From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting

Griffin Adams[♠,♣] griffin.adams@columbia.edu
Alexander R. Fabbri[♢] afabbri@salesforce.com
Faisal Ladhak [♠] faisal@cs.columbia.edu
Eric Lehman[♡] lehmer16@mit.edu
Noémie Elhadad[♠,♣] noemie.elhadad@columbia.edu
Columbia University: CS[♠]**, Biomedical Informatics**[♣] **Salesforce AI**[♢] **MIT**[♡]

A GPT-4 Metrics

For the GPT-4 Likert-style evaluation, we use the following prompt template.

```
Article: {{Article}}

Summary: {{Summary}}

Please rate the summary
(1=worst to 5=best) with
respect to {{Dimension}}.

{{Definition}}
```

Below, we present the definitions provided for each quality metric.

- **Informative**: An informative summary captures the important information in the article and presents it accurately and concisely.
- **Quality**: A high quality summary is comprehensible and understandable.
- **Coherence**: A coherent summary is well-structured and well-organized.
- **Attributable**: Is all the information in the summary fully attributable to the Article?
- **Overall Preference**: A good summary should convey the main ideas in the Article in a concise, logical, and coherent fashion.

The **Quality** and **Coherence** prompts do not include the Article in the prompt. These definitions were paraphrased from previous summarization annotation efforts: (**??**).

Meta-Evaluation. To compute the summary-level correlation, we first turned the preference data into a vector representing the number of times that summary received a first-placed vote. Table 1 demonstrates, unsurprisingly, that a prompt designed to capture overall summary rating has the highest summary-level Pearson correlation to overall preferences (31), yet overall correlations are still low.

Dimension	Correlation
Informative	0.215
Quality	0.120
Coherence	0.178
Attributable	0.245
Overall	**0.311**

Table 1: Summary-Level Pearson Correlation coefficient between human preferences and GPT-4 Likert ratings.