Evaluation of Summaries

Cillian Berragan

This document compares summaries written by Cambridge, to the summaries generated automatically by our model.

Overview

For all representations, the original summary was compared with the generated summary provided by the LLM. A separate LLM call was used to determine which of these two summaries was preferred, based on set criteria:

> A good summary should:
> 1. **Be accurate** – It should not include information that is not present in the source document.
> 2. **Be comprehensive** – It should reflect all key points in the source document without omitting important details.
> 3. **Be well-grounded** – It should be based entirely on the source document without adding interpretations, opinions, or external information.

This model was given the option to return 4 different scores; 0 meaning neither summaries are suitable, 1 meaning the original summary is preferred, 2 meaning the LLM-generated summary is preferred, or 3 meaning both summaries are suitable.

Table 1 gives the results of this processing. We can see that the majority of the preferred summaries are those generated by the LLM (2). There are however 8 cases where the original summary is considered better, and 17 where both summaries are considered suitable.

Table 1

Original	LLM-generated	Both
8	65	17