For baseline:

- The system you used DID NOT HAVE explainability features. Please recall your actions when interacting with the system.
    - In general, how did you decide whether to agree or disagree with the system?
        - What was your general strategy or approach?
        - What types of evidence did you use to inform you decision to agree or disagree?
        - How did you use these types of evidence?

For conf+sent:

- The system you used HAD explainability features. Please recall your actions when interacting with the system.
    - In general, how did you decide whether to agree or disagree with the system?
        - What was your general strategy or approach?
        - What types of evidence did you use to inform you decision to agree or disagree?
        - How did you use these types of evidence?

- This system displayed its confidence value. For any of the articles that you judged, did you finding yourself looking at these confidence values?
    - What were you trying to achieve by engaging with the confidence value feature of the system?
    - How did the confidence value feature of the system help you decide whether to agree or disagree with the system?
    - Did you experience any difficulties while engaging with the confidence value feature of the system? Can you elaborate on these difficulties?

- This system allowed you to highlight which sentences it considered to be most influential. For any of the articles that you judged, did you engage with the sentence highlighting feature?
    - What were you trying to achieve by engaging with the sentence highlighting feature of the system?
    - How did the sentence highlighting feature of the system help you decide whether to agree or disagree with the system?
    - Did you experience any difficulties while engaging with the sentence highlighting feature of the system? Can you elaborate on these difficulties?

Can you explain your rationale for agreeing/disagreeing with the system? How did you make your decision? Participants were shown their decisions and were asked 4 times in each interface condition = 8 times in total. Each decision had the highest Confidence in that category.

Agree (TP); Disagree (TP); Agree (FP); Disagree (FP)