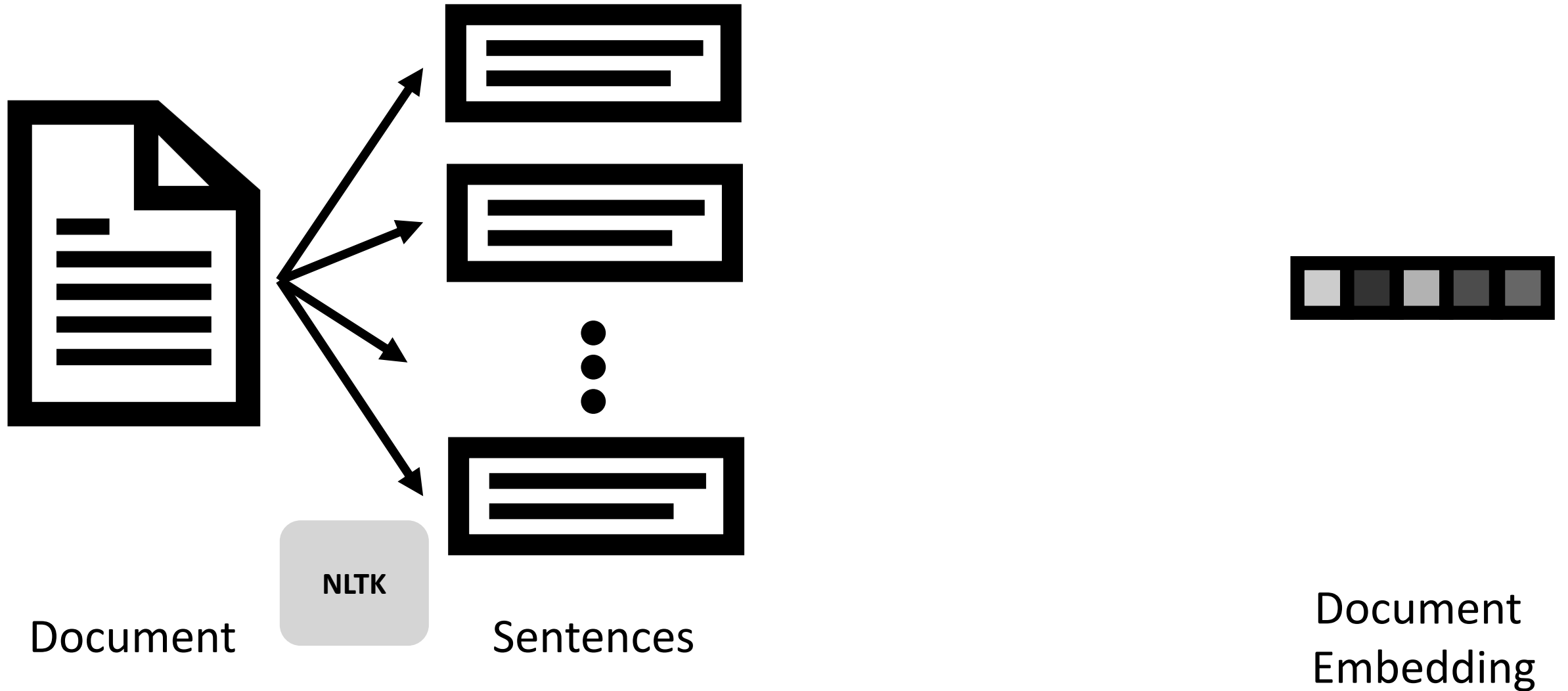


Document

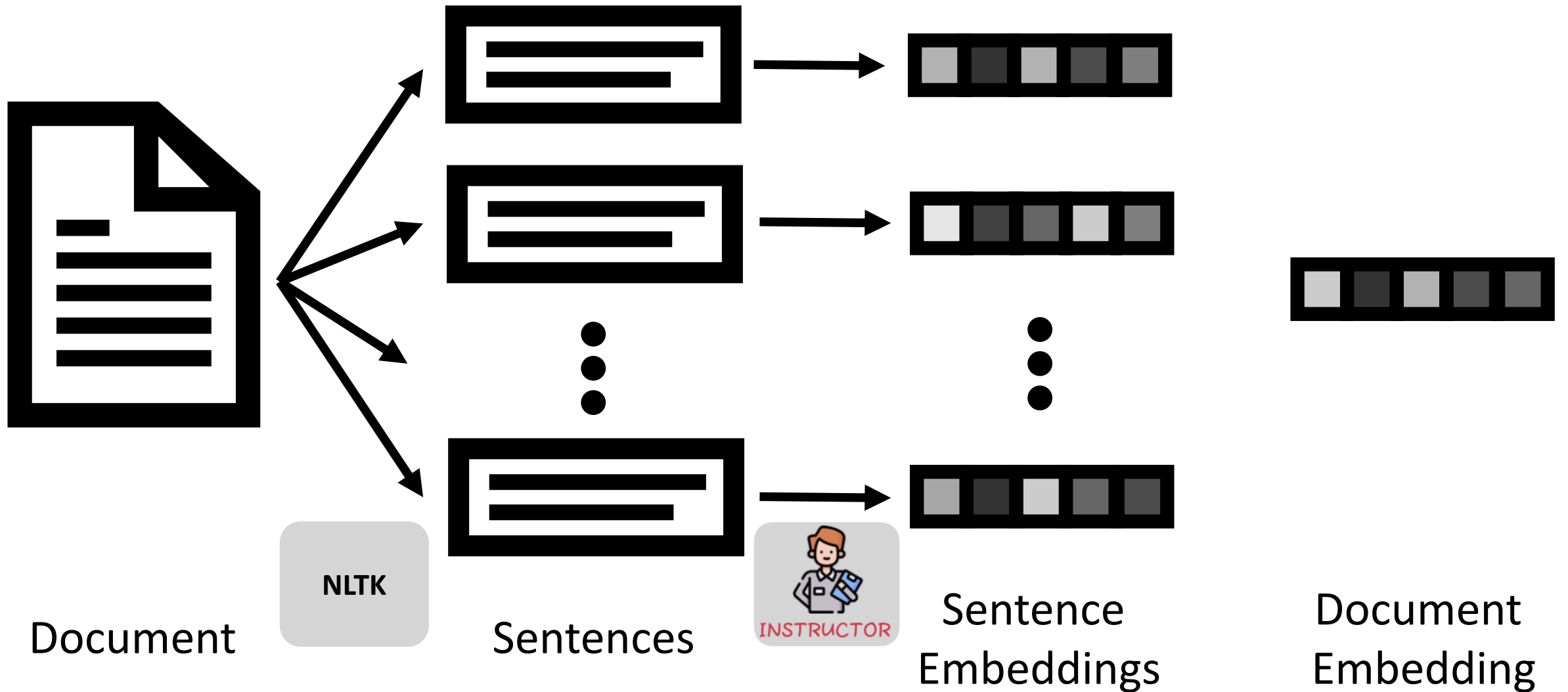


Document
Embedding

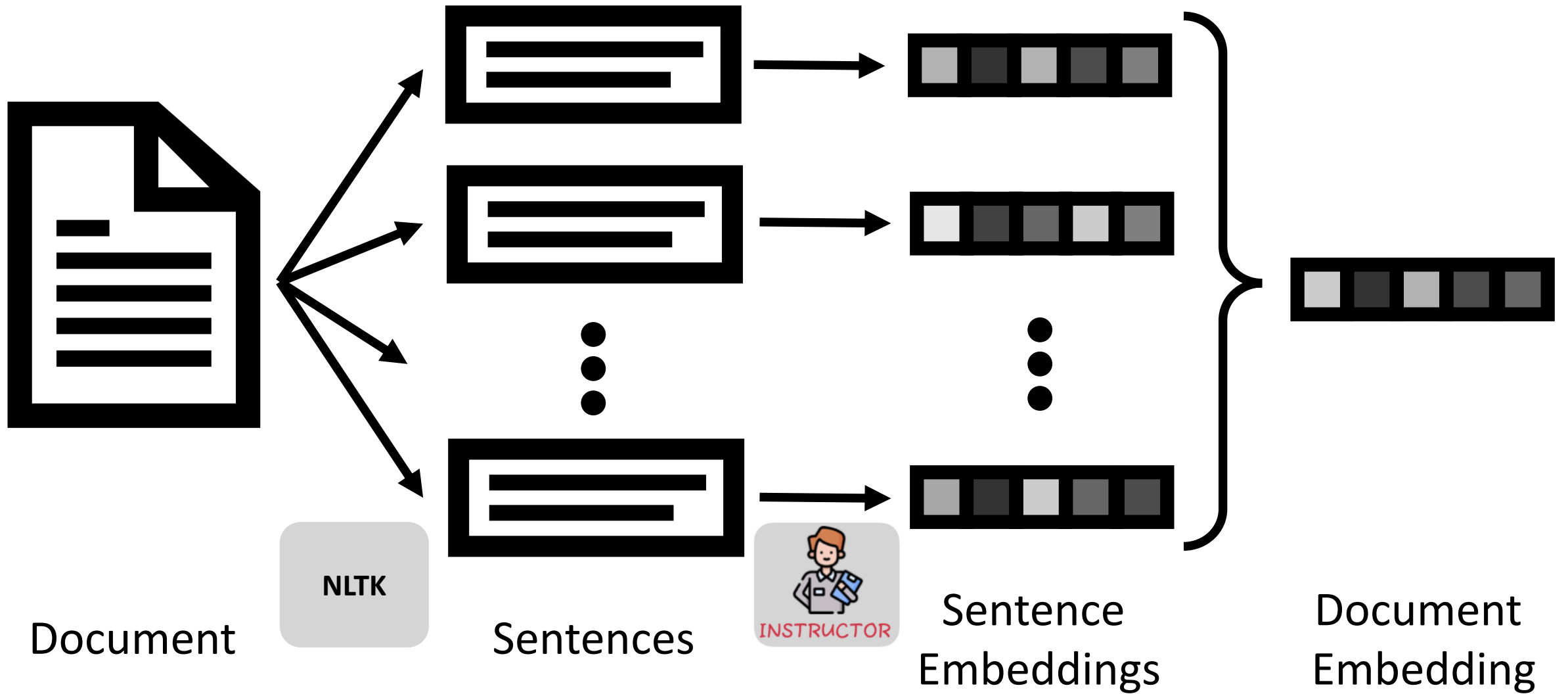
Quality prediction: Automatic Features



Quality prediction: Automatic Features



Quality prediction: Automatic Features

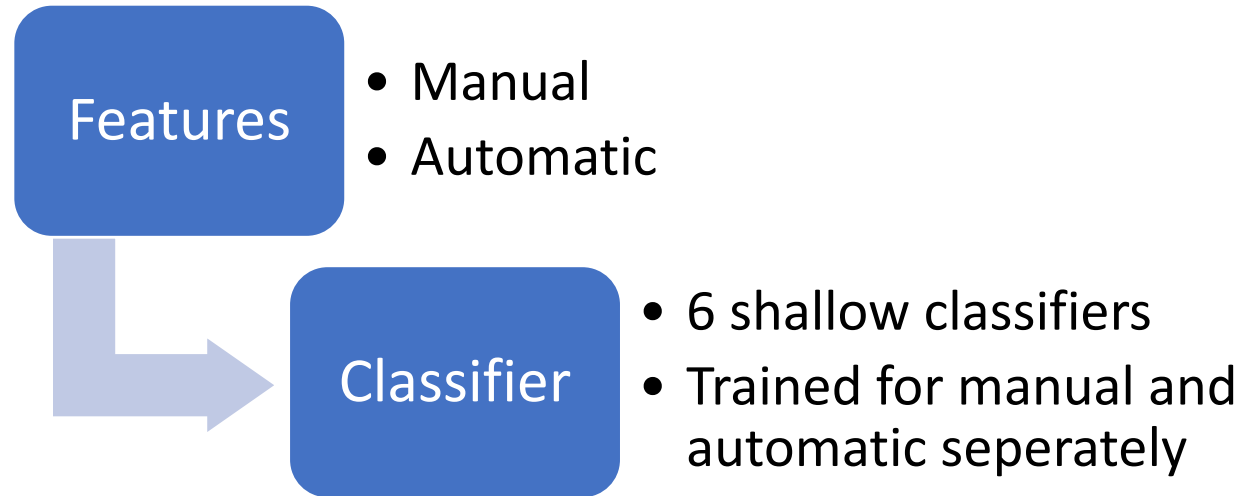


Quality prediction: Classifier

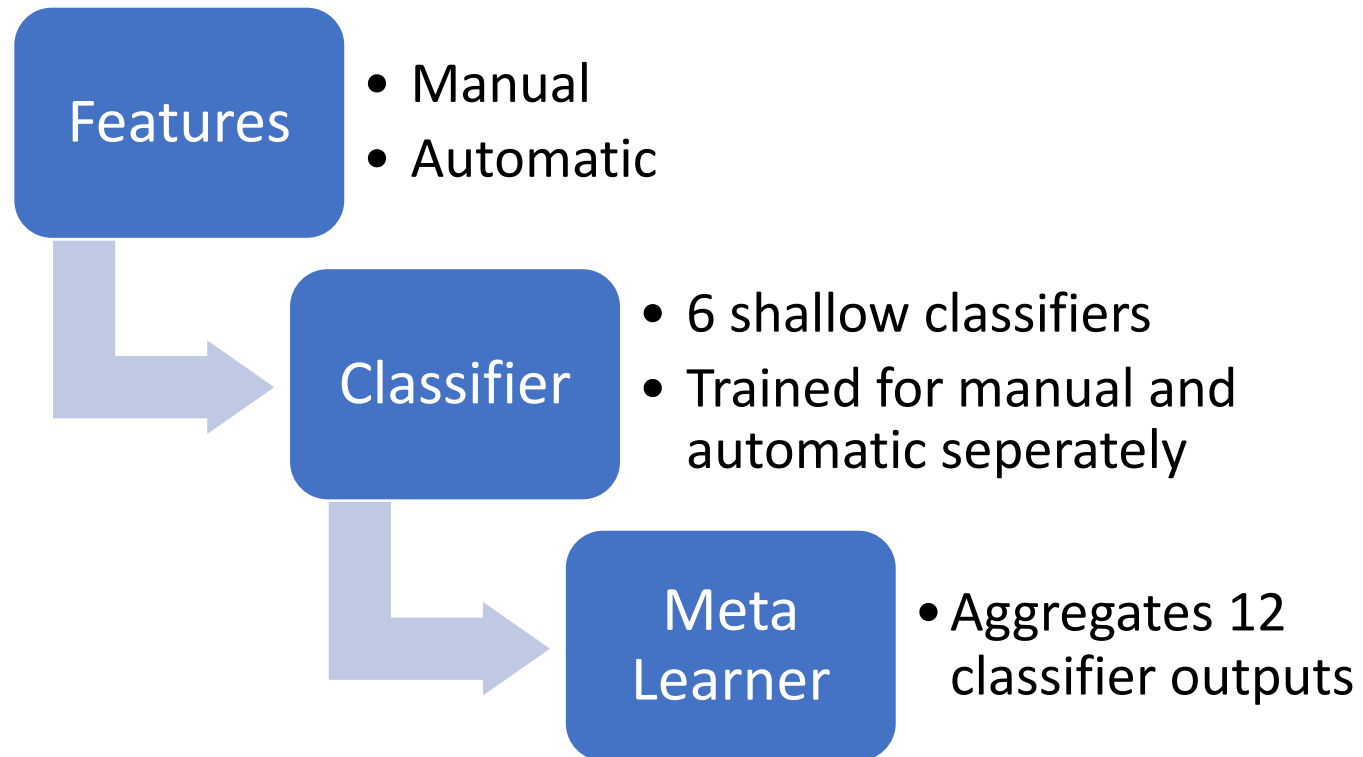
Features

- Manual
- Automatic

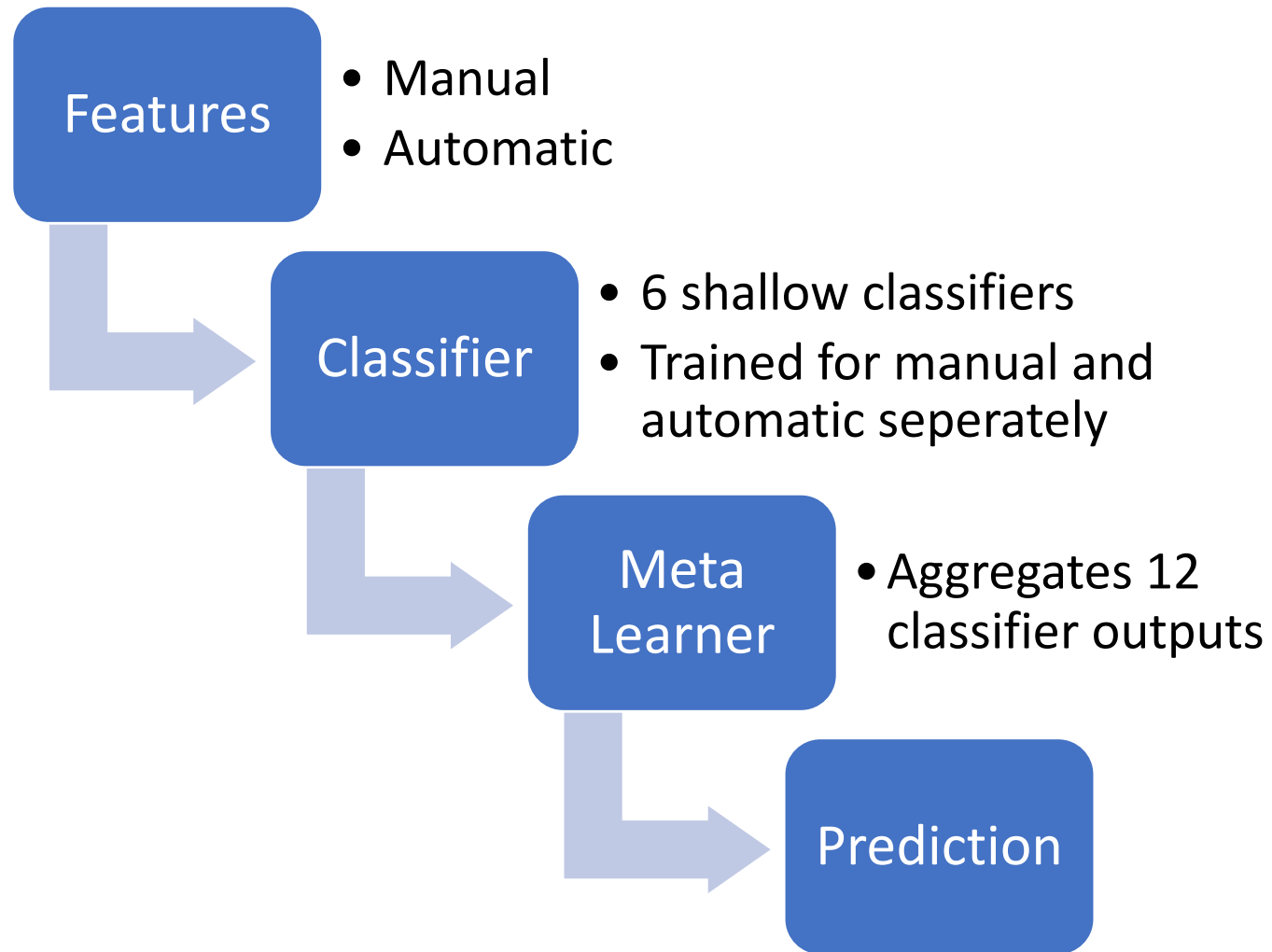
Quality prediction: Classifier



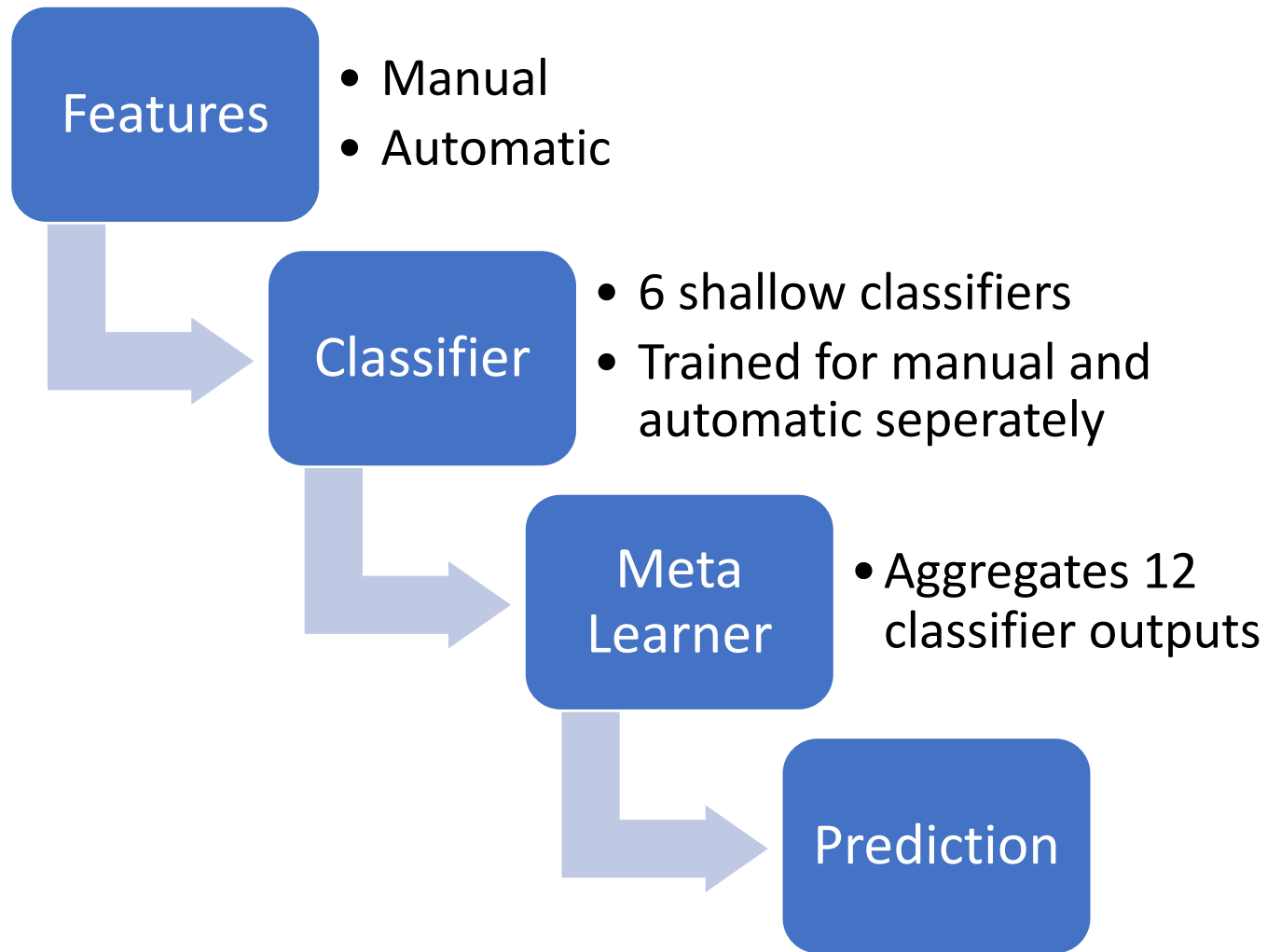
Quality prediction: Classifier



Quality prediction: Classifier




Quality prediction: Classifier







- Data from Touché 2021 Task 1
- 3 Quality Labels: *low, medium and high*
- Cross-topic split




Quality prediction: ChatGPT


 Your task is to predict in a given text the rhetorical argument quality, i.e. "well-writtnenes":
(1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc.
You should return one of the three labels: "high", "medium" and "low".




Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.

 high  

 [More examples...]

 [More correct labels...]  

 [Document]

 [Predicted label]  

Quality prediction: ChatGPT

The screenshot shows a chat interface with a dark background. At the top, a user message (indicated by a small profile picture icon) contains the following text: "Your task is to predict in a given text the rhetorical argument quality, i.e. 'well-writtnenes': (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc. You should return one of the three labels: 'high', 'medium' and 'low'." Below this, a text sample is provided: "Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society." The ChatGPT response (indicated by the OpenAI logo) is the word "high". Below the response are several placeholder messages: "[More examples...]", "[More correct labels...]", "[Document]", and "[Predicted label]". Each message has a thumbs-up and thumbs-down icon to its right.

Instruction

Quality prediction: ChatGPT

Few-shot
(9 examples)

Instruction



Your task is to predict in a given text the rhetorical argument quality, i.e. "well-writtnenes":
(1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc.
You should return one of the three labels: "high", "medium" and "low".

Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.



high



[More examples...]



[More correct labels...]



[Document]



[Predicted label]



Quality prediction: ChatGPT

Few-shot
(9 examples)

The screenshot shows a chat interface with a dark background. At the top, a user message (indicated by a small profile picture) contains the following text: "Your task is to predict in a given text the rhetorical argument quality, i.e. 'well-writtnenes': (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc. You should return one of the three labels: 'high', 'medium' and 'low'." Below this is a document text: "Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society." The next message is from ChatGPT (indicated by the OpenAI logo) and says "high". Below this are three placeholder messages: "[More examples...]", "[More correct labels...]", and "[Document]". The final message is from ChatGPT and says "[Predicted label]".

Instruction

Document

Quality prediction: ChatGPT

Few-shot
(9 examples)

Prediction

The screenshot shows a chat interface with a dark background. At the top, a user message (indicated by a profile picture icon) contains the following text: "Your task is to predict in a given text the rhetorical argument quality, i.e. 'well-writtnenes': (1) whether the document contains arguments and whether the argument text has a good style of speech, (2) whether the text has a proper sentence structure and is easy to follow, (3) whether it includes profanity, has typos, etc. You should return one of the three labels: 'high', 'medium' and 'low'." Below this is a block of text representing the document: "Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society." The next message is from the AI (indicated by the OpenAI logo), which has responded with the word "high". Below this are three more messages from the user, each with a placeholder icon and text: "[More examples...]", "[More correct labels...]", and "[Document]". The final message is from the AI, which has responded with "[Predicted label]".

Instruction

Document

Stance prediction: ChatGPT

Given a query, your task is to predict the stance of a given text. You can give one of the following four labels:

- pro: The text provides overall strong pro argumentation towards the topic in the query.
- con: The text provides overall strong con argumentation towards the topic in the query.
- neutral: The text contains both pro and con arguments, such that overall the stance can be considered as neutral.
- none: The text does not contain arguments or opinions towards the topic in the query, does not take the stance, and mostly contains factual information.

You should return one of the four labels: "pro", "con", "neutral" and "none".

Query: Do animals have rights?
Text: Animal testing and the subjugation of animals undermines a fundamental scientific reality; that humans and animals are kin. With humans and Chimpanzees sharing 99.4% of their genetic code, and humans and mice sharing 99% of their genetic code, it is important to recognize that humans are, on a scientific basis, the kin of animals. The testing of animals undermines this scientific understanding by subjugating animals. This is harmful to broader scientific progression in society.

pro

[More examples...]

[More correct labels...]

Query: [Query]
Text: [Document]

[Predicted label]

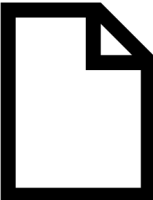
Instruction

Few-shot
(4 examples)

Prediction


Document
& Query

Initial Retrieval

 **Relevance: 3**
Quality: 2
Stance: None

 **Relevance: 2**
Quality: 3
Stance: Pro

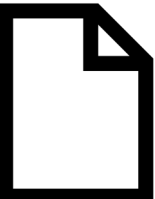


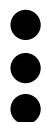
 **Relevance: 1**
Quality: 1
Stance: Con



Re-ranking by Quality

 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 3
Quality: 2
Stance: None



 Relevance: 1
Quality: 1
Stance: Con

Re-ranking by Stance & Quality

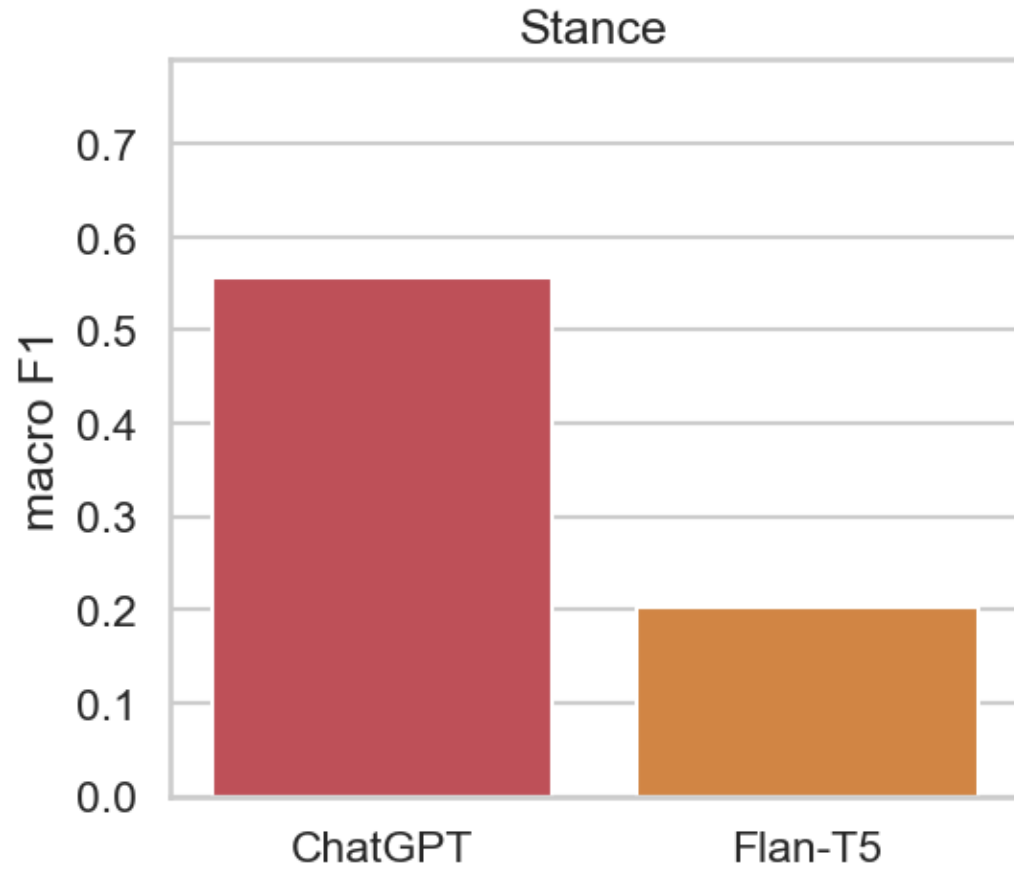
 Relevance: 2
Quality: 3
Stance: Pro

 Relevance: 1
Quality: 1
Stance: Con

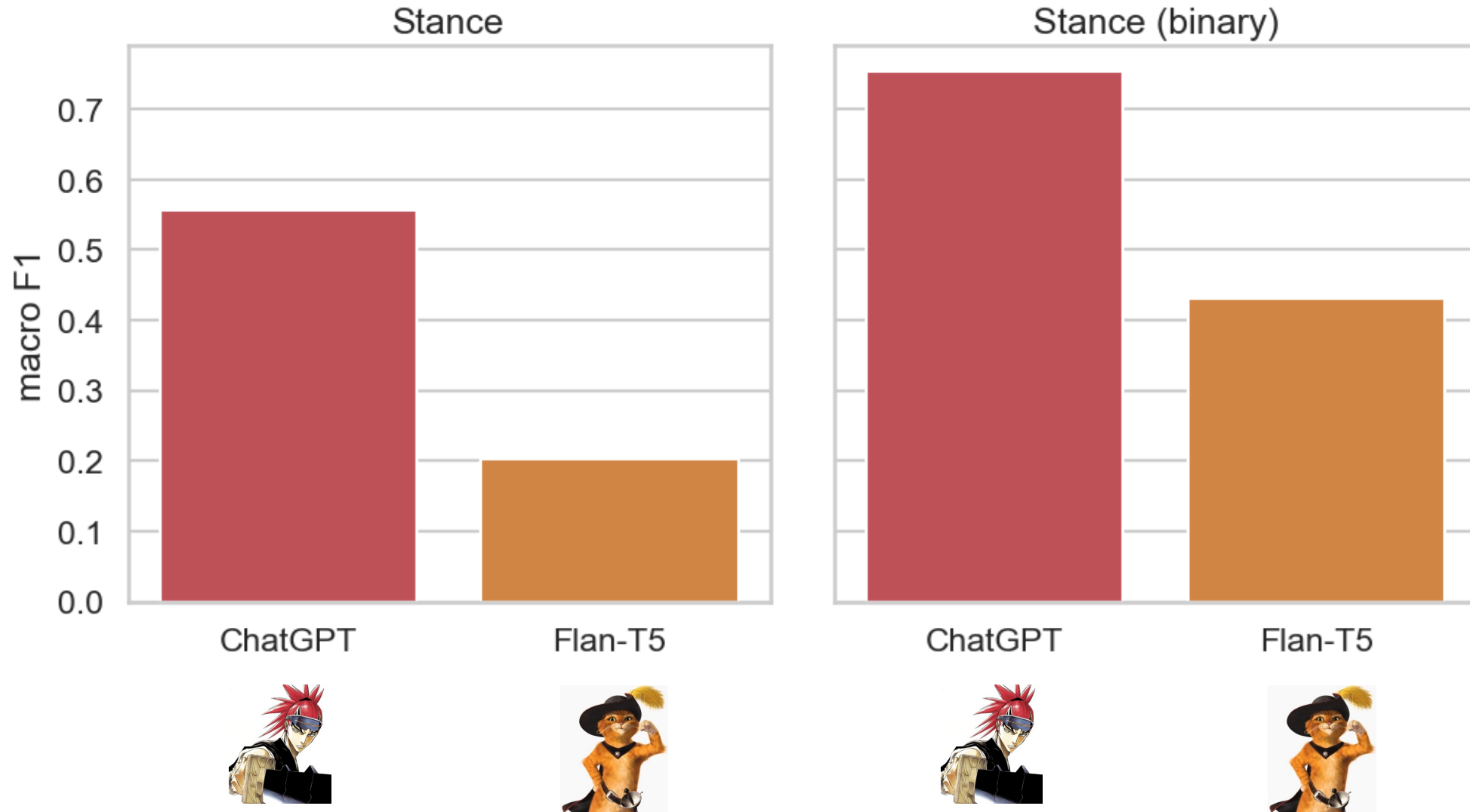


 Relevance: 3
Quality: 2
Stance: None

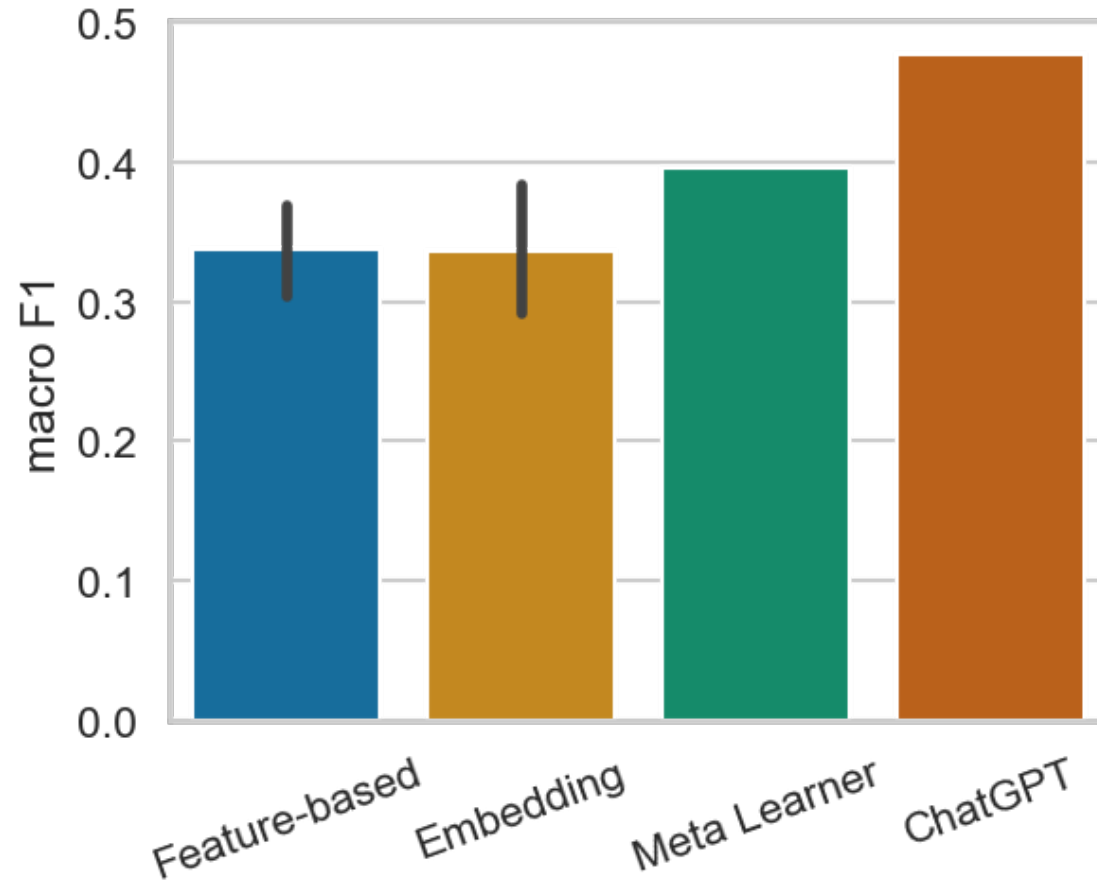
Results – Stance



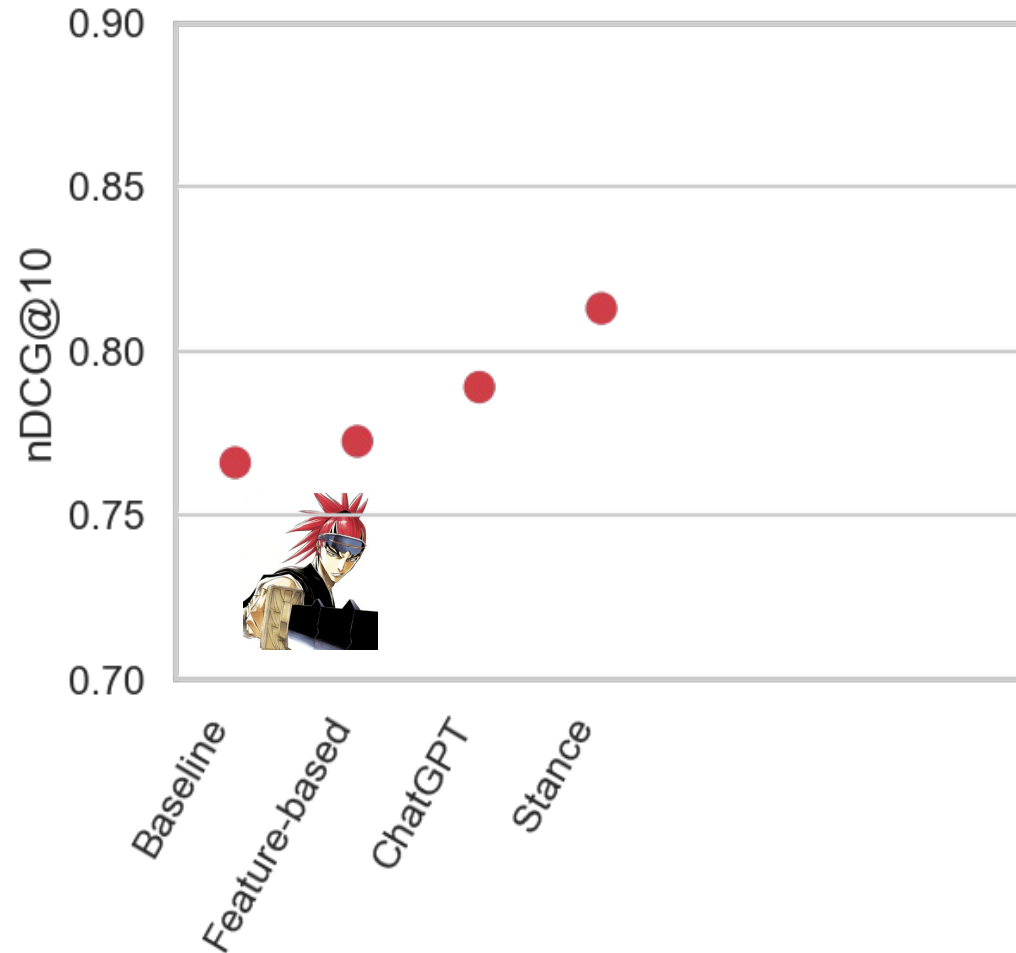
Results – Stance



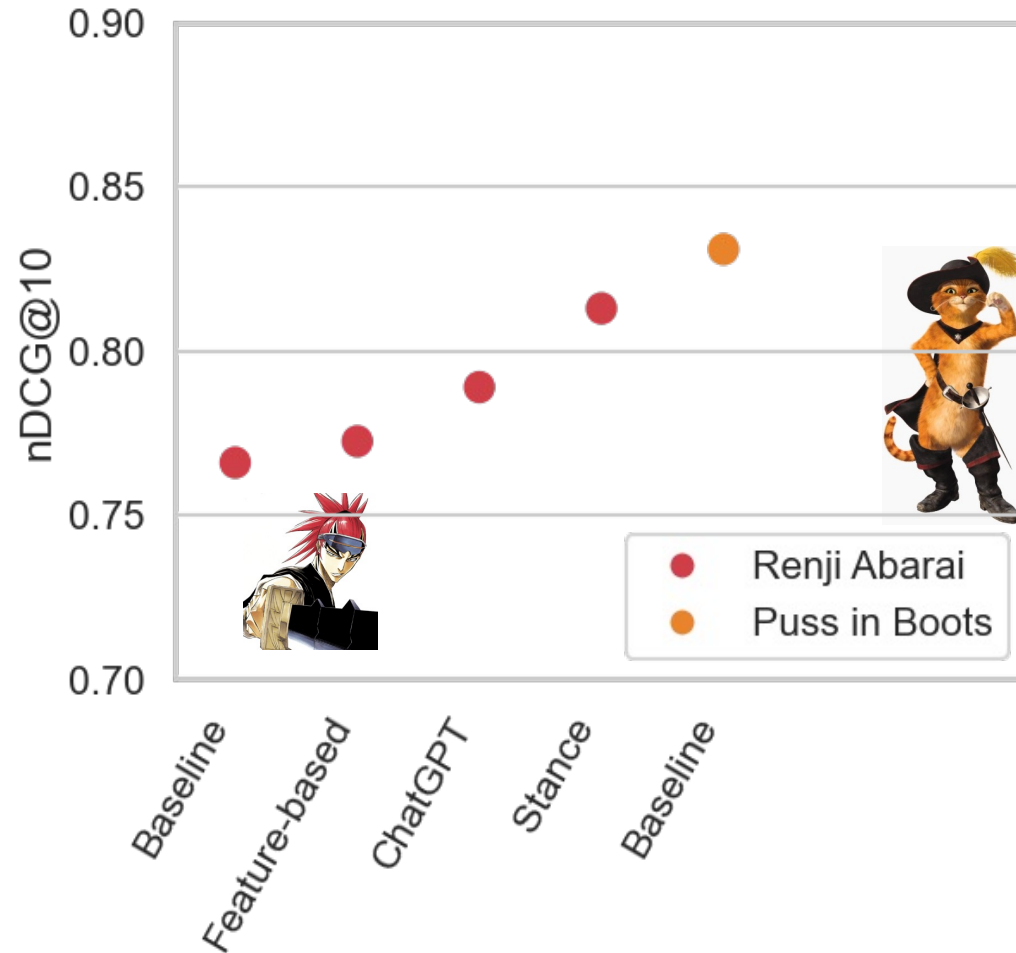
Results – Quality Classification



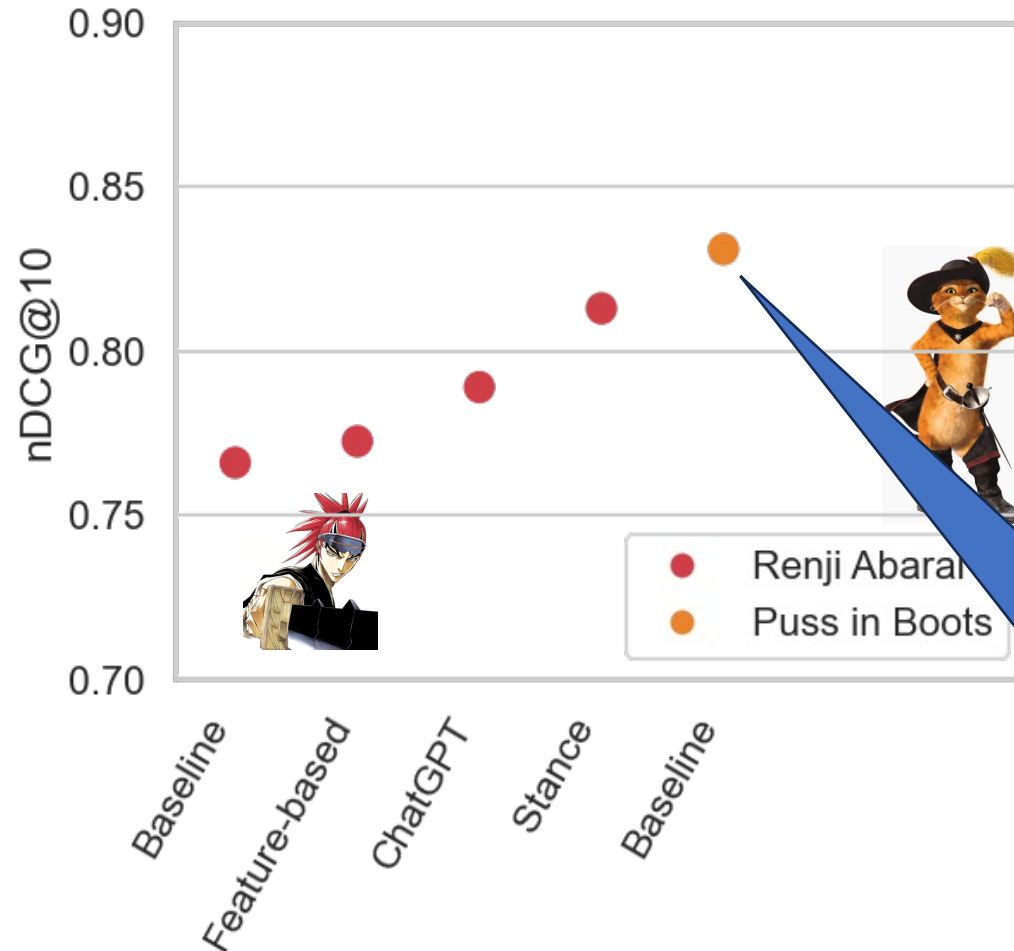
Results – Quality Ranking



Results – Quality Ranking

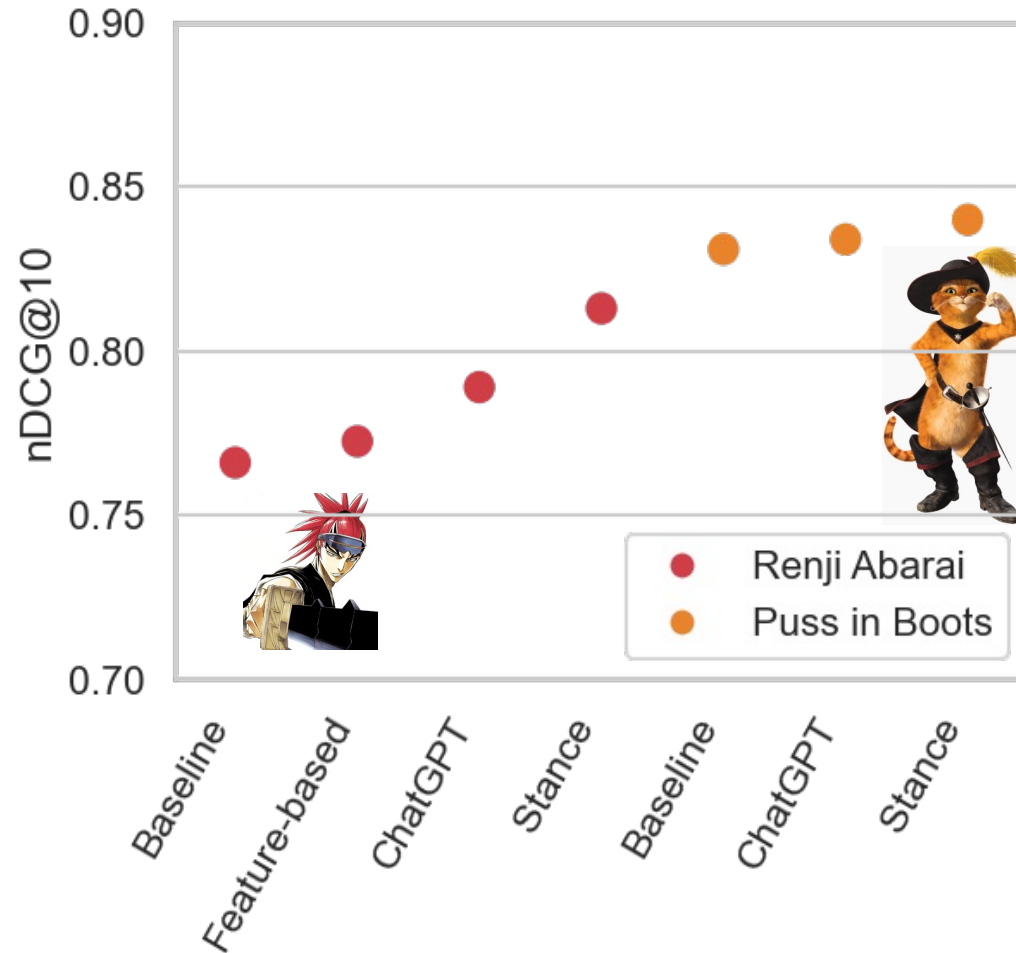


Results – Quality Ranking

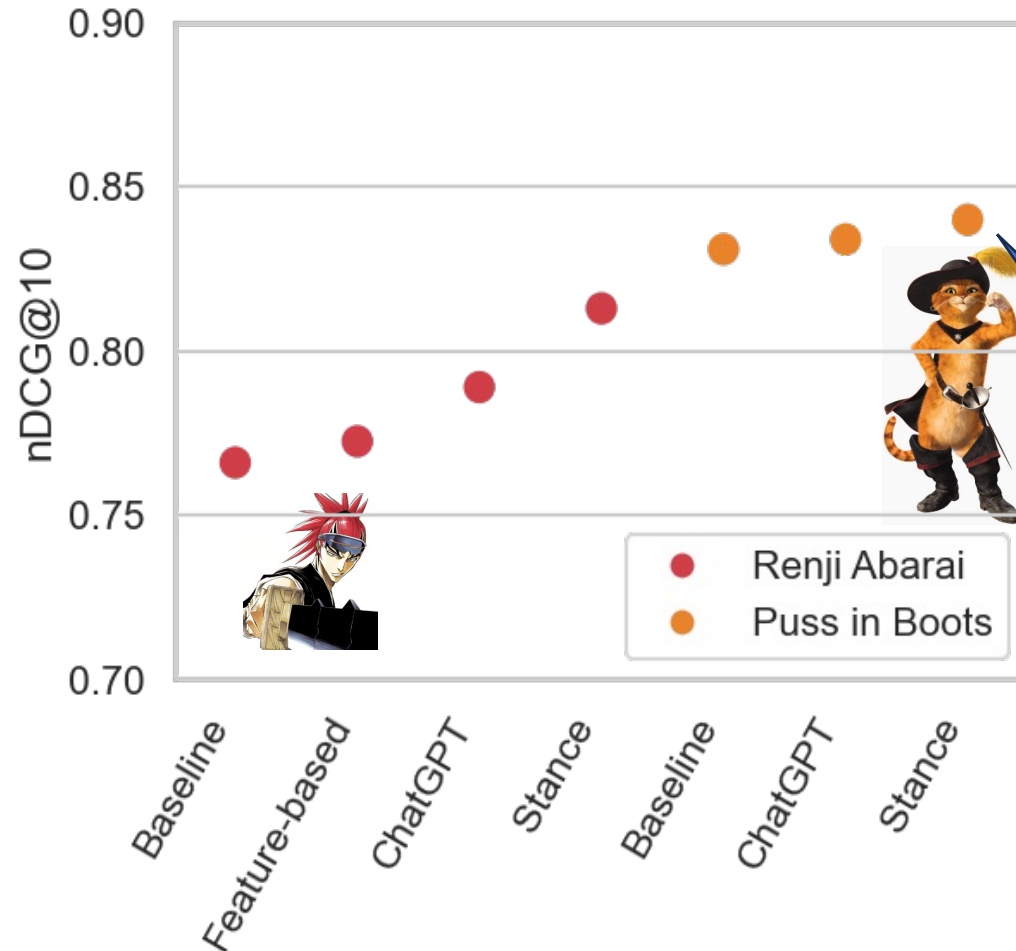


Task baseline
outperforms all
submissions

Results – Quality Ranking

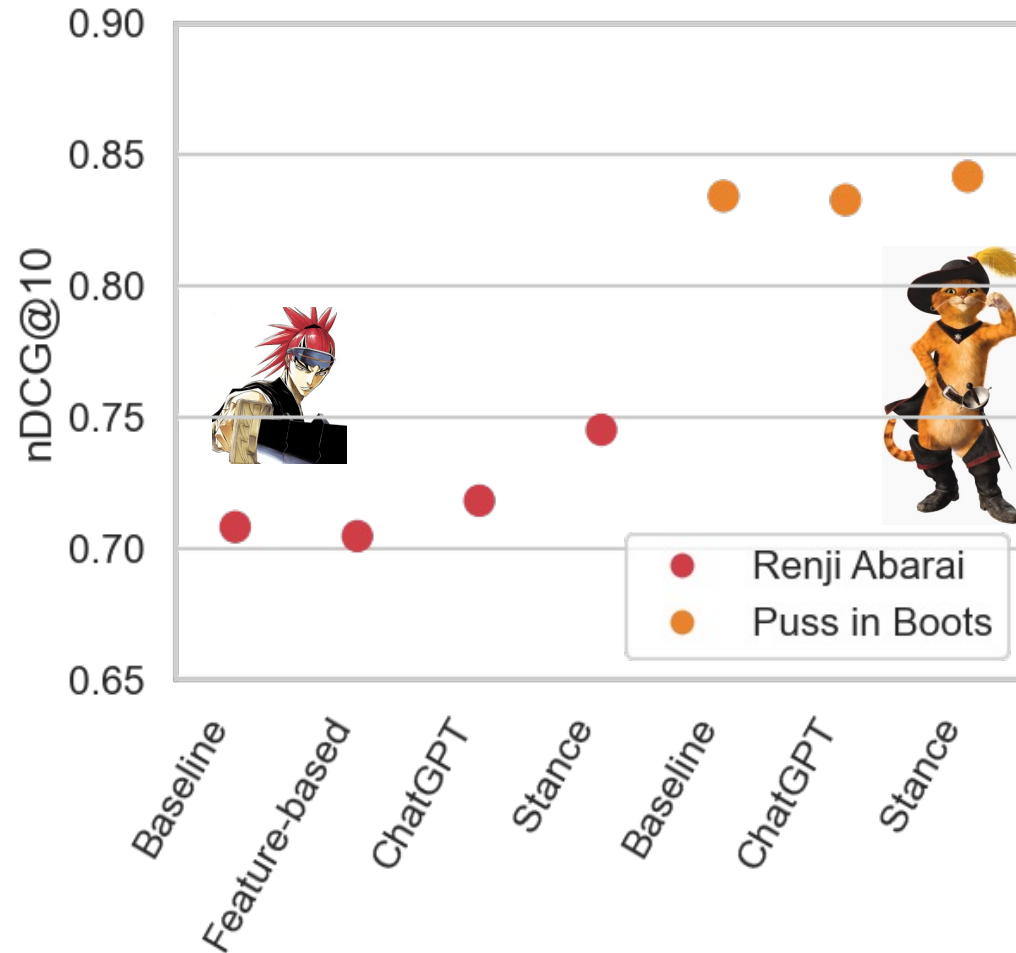


Results – Quality Ranking

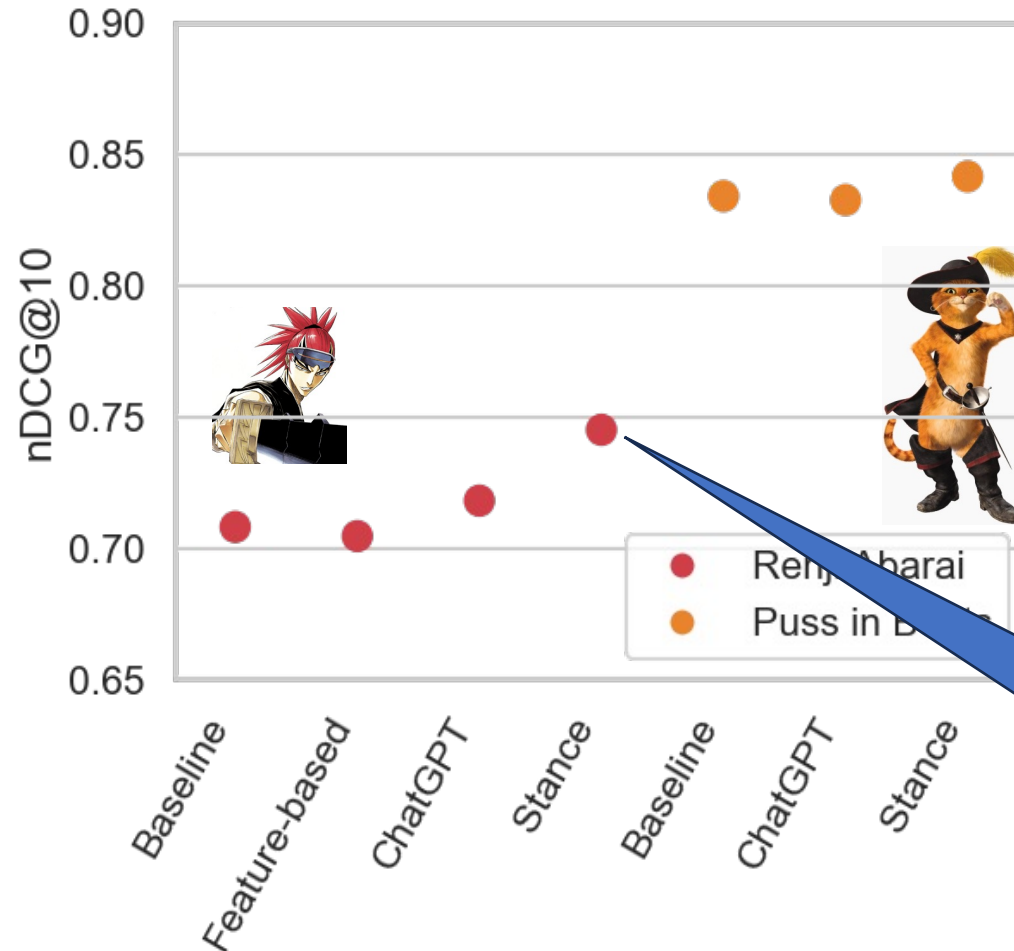


Re-ranking
improves task
baseline

Results – Relevance Ranking



Results – Relevance Ranking



Most improvement
from **Stance**

Conclusion

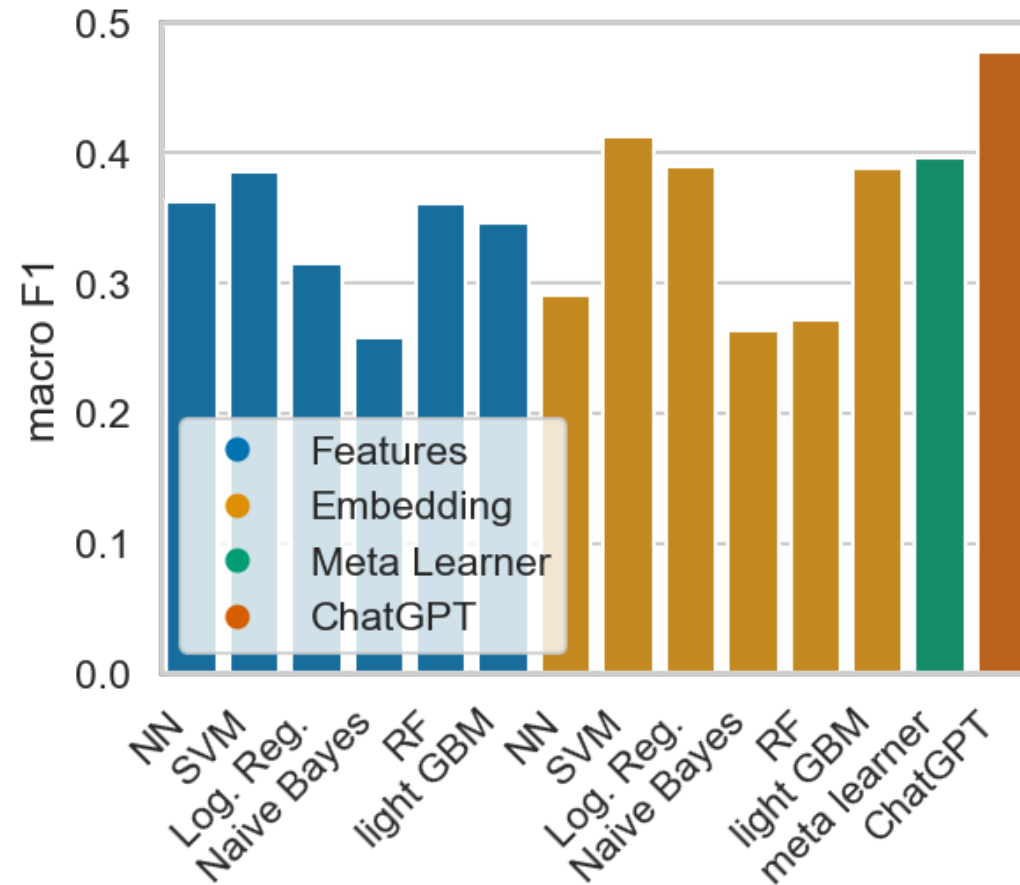
- **BM25F** is a strong baseline
- **Quality** consistently improves with re-ranking
- **Relevance** improves mostly from considering the stance in re-ranking



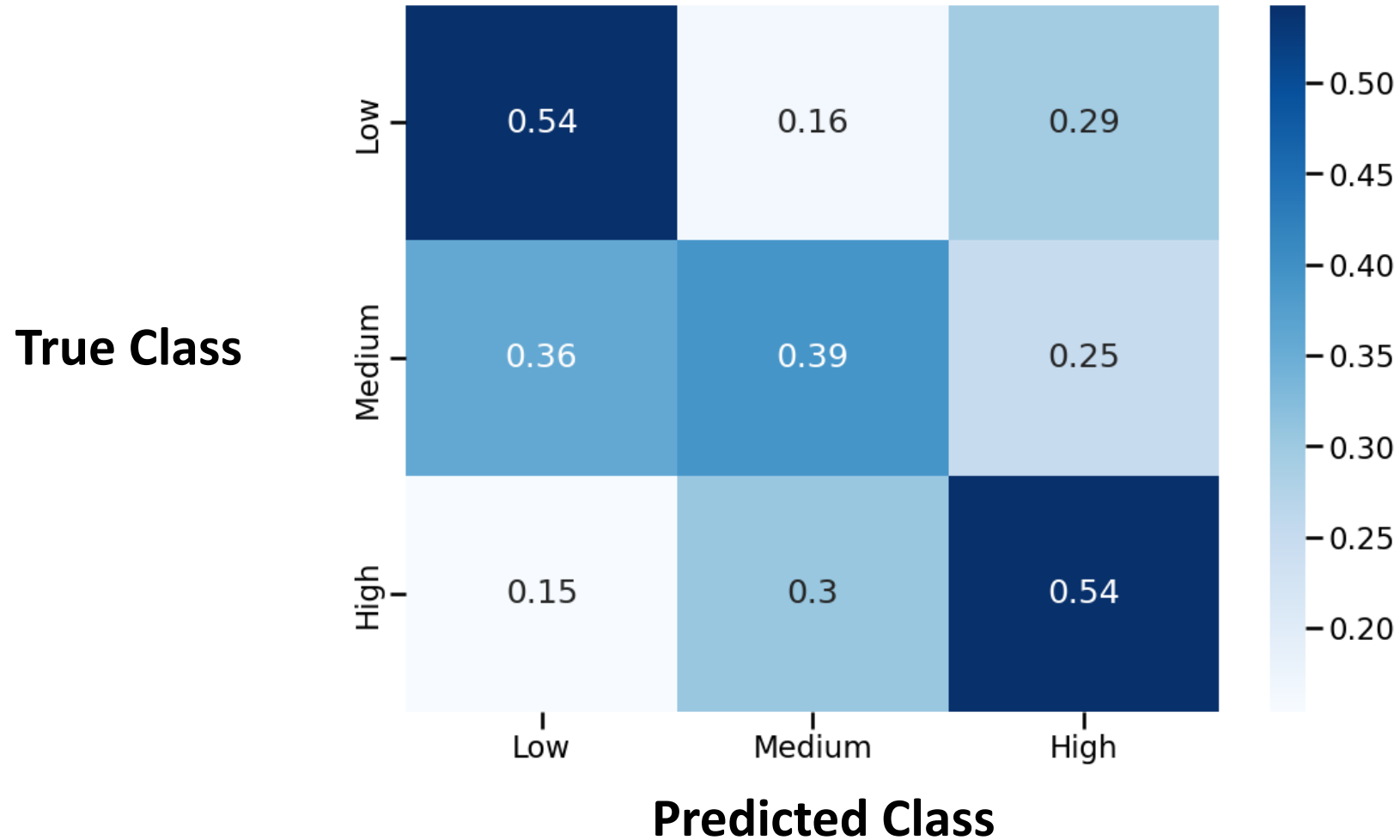
Appendix – Table

		nDCG@10		macro F1	
Configuration (run)		Quality	Relevance	Stance	Stance bin.
Ours w/ BL	stance_ChatGPT	0.840	0.841	0.556	0.754
	stance-certainNO_ChatGPT	0.840	0.842	0.557 ^(*)	0.762 ^(*)
	ChatGPT_mmGhl	0.834	0.833	0.556	0.754
	ChatGPT_mmEQhl	0.834	0.832	0.556	0.754
BL	ChatNoir [9] / Flan-T5 (stance) [38]	0.831	0.834	0.203	0.432
Ours	stance_ChatGPT	0.815 [†]	0.744	0.599	0.780
	stance-certainNO_ChatGPT	0.811 [†]	0.746	0.604 ^(*)	0.783 ^(*)
	ChatGPT_mmGhl	0.789	0.718	0.599	0.780
	ChatGPT_mmEQhl	0.789	0.718	0.599	0.780
	meta_qual_prob	0.774	0.697	0.599	0.780
	meta_qual_score	0.771	0.712	0.599	0.780
	baseline	0.766	0.708	0.599	0.780

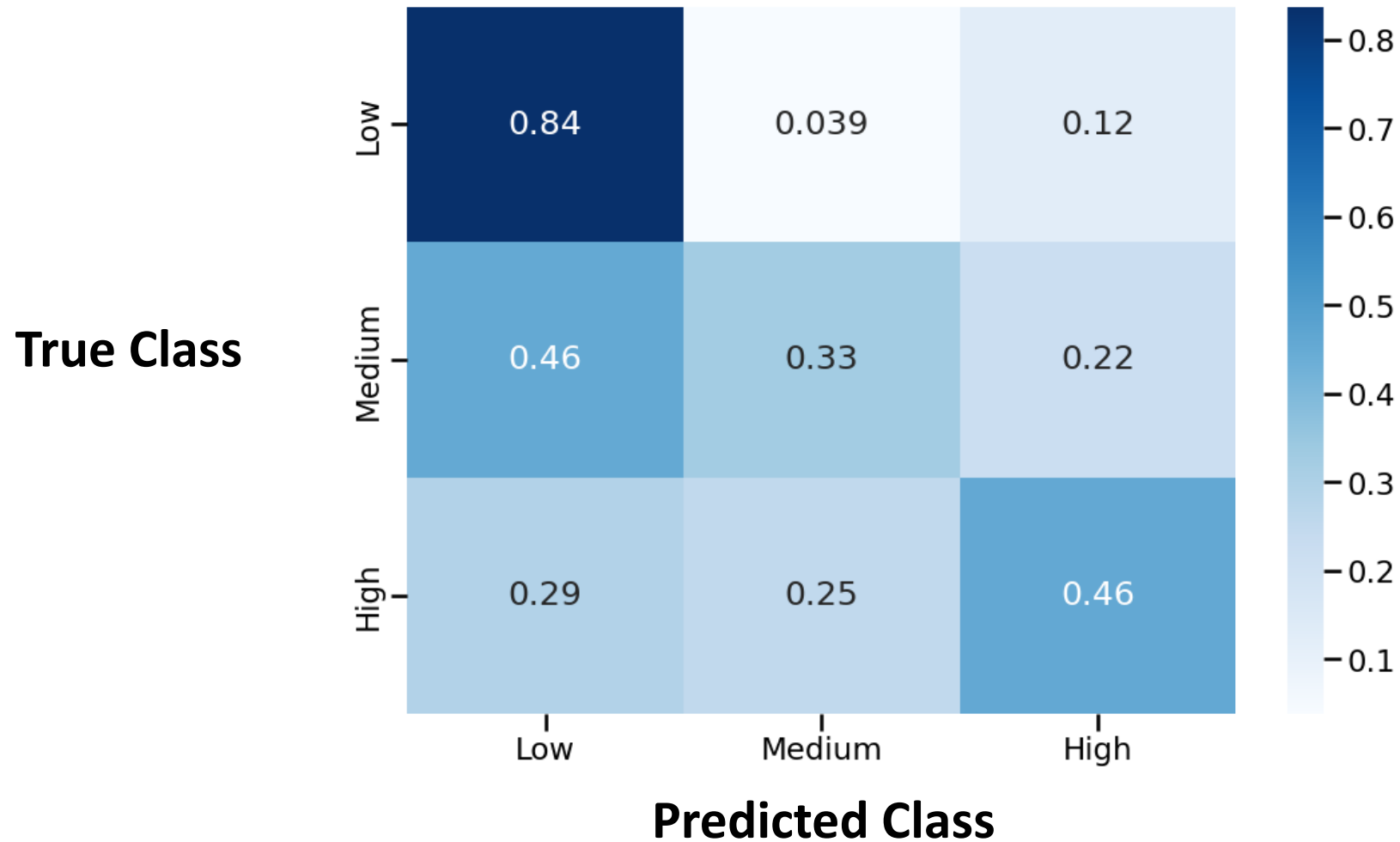
Appendix – Quality Classification Detailed



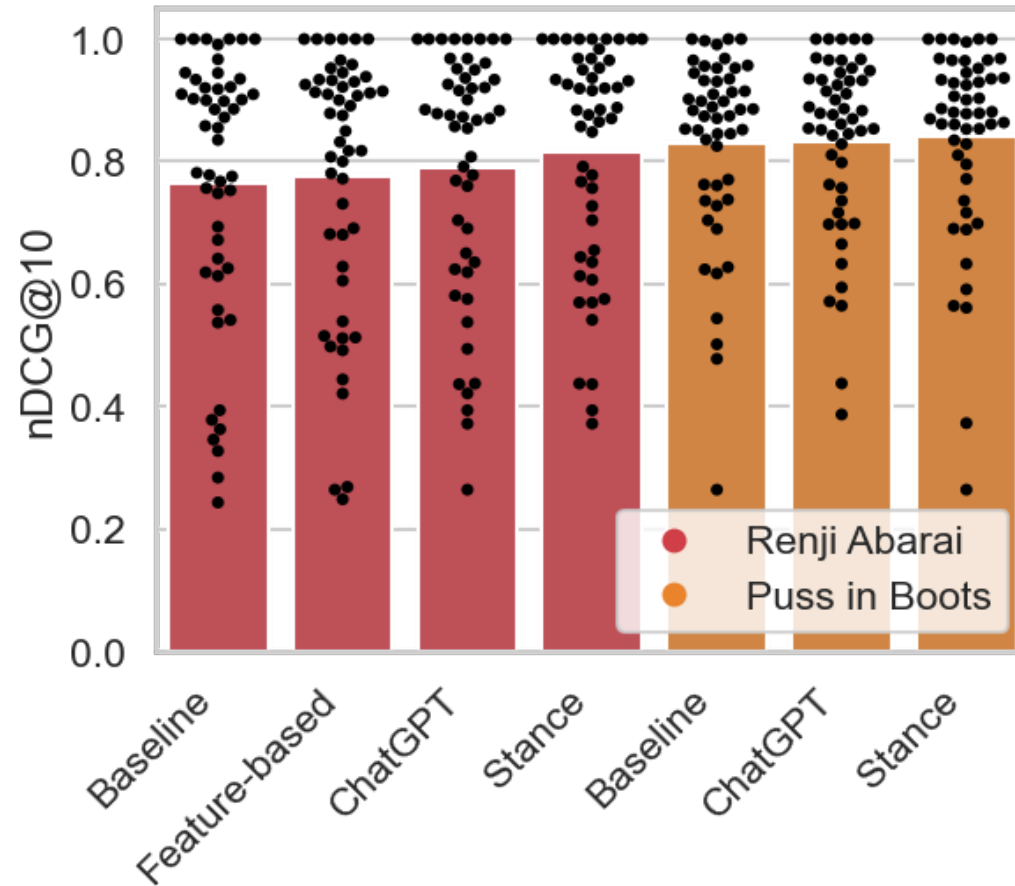
Appendix – Quality Classification ChatGPT



Appendix – Quality Classification with Stance ChatGPT



Appendix – Quality Ranking Detailed



Appendix – Relevance Ranking Detailed

