W3: Regarding experiment comparison, I highly recommend comparing AutoSurvey to not only naive RAG but more advanced methods, which could make the results more convincing.

A3:

To address your concerns regarding experiment comparison, we have supplemented the original naive RAG-based experiments by including refinement and query rewriting stages. In Naive RAG + Refinement, the LLM is required to enhance the continuity of the written content with previous sections and check for factual errors based on the references retrieved. In Naive RAG + Query Rewriting, references are first retrieved using the topic, after which the LLM rewrites the query based on the references to assist in writing subsequent content.

Survey Length (#tokens)	Methods	Recall	Precision	Coverage	Structure	Relevance
8k	Naive RAG-based LLM generation + Refinement	82.25	76.84	4.46	4.02	4.86
	Naive RAG-based LLM generation + Query Rewriting	80.99	71.83	4.84	4.05	4.88
16k	Naive RAG-based LLM generation + Refinement	79.67	73.73	4.57	4.28	4.83
	Naive RAG-based LLM generation + Query Rewriting	77.73	66.29	4.70	3.67	4.79
32k	Naive RAG-based LLM generation + Refinement	80.50	72.18	4.82	4.08	4.49
	Naive RAG-based LLM generation + Query Rewriting	76.56	65.36	4.61	3.96	4.88
64k	Naive RAG-based LLM generation + Refinement	73.12	68.36	4.66	4.06	4.76
	Naive RAG-based LLM generation + Query Rewriting	69.77	62.21	4.45	3.88	4.69

After adding the refinement stage, both citation quality and structure improved. The effect of query rewriting is not obviously enhanced, possibly due to the model's lack of a clear planning of the content to be written, leading to lower quality of rewritten queries. Overall, these baselines still lag behind AutoSurvey, especially when surveys get longer. This gap may be attributed to the streaming generation process, where each step must reference previous content, leading to the accumulation of errors. To validate this, we segmented the extracted claims into 20% intervals and calculated the citation recall for each segment. The results indicate that the recall of Naive RAG gradually decreases as the generated text length increases, while AutoSurvey maintains stable performance.

Claims	20%	20%~40%	40%~60%	60%~80%	80%~100%
Naive RAG-based LLM generation (64k)	76.79	73.17	71.52	64.08	49.85
AutoSurvey (64k)	82.86	84.89	79.04	82.27	82.29

References: [1] SELF-RAG: Learning to Retrieve, Generate and Critique through self-reflection (ICLR 2024) [2] Query Rewriting in Retrieval-Augmented Large Language Models (EMNLP 2023) [3] Long Text Generation by Modeling Sentence-Level and Discourse-Level Coherence (ACL 2021)