What is a quote?

A quotation is the repetition of a sentence, phrase, or passage from speech or text that someone has said or written. In oral speech, it is the representation of an utterance that is introduced by a quotative marker, such as a verb of saying. For example: John said: "I saw Mary today". In written text, quotations are signaled by quotation marks.

The model is trained to recognise **3 entities**:

SOURCE - the speaker, which might be a person, an organisation etc.
CUE - usually a verb phrase, indicating the act of speech or expression
CONTENT - the direct quote, including quotation marks

SOURCE and/or CUE must always be accompanied by CONTENT, but CONTENT can exist without a SOURCE or CUE.

Examples of different quotation styles:

1. She said: “This is a nicely structured quote.”
2. She also said: “Sometimes quotes are very short. But sometimes they include multiple sentences or even paragraphs.”
3. “Quotative inversion occurs where the direct quotation occurs before a cue-verb,” he added.
4. He has told us: “In this case we have an auxiliary verb which we do not include as ‘cue’.”
5. The annotator noted “Sometimes they omit colons from quotes.”
6. Jane Doe, the journalist, rejected the statement, saying “that isn’t true”.
7. They were criticised by the annotator because “their use of quotation marks is not consistent”.
8. The annotators said they were puzzled. “Really? Yet another quote format?”
9. He said that for a machine “to distinguish between paraphrases and quoted terms will be difficult”.
10. “It's difficult indeed,” the annotator said. He looked puzzled. “What even is a quote really?
    “Sometimes it spans paragraphs without closing quotation marks.”
11. It is not clear if this counts as a “real quote” or not, she said.
12. He said they were not sure either.
    “And then the next paragraph is a quote.”

13. After she warned that quotes “would be hard to detect”, she added - “we will try to do our best”.
14. The annotator got annoyed and said: “When we thought we had listed all the quote styles we found this…”, he said. 🙅
15. “And this!” Another annotator screamed.

Exclusions:

- She said that sometimes journalists used paraphrases without quotation marks. We decided not to label text without quotation marks as a quote.

  Kwarteng said the number of Covid-related deaths was not yet causing concern because it was significantly lower than during the third wave.

- This is not a quote, it just uses quotation marks to indicate a non-standard English term like “woke”.

  Far from offering today's refugees “relief and salvation”, they are detained in camps and our doors are closed to lone child refugees stranded in Europe, even those whose only family is here and desperate to offer them a home.

- Sometimes quotation marks are used for dramatic effect or to indicate hypothetical speech. “Why am I doing this?” the annotator thought.

  May embellished this theme, suggesting that Scottish police officers never have occasion to say “You're nicked, sunshine.”

- The annotator’s motto was “hope for the best, prepare for the worst” but that’s not a quote.

  “Whatever gets the job done” is my new motto.

- People annotating articles often said a combination of a “broad source and vague content” does not qualify as a quote.

  Brooks Koepka took umbrage with the golfing media’s coverage of recent comments that team sports were “just maybe not in his DNA”, which have snowballed into public questioning of his commitment to Team USA and the Ryder Cup.

Additional rules when annotating:

1. Label quotation marks and other punctuation inside the marks as CONTENT. Do not label paraphrases without quotation marks as CONTENT.

   Although 223 daily Covid-related UK deaths were reported on Tuesday – higher than on the same day last autumn – Kwarteng said: “I don’t see any cause for changing the course at this minute.”

2. Annotate as CONTENT only the text inside quotation marks and the quotation marks (i.e. exclude text after CUE and before quotation marks).

   However, the NHS Confederation, which represents the healthcare system in England, Wales and Northern Ireland, told the Guardian that immediate action was required to prevent the NHS "stumbling into a crisis" where the elective care recovery would be jeopardised.

3. Annotate quotes which span multiple paragraphs with a single CONTENT label (one quote).

   He told BBC Radio 4's Today programme: "We're looking at data on an hourly basis … For now, we think that the policy is working. Yes, increases of infection rates are being seen. But at the same time we're very closely monitoring hospitalisations and death rates. Mercifully, they're much, much lower than they were at the beginning of the year.

   "That doesn't mean we're being complacent. But we do feel that the vaccination rollout has been successful, it's allowed us to reopen the economy, it's allowed people to get back to some semblance of normality."

4. When a quote is split by words such as *and* or additional *cue-verbs,* annotate each part of the quote with separate CONTENT labels (multiple quotes).

   The CMA said it was the first time a company had "consciously refused" to supply information under an IEO and said it "considers that Facebook's failure to comply was deliberate".

5. Do not label auxiliary verbs as CUE (has told, was accused).

   Last year, Facebook was criticised by the Competition Appeal Tribunal and the court of appeal for employing "what might be regarded as a high-risk strategy" by not cooperating fully with the CMA and the IEO.

6. Label indirect cues, such as in his words or according to, as CUE.

   According to the Center for American Progress, 30 US senators and 109 representatives "refuse to acknowledge the scientific evidence of human-caused climate change".

7. Phrasal verbs, like called for, should be labelled as CUE in their entirety.

   Jane Doe called for a "detailed investigation into the issue, to determine …"

8. If a title/function is included alongside the name, include it in SOURCE. ("These are the rules", said PERSON, the annotator).

   Saffron Cordery, the deputy chief executive of NHS Providers, which represents NHS hospitals, ambulance, community and mental health services, said "hard decisions" may have to be made about which patients to prioritise if Covid cases continue to rise.

9. If a SOURCE is named earlier, but there is a personal pronoun closer to the CONTENT, the pronoun should be labelled.

England's chief medical officer, Professor Chris Whitty, stressed the importance of mask-wearing and also encouraged people to take up the offer of a vaccination. "Covid-19 cases are rising and winter is drawing closer," he said. "Ventilation, masks in crowded indoor spaces and hand-washing remain important."

10. Do not include adjectives or additional description in SOURCE.

    Chris Garrard, of the campaign group Culture Unstained, which obtained the emails under freedom of information legislation, said: "For years oil companies have been given prominent platforms at the UN climate negotiations, promoting themselves as climate leaders while they continued to pour millions into new fossil fuels, so this is a big step forward."

11. Do not annotate as SOURCE people mentioned previously without a direct indication of speech attached to the CONTENT. This would be an orphan quote.

    Looking out for others and being a team player is important to Woodward, and her advice to those early in their editing careers shows the value she places on nurturing relationships. "If you prove to people that you're hardworking and nice to be around, they will want to keep you close by, and take you on to the next job with them. The industry is so welcoming, and everyone is willing to share – we have so many WhatsApp groups where we help each other."

12. For quotes within quotes, annotate the outermost quote as CONTENT.

    "She said to me: 'She and Steve are up there looking after us', which is a lovely thought. I like the idea that he sent this to us to lift our spirits."

Notes:

- The decision to extract ONLY CONTENT inside quotation marks might be controversial as we risk losing the meaning and context of the quote. It should be noted that we will be able to extract context if needed once the quote is identified in the text.
- The decision to extract names with job titles means there needs to be an extra step to identify personal and company names from SOURCE but we believe that this approach will give us a more accurate output.