A Pilot Study on Annotation Interfaces for Summary Comparisons

Sian Gooding **Lucas Werner** **Victor Cărbune**
Google Research
{sgooding|lucaswerner|vcarbune}@google.com

Abstract

The task of summarisation is notoriously difficult to evaluate, with agreement even between expert raters unlikely to be perfect. One technique for summary evaluation relies on collecting comparison data by presenting annotators with generated summaries and tasking them with selecting the best one. This paradigm is currently being exploited in reinforcement learning using human feedback, whereby a reward function is trained using pairwise choice data. Comparisons are an easier way to elicit human feedback for summarisation, however, such decisions can be bottle necked by the usability of the annotator interface. In this paper, we present the results of a pilot study exploring how the user interface impacts annotator agreement when judging summary quality.

1 Introduction

As language models become more powerful, training and evaluation are increasingly limited by the data and metrics used for a particular task (Stiennon et al., 2020). Human evaluation has traditionally been used in the field of summarisation as a gold standard when assessing the quality of model outputs and for corroborating automated evaluation techniques. However, ensuring high quality evaluation with human annotators is difficult due to the subjective and task-dependent paradigm of summarisation. As model refinement will increasingly rely on human feedback it is important to consider how to best elicit high quality signal from human annotators.

One technique to judge the quality of summaries is the use of human preferences via comparison or ranking. Such rankings can also be used to improve summary quality by training models in a reinforcement learning paradigm. For instance, Stiennon et al. (2020) show that human preference data can be used to improve the capability of large language models (LLMs) for summarisation via a technique referred to as Reinforcement Learning from Human Feedback (RLHF). However, human feedback does not always provide a gold standard for summarisation when the task is not clearly defined. It has been established that linguistically trained, expert raters, provide the gold standard in summarisation evaluation and the reliability of non-experts has been repeatedly questioned (Lloret et al., 2018). For instance, it has been found that crowd workers should not be used to evaluate summary quality because of a non-correlation with experts (Gillick and Liu, 2010; Fabbri et al., 2021). Furthermore, even for expert annotations mediation meetings are necessary to assure reliability (Iskender et al., 2021). In short, evaluating the quality of a summary is not an easy or straightforward task.

The use of RLHF to train LLMs is becoming increasingly common. However, training a model from human feedback relies on the collection of data via user interfaces for the chosen task. Increasingly then, natural language processing applications are heavily influenced by the human computer interaction that takes place when collecting preference data. Recent work in RLHF for summarisation overlooks how critical the user interface is in this process, with little to no discussion of the design decisions made.

In this paper, we present the findings from a pilot study introducing a novel user interface for summary comparisons. We document how the introduction of the new interface impacts annotator engagement as well as investigate the following research questions:

- **RQ1:** Does the annotator agreement for the task of summary comparison change based on the user-interface and task conceptualisation?
- **RQ2:** Does allowing the highlighting of tokens improve the agreement of summary comparisons?

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 179–187
July 13, 2023 ©2023 Association for Computational Linguistics

2 Background

It is widely understood that machine learning systems are limited by the quality of the labelled training data (Gooding et al., 2019). One approach to improving the performance of such systems is to treat the human labeller(s) as a source of noise (Frénay and Verleysen, 2014) who can be modelled statistically (Yan et al., 2010) in order to more accurately identify an underlying ground truth. Noise estimation can be improved if multiple labels are obtained for each item in the training set in order to model inconsistency (Ipeirotis et al., 2014), or if a distribution of label values can be used as a basis for rejecting outliers (Brodley and Friedl, 1999). More recent approaches have relied on probabilistic methods for training deep classifiers under input-dependent label noise (Collier et al., 2021).

However, these approaches focus on dealing with noise post-annotation, whereas it is known that the quality and clarity of the user interface itself, as well as the task formulation has large implications for the annotator agreement. For instance, several of the human factors can be addressed through the use of pairwise comparison, where labellers make relative judgments to compare training items, rather that attempting to characterize each item independently against an abstract conceptual category, for which they are expected to have a stable definition and associated membership criteria. In the context of labelling, comparative judgments are used to compare how well the training items correspond to the required concept. Carterette et al. (2008) demonstrate that this method can facilitate judgments for information retrieval applications. Comparative judgments have also been used in gamified labelling (Bennett et al., 2009), where cooperating players reduce the set of alternative items until agreement is reached.

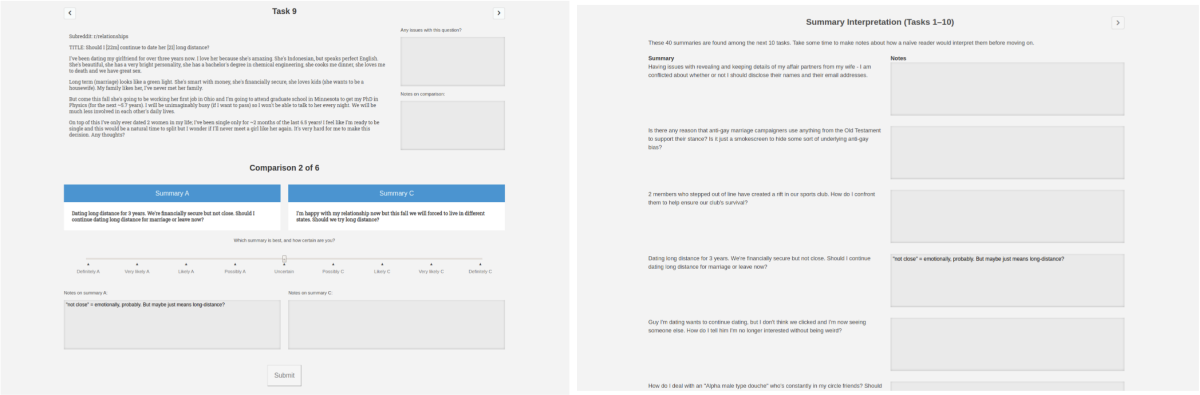
Recent work has looked into the application of comparative judgments to labelling as opposed to assignment of categorical values or scores on a scale (Simpson et al., 2019; Yang and Chen, 2011; Kingsley and Brown, 2010). Simpson et al. (2019) note that comparative judgments are suitable for abstract linguistic properties, whose nature can cause inconsistencies in the assigned numerical scores.

2.1 Agreement in summarisation

Text summarisation is the process of generating short, fluent, and factually accurate summaries of longer documents. As with most natural language generation tasks, evaluation of generated summarisation is difficult, with automated metrics often falling short. Human evaluation on summarisation has been broadly classified into two types: intrinsic and extrinsic (Jones and Galliers, 1995; Belz and Reiter, 2006; Steinberger and Jezek, 2009). In intrinsic evaluation, the summarisation quality is measured based on the resulting summary itself without considering the source. Generally, it has been carried out as a pair comparison task (generated output to expert summaries) or using absolute scales without showing a reference summary. Extrinsic evaluation, also known as task-based evaluation, aims to measure the summary's impact using a task based on the source document (Mani, 2001). (Reiter and Belz, 2009) have argued that the extrinsic evaluation is more useful than intrinsic because the summarization systems are developed to satisfy the information need from the source text in a condensed way, but van der Lee et al. (2021) have reported that only 3% of summarisation papers employ extrinsic evaluation. Extrinsic evaluation is important because it is rooted in the fundamental application of summarisation models. Across papers, guidelines provided to annotators on what constitutes a good summary have a high degree of variation. For instance, Howcroft et al. (2020) found over 200 variations in terminology when analysing annotator guidelines.

2.2 Summary Comparisons for RLHF

Stiennon et al. (2020) show that human preference data can be used to improve summary quality by training the model to optimise for human preferences instead of using coarse proxies like ROUGE. This is achieved via RLHF whereby a large dataset of human preferences between generated summaries is collected, and a reward model trained using this data. The annotations collected are from researchers (experts) and human annotations with the agreement rate between researchers ranging from about 65% on the most difficult comparisons, to approximately 80% on the easiest comparisons (comparing a high-temperature sample from a supervised baseline to the human reference summary). For cases where annotators discussed the comparisons with each other the agreement reached 95%. The paper states that: substantial noise comes from comparisons being quite difficult and subjective. In the entire corpus, labellers agree with each other 72% of the time. Using the modal



Subreddit: r/relationships

TITLE: Should I [22m] continue to date her [21] long distance?

I've been dating my girlfriend for over three years now. I love her because she's amazing. She's Indonesian, but speaks perfect English. She's beautiful, she has a very bright personality, she has a bachelor's degree in chemical engineering, she cooks me dinner, she loves me to death and we have great sex.

Long term (marriage) looks like a green light. She's smart with money, she's financially secure, she loves kids (she wants to be a housewife). My family likes her, I've never met her family.

But come this fall she's going to be working her first job in Ohio and I'm going to attend graduate school in Minnesota to get my PhD in Physics (for the next ~5.7 years). I will be unimaginably busy (if I want to pass) so I won't be able to talk to her every night. We will be much less involved in each other's daily lives.

On top of this I've only ever dated 2 women in my life; I've been single only for ~2 months of the last 6.5 years! I feel like I'm ready to be single and this would be a natural time to split but I wonder if I'll never meet a girl like her again. It's very hard for me to make this decision. Any thoughts?

Summary A: Dating long distance for 3 years. We're financially secure but not close. Should I continue dating long distance for marriage or leave now?

Summary C: I'm happy with my relationship now but this fall we will forced to live in different states. Should we try long distance?

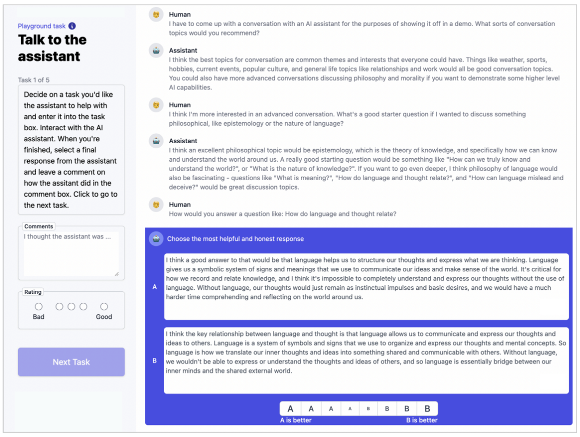
Which summary is best, and how certain are you?

Notes on summary A: "not close" = emotionally, probably. But maybe just means long-distance?

These 40 summaries are found among the next 10 tasks. Take some time to make notes about how a naïve reader would interpret them before moving on.

Summary	Notes
Having issues with revealing and keeping details of my affair partners from my wife - I am conflicted about whether or not I should disclose their names and their email addresses.	
Is there any reason that anti-gay marriage campaigners use anything from the Old Testament to support their stance? Is it just a smokescreen to hide some sort of underlying anti-gay bias?	
2 members who stepped out of line have created a rift in our sports club. How do I confront them to help ensure our club's survival?	
Dating long distance for 3 years. We're financially secure but not close. Should I continue dating long distance for marriage or leave now?	"not close" = emotionally, probably. But maybe just means long-distance?
Guy I'm dating wants to continue dating, but I don't think we clicked and I'm now seeing someone else. How do I tell him I'm no longer interested without being weird?	
How do I deal with an "Alpha male type douche" who's constantly in my circle friends? Should	

Figure 1: Stiennon et al. (2020) user interface to collect preference data from annotators (left) and interface to collect interpretations of summaries on (right)



Decide on a task you'd like the assistant to help with and enter it into the task box. Interact with the AI assistant. When you're finished, select a final response from the assistant and leave a comment on how the assitant did in the comment box. Click to go to the next task.

Human
I have to come up with a conversation with an AI assistant for the purposes of showing it off in a demo. What sorts of conversation topics would you recommend?

Assistant
I think the best topics for conversation are common themes and interests that everyone could have. Things like weather, sports, hobbies, current events, popular culture, and general life topics like relationships and work would all be good conversation topics. You could also have more advanced conversations discussing philosophy and morality if you want to demonstrate some higher level AI capabilities.

Human
I think I'm more interested in an advanced conversation. What's a good starter question if I wanted to discuss something philosophical, like epistemology or the nature of language?

Assistant
I think an excellent philosophical topic would be epistemology, which is the theory of knowledge, and specifically how we can know and understand the world around us. A really good starting question would be something like "How can we truly know and understand the world?", or "What is the nature of knowledge?". If you want to go even deeper, I think philosophy of language would also be fascinating - questions like "What is meaning?", "How do language and thought relate?", and "How can language mislead and deceive?" would be great discussion topics.

Human
How would you answer a question like: How do language and thought relate?

Choose the most helpful and honest response

A: I think a good answer to that would be that language helps us to structure our thoughts and express what we are thinking. Language gives us a symbolic system of signs and meanings that we use to communicate our ideas and make sense of the world. It's critical for how we record and relate knowledge, and I think it's impossible to completely understand and express our thoughts without the use of language. Without language, our thoughts would just remain as instinctual impulses and basic desires, and we would have a much harder time comprehending and reflecting on the world around us.

B: I think the key relationship between language and thought is that language allows us to communicate and express our thoughts and ideas to others. Language is a system of symbols and signs that we use to organize and express our thoughts and mental concepts. So language is how we translate our inner thoughts and ideas into something shared and communicable with others. Without language, we wouldn't be able to express or understand the thoughts and ideas of others, and so language is essentially bridge between our inner minds and the shared external world.

Figure 2: Bai et al. (2022) conversational interface for annotators to select helpful LLM responses

output from 3 labellers can increase this agreement rate from 72% to 77%. However, this is not used as the work prioritises label throughput with summaries receiving on average 1 annotation. Figure 1 shows the interface used by annotators to collect the preference data. The main focus of this work was to prove the efficacy of RLHF for summarisation, as such there is little discussion on how the user interface was designed or how this may be impacting the engagement or agreement of raters for this task.

Finally, Bai et al. (2022) apply preference modelling and reinforcement learning from human feedback (RLHF) to fine tune language models to act as 'helpful and harmless assistants'. They explicitly outline summarisation as an example of a helpful task. They state that they found poor average agreement between researchers from Anthropic[1] and crowd sourced data and found that author-rater agreement wasn't a good guide for assessing overall conversation quality. Similarly to the work by Stiennon et al. (2020), the discussion of the interface, shown in Figure 2, is limited.

Both works are valuable in setting the groundwork for RLHF as a technique for LLM task alignment. However, the annotator agreement is a factor which is highlighted as unstable for differing tasks and settings in both papers. We argue that the design and usability of the interfaces used should be considered as a much more critical component in the paradigm of RLHF research.

3 Experimental Design

Our study compares the use of a baseline interface for summary comparisons with an novel interface designed in conversation with annotators. We investigate how the use of the new interface impacts both annotator engagement and agreement using specially trained annotators for the task of summary comparison.

3.1 Methodology

The study relies on two experimental settings: in the baseline setup, annotators are tasked with selecting the best summary from a set of 5 generated summaries using a standard interface. The summary selections include generations from a range of LLMs as well as human-written gold standard summaries. Further details on summary generation are provided in § 3.3. Prior to the study, annotators have worked with the initial interface for 6 months. In the second setting, annotators are asked to select the top n summaries in a ranked order, where n can be chosen by the annotator. Annotators interact with a novel user interface which has

[1] https://www.anthropic.com/

been designed in conversation with annotators to improve readability and aid in annotation judgements. We collect annotations for 500 documents (in each UI setting). Every document and set of summaries are annotated 3 times in total.

3.2 Annotators

As emphasised, the task of judging summary quality is non-trivial and the best results are attained using judgements from trained annotators. In our experiments, we use a team of 6 in-house annotators who have been trained on the task of judging summary quality. Annotators are paid at a daily rate, irrespective of their throughput, to incentivise high quality judgements. Annotator demographic information is included in Table 1. All annotators have worked on the task of summary evaluation for a minimum of 6 months prior to the study.

3.3 Data

The datasets provided to the annotators consisted of two batches containing 500 samples each. Each batch contains articles which have been scraped from the web. Summaries for these datasets were produced by fine-tuning the following language models: LaMDA (150B), Pegasus (500M) and FlanT5 XXL. All models used in this study were Transformer Encoder models with six layers and LSTM Decoders with two layers, containing approximately 27 million parameters, resulting in a file size of around 35MB. These small models are designed to run on-device and the data used to fine-tune the models for the task of summarisation were between 1-10K gold standard summaries.

3.4 Interface design

Annotators had been working with the baseline interface (interface 1) for 6 months prior to the study. To understand which features may improve the annotation experience we conduct a qualitative interview to identify pain points. Feedback from annotators was then used to design a novel interface which addressed the following two issues; (1) the readability of text and (2) the ability to highlight tokens in either the original or in the generated summaries to identify overlap more quickly.

Figure 3 shows screenshots of the baseline and updated interface. As demonstrated in the screenshot, the highlight interface allows for the selection of words within the original text and corresponding summaries. If a token is selected, all instances of the token will be highlighted to emphasise the overlap of summaries with the original. Annotators can select and de-select as many tokens as they want, an analysis of this behaviour is presented in Section 4.1.1.

4 Results

The following results section presents the findings related to annotator behavior and agreement in the decision-making tasks. The initial analysis focuses on measuring the time taken by annotators to make decisions in two task settings. Additionally, we examine annotators' engagement with the highlighting tool in the second interface, assessing the overlap between annotators and the tool's usage patterns. Lastly, we evaluate the level of agreement among annotators when selecting the best summary in both scenarios. These results provide insights into annotator decision-making processes, tool engagement, and the consensus achieved, contributing to a better understanding of the task dynamics and effectiveness in each setting.

4.1 Annotator engagement

The time taken for annotators to perform both tasks is recorded for each set of summaries presented. Figure 4 shows a plot of the time distributions normalised by the length of the texts and summaries for each interface. Interface 1 represents the original interface used by annotators and 2 is the highlight interface. Both time distributions are binned into 200 buckets and the density of occurrences for each bucket is plotted.

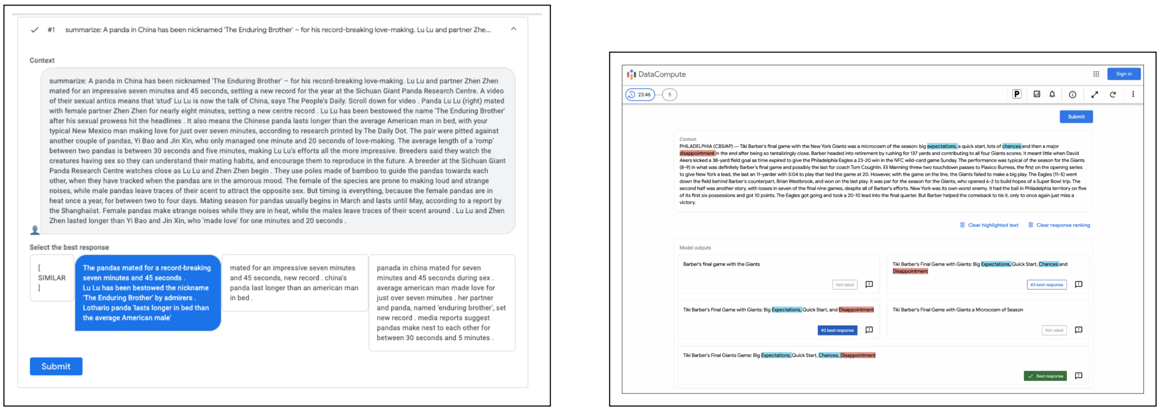
The task presented to annotators in the highlight interface is more cognitively demanding as annotators can select n best summaries instead of the best one. This is reflected in the proportion of time taken to perform the task as the histogram shows that the task takes longer to perform by annotators. There is a larger spread of time taken for annotators using the highlight interface which may be due to a lack of familiarity with the new set-up. However, from analysing the highlight behaviour of annotators we can see that the extent to which users are interacting with the highlighting tool differs greatly and this will contribute to a larger spread of times.

4.1.1 Token selection

Figure 5 presents histograms showing the varying degrees of engagement exhibited by annotators in response to the annotation task, as determined by the number of selected tokens. The histograms

Proficiency		Education		Age range		Hours reading English per week	
Native	1/6	Graduate	5/6	18 - 24	3/6	0 - 5	1/6
Near native	0/6	Undergraduate	1/6	25 - 34	3/6	5 - 10	2/6
Advanced	5/6	High School	0/6	35 - 44	0/6	10 - 15	1/6
Intermediate	0/6	Vocational Training	0/6	45 - 54	0/6	15 - 20	0/6
Beginner	0/6	No formal education	0/6	55+	0/6	20 +	2/6

Table 1: Background statistics for annotators in study


✓ #1 summarize: A panda in China has been nicknamed 'The Enduring Brother' – for his record-breaking love-making. Lu Lu and partner Zhe...

Context

summarize: A panda in China has been nicknamed 'The Enduring Brother' – for his record-breaking love-making. Lu Lu and partner Zhen Zhen mated for an impressive seven minutes and 45 seconds, setting a new record for the year at the Sichuan Giant Panda Research Centre. A video of their sexual antics means that 'stud' Lu Lu is now the talk of China, says The People's Daily. Scroll down for video . Panda Lu Lu (right) mated with female partner Zhen Zhen for nearly eight minutes, setting a new centre record . Lu Lu has been bestowed the name 'The Enduring Brother' after his sexual prowess hit the headlines . It also means the Chinese panda lasts longer than the average American man in bed, with your typical New Mexico man making love for just over seven minutes, according to research printed by The Daily Dot. The pair were pitted against another couple of pandas, Yi Bao and Jin Xin, who only managed one minute and 20 seconds of love-making. The average length of a 'romp' between two pandas is between 30 seconds and five minutes, making Lu Lu's efforts all the more impressive. Breeders said they watch the creatures having sex so they can understand their mating habits, and encourage them to reproduce in the future. A breeder at the Sichuan Giant Panda Research Centre watches close as Lu Lu and Zhen Zhen begin . They use poles made of bamboo to guide the pandas towards each other, when they have tracked when the pandas are in the amorous mood. The female of the species are prone to making loud and strange noises, while male pandas leave traces of their scent to attract the opposite sex. But timing is everything, because the female pandas are in heat once a year, for between two to four days. Mating season for pandas usually begins in March and lasts until May, according to a report by the Shanghaiist. Female pandas make strange noises while they are in heat, while the males leave traces of their scent around . Lu Lu and Zhen Zhen lasted longer than Yi Bao and Jin Xin, who 'made love' for one minutes and 20 seconds .

Select the best response

[SIMILAR]

The pandas mated for a record-breaking seven minutes and 45 seconds .
Lu Lu has been bestowed the nickname 'The Enduring Brother' by admirers .
Lothario panda 'lasts longer in bed than the average American male'

mated for an impressive seven minutes and 45 seconds, new record . china's panda last longer than an american man in bed .

panada in china mated for seven minutes and 45 seconds during sex . average american man made love for just over seven minutes . her partner and panda, named 'enduring brother', set new record . media reports suggest pandas make nest to each other for between 30 seconds and 5 minutes .

Submit

DataCompute

Sign in

23:46 5

Submit

Context

PHILADELPHIA (CBS/AP) — Tiki Barber's final game with the New York Giants was a microcosm of the season: big expectations, a quick start, lots of chances and then a major disappointment in the end after being so tantalizingly close. Barber headed into retirement by rushing for 137 yards and contributing to all four Giants scores. It meant little when David Akers kicked a 38-yard field goal as time expired to give the Philadelphia Eagles a 23-20 win in the NFC wild-card game Sunday. The performance was typical of the season for the Giants (8-9) in what was definitely Barber's final game and possibly the last for coach Tom Coughlin. Eli Manning threw two touchdown passes to Plaxico Burress, the first on the opening series to give New York a lead, the last an 11-yarder with 5:04 to play that tied the game at 20. However, with the game on the line, the Giants failed to make a big play. The Eagles (11-5) went down the field behind Barber's counterpart, Brian Westbrook, and won on the last play. It was par for the season for the Giants, who opened 6-2 to build hopes of a Super Bowl trip. The second half was another story, with losses in seven of the final nine games, despite all of Barber's efforts. New York was its own worst enemy. It had the ball in Philadelphia territory on five of its first six possessions and got 10 points. The Eagles got going and took a 20-10 lead into the final quarter. But Barber helped the comeback to tie it, only to once again just miss a victory.

Clear highlighted text   Clear response ranking

Model outputs

Barber's final game with the Giants — Not rated

Tiki Barber's Final Game with Giants: Big Expectations, Quick Start, Chances and Disappointment — #3 best response

Tiki Barber's Final Game with Giants: Big Expectations, Quick Start, and Disappointment — #2 best response

Tiki Barber's Final Game with Giants a Microcosm of Season — Not rated

Tiki Barber's Final Giants Game: Big Expectations, Quick Start, Chances, Disappointment — ✓ Best response

(a) Interface 1: baseline

(b) Interface 2: highlight interface

Figure 3: Screenshots of the annotation interfaces – the baseline interface is presented on the left and the highlight interface on the right.

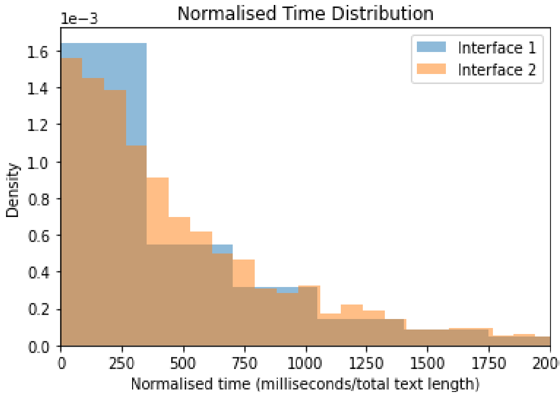


Figure 4: Histogram showing the time distribution for annotators to complete labelling using both interfaces.

depict the average number of highlighted words per annotator and reveal discrepancies in uptake among individuals. This data provides an insight into the highlighting practices of annotators. We see that the adoption varies, with one annotator (A3) not selecting any tokens during annotation compared with annotator (A5) who selects 509 tokens. The total tokens highlighted by all 6 annotators was 1836.

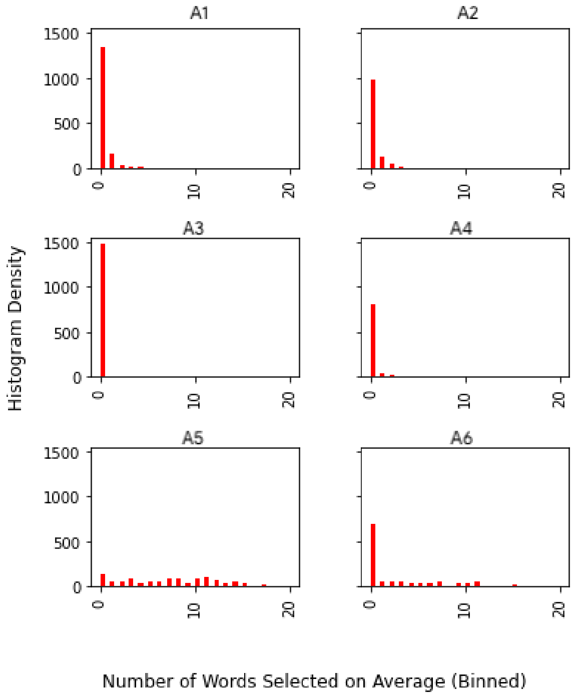


Figure 5: Histograms displaying the number of highlighted words for each annotation, labeled by annotator ID.

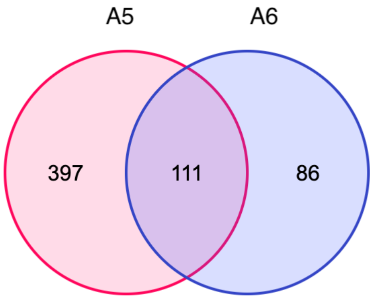


Figure 6: Venn diagram illustrating agreement between two most active annotators when selecting tokens for shared documents.

Token selection agreement Due to the variation in highlighting engagement we identify the two annotators who were most active in their use of the highlighting tool, as demonstrated in the bottom two histograms of Figure 5. To examine their degree of highlighting agreement in the annotation process, we focused on the subset of documents that were highlighted by both annotators, and investigated the crude overlap in terms of the tokens that were selected. The resulting venn diagram, shown in Figure 6, provides a visual representation of the extent of overlap between the two annotators' selected tokens for the 71 documents annotated by both.

Annotator A5 highlights a larger number of words in total than annotator A6. The overlap between their highlighted words was 22% of annotator A5's total and 56% of annotator A6's total. Of the words selected by both annotators for the shared documents 63% were nouns. It is worth noting that a more comprehensive analysis of token agreement will require a longer-term study, as annotator adoption of the highlighting tool is expected to increase over time.

4.2 Token selection

Using the total 1836 tokens selected across annotators, we find that there is a statistically significant correlation ($p < 0.01$) in the number of times a token is selected and the number of occurrences of that token in the original article. The total proportion of nouns selected is 64% which implies that the search of noun specific content words is most useful when considering whether generated summaries are high quality.

4.2.1 Annotator agreement

Table 2 shows the results of the pairwise Kappa agreement for annotators in both interface settings. The first interface yields a higher overall agreement compared to the second and the values range from 0.36 to 0.74 with an average of 0.59, while the values for the second setting range from 0.32 to 0.65 with an average of 0.46. These results show that there was higher inter-annotator agreement for interface 1 than for interface 2. In general, whilst there were some pairs of annotators who agreed more strongly than others for both interfaces, the results indicate that there is some variability in the inter-annotator agreement. Further efforts are needed to increase the consistency of annotations for the task, especially for Interface 2.

We posit that the lower annotator agreement in the second setting is for two reasons. Firstly, annotators are much less familiar with the new interface as this is the first experience they have with labelling via the new tool. Additionally, the new task requires a higher cognitive load as it involves selecting the best set of n summaries, as opposed to a single best summary. We found a substantial drop in the average agreement for annotators in the second setting, which raises questions about the stability of annotation and the complexity of the task. While this not the expected result, it provides an opportunity to investigate the task further. We plan to conduct a longitudinal study to examine whether annotator agreement improves with experience. Our preliminary results from this study are encouraging, showing that agreement increases as annotators become more familiar with the tool (an average kappa of 0.52 for the last 100 annotations in interface 2).

The highlight interface has an advantage in that it is designed to capture a more comprehensive range of behavioural information during the annotation process. One such behaviour is the frequency with which annotators change their top choice of summary. This is particularly useful when judging the difficulty of the decision, as it indicates the level of uncertainty for annotators. We investigate whether the level of agreement among annotators differs significantly based on the number of times they reselect their top choice in the highlight annotation interface. To do this we calculate significance using Satterthwaite's method (Kuznetsova et al., 2017), applied to a mixed-effects model that treats participants and the specific annotation task as crossed

	Annotator	A1	A2	A3	A4	A5	A6	Avg
Interface 1	A1	1.00	0.73	0.36	0.71	0.74	0.58	0.62
	A2	0.73	1.00	0.34	0.61	0.61	0.64	0.59
	A3	0.36	0.34	1.00	0.59	0.43	0.49	0.44
	A4	0.71	0.61	0.59	1.00	0.63	0.60	0.63
	A5	0.74	0.61	0.43	0.63	1.00	0.49	0.58
	A6	0.58	0.64	0.49	0.60	0.49	1.00	0.56
Interface 2	A1	1.00	0.45	0.52	0.39	0.44	0.32	0.42
	A2	0.45	1.00	0.45	0.42	0.65	0.38	0.47
	A3	0.52	0.45	1.00	0.42	0.48	0.45	0.46
	A4	0.39	0.42	0.42	1.00	0.36	0.40	0.40
	A5	0.44	0.65	0.48	0.36	1.00	0.45	0.48
	A6	0.32	0.38	0.45	0.40	0.45	1.00	0.40

Table 2: Kappa agreement between annotators for interface 1 (baseline) and interface 2 (highlight): results show a higher degree of agreement for annotators when using interface 1

random effects.[2] We find that there is a statistically significant relationship between the agreement of annotators and the frequency of changes made during the annotation process. This finding suggests that there are inherent indicators of annotator uncertainty in their behaviour prior to making a final decision.

5 Discussion

After receiving written feedback from annotators following the adoption of the new user interface, it was noted that all annotators stated the highlighting feature was useful. However, when analysing annotator behaviours not all annotators are using the tool. This presents an interesting issue of misaligned incentives, where annotators may feel the need to praise the new interface to maintain their employment status, even if they don't actually find it useful. While it's beneficial to have a consistent pool of annotators for engagement purposes, it's challenging to eliminate the power dynamic that arises from employing them directly. Therefore, performing an interaction-based analysis is valuable as it shows the true nature of tool adoption by annotators. It is possible that the lack of adoption among some annotators is due to unfamiliarity rather than a lack of utility, which may change over time.

In the second setting, we observed a reduction in annotator agreement compared to the first setting, which we attribute to both the change in interface and the new annotation task. Rather than selecting a single best summary, annotators were now allowed to choose multiple summaries, which increased the cognitive load. To determine if the decrease in agreement was due to the interface design or the increased cognitive load, we plan to conduct further experiments while controlling for the task. We also observed that annotator agreement tends to increase with greater exposure to the new interface, which suggests that familiarity with the tool is an important factor to consider.

The new interface has a significant advantage in that it enables us to use annotation behaviour to gain a better understanding of the task of summary comparison. For instance, we have observed that annotators who use the highlight option tend to prefer nouns as their preferred token type to search for. Furthermore, we have found that the stability of annotator choices during annotation (i.e. the frequency of deselecting an option) is a reliable indicator of annotator uncertainty and is strongly correlated with the level of agreement among annotators. These behaviours are statistically significant and can be used to predict the likelihood of achieving high agreement.

6 Limitations and Future Work

The authors would like to emphasise that this paper presents an initial pilot study aimed at documenting the process of updating an internal annotation tool. Our main contribution lies in emphasizing the

[2] Using R formula notation, the model is: $agreement \sim uncertainty + (1|participant) + (1|task)$. Tests were performed using the lme4 and lmerTest R packages by Bates et al. (2014).

influence of task conceptualization and interface design on annotator agreement. Additionally, we draw attention to the significant impact of the interface used for annotating summaries in the current human feedback reinforcement learning paradigm, which is often overlooked.

While there is a distinction between binary selection of the best summary and n-ary ranking, it is still the case that both scenarios involve selecting a preferred top candidate. Therefore, the substantial difference in agreement rates raises questions about the stability of the task and how the experimental setting can affect annotators' perception of summary quality, even among experienced and trained individuals. It is important to acknowledge that due to the variations in experiment settings and interfaces between the two task formats, it is difficult to draw definitive conclusions about the primary factor impacting annotator agreement. However, as an initial exploratory pilot study, our focus was primarily on assessing the tool's robustness and comparing the relative times taken in the different scenarios as well as measuring the annotator's usage of the new tool.

In future work, we will investigate how annotator behaviors can provide insights into the task difficulty and likelihood of agreement. This will involve analysing the interactions with the new interface, such as the time taken to complete the task, the frequency of selecting tokens, and the number of summary selections. By gaining a better understanding of the cognitive processes involved in annotation and how they affect annotator agreement, we can improve the development of annotation tools and methodologies for more accurate reward models.

7 Conclusion

The results of this pilot study emphasise how subtle variations in an annotation task can impact annotator agreement. Even highly experienced annotators can experience fluctuations in agreement as a result of interface changes. To aid in the annotation of summary comparison, we developed a new interface that allows tokens to be selected and displayed across resulting summaries and observed patterns in the types of tokens highlighted by annotators. Moving forward, we plan to conduct additional studies to explore the use of implicit interaction signals in predicting annotator agreement.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320.

Paul N. Bennett, David Maxwell Chickering, and Anton Mityagin. 2009. Learning consensus opinion: mining data from a labeling game. In *Proceedings of the 18th international conference on World wide web*, pages 121–130. ACM.

Carla E. Brodley and Mark A. Friedl. 1999. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, pages 131–167.

Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or there preference judgments for relevance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4956 LNCS:16–27.

Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. 2021. Correlated input-dependent label noise in large-scale image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1551–1560.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Benoit Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151.

Sian Gooding, Ekaterina Kochmar, Advait Sarkar, and Alan Blackwell. 2019. Comparative judgments are more consistent than binary classification for labelling word complexity. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 208–214, Florence, Italy. Association for Computational Linguistics.

David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definition. Association for Computational Linguistics (ACL).

Panagiotis G. Ipeirotis, Foster Provost, Victor S. Sheng, and Jing Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Karen Sparck Jones and Julia R Galliers. 1995. Evaluating natural language processing systems: An analysis and review.

David C. Kingsley and Thomas C. Brown. 2010. Preference Uncertainty, Preference Learning, and Paired Comparison Experiments. *Land Economic*, 86:530–544.

Alexandra Kuznetsova, Per B Brockhoff, Rune HB Christensen, et al. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13):1–26.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52:101–148.

Inderjeet Mani. 2001. Automatic summarization. *Automatic Summarization*, pages 1–298.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics.

Josef Steinberger and Karel Jezek. 2009. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer G. Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *International conference on artificial intelligence and statistics*, pages 932 – 939.

Yi-Hsuan Yang and Homer H. Chen. 2011. Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:762–774.