Hypothesis

The number of tokens produced in the output of a Large Language Model (LLM) will not be significantly impacted by adjustments to the temperature parameter, which regulates the randomness and creativity of its responses.

Design

- Objective: The objective of this experiment is to determine whether the temperature parameter, which regulates the model's inventiveness, affects the number of tokens that are used in the answer.
- Model: Throughout the experiment, we will be using Google's "gemma-3-27b-it" model.
- Input data: A fixed set of 10 pre-defined review clusters which is stored in snowflake database will be used as the data for this experiment.
- Temperature settings: We will be conducting this experiment in the temperature values – 0, 0.25, 0.5, 0.75, 1
- System prompt: For this experiment, we will be following a common system prompt (attached in the appendix section) for all the scenarios.
- Evaluation: We will assess whether the temperature parameter affects token usage by computing summary statistics (mean, median, standard deviation, range) of total tokens for each temperature and analysing the correlation between temperature and total tokens.

Procedure:

1. Select temperature settings:
   - Define the set of temperatures to test: 0.0, 0.25, 0.5, 0.75, 1.0
2. Prepare the input data for LLM:
   - Retrieve all reviews and centroid-closest reviews from the snowflake.
   - For each cluster, select the centroid reviews (30 number) and 20 random reviews (other than the 30 centroid reviews) from the same cluster.
   - Concatenate centroid and random reviews into the input text for the LLM.
3. Summarization with LLM:
   - For each cluster in the temperature setting, we will pass a fixed prompt containing cluster ID, selected reviews, and task instructions

- Pass this to the LLM model (gemma- 3-27b-it)
- Record the title, description, and token usage.

4. Text quality evaluation:
   - For each summary, calculate the BLEU, BERT and ROUGE scores.
5. Store the findings:
   - Use a Snowflake database to store the results in the table TEMP_EXP, with the columns: Cluster ID, title, description, top reviews, evaluation metrics, temperature, and total tokens.

Result

Summary Statistics

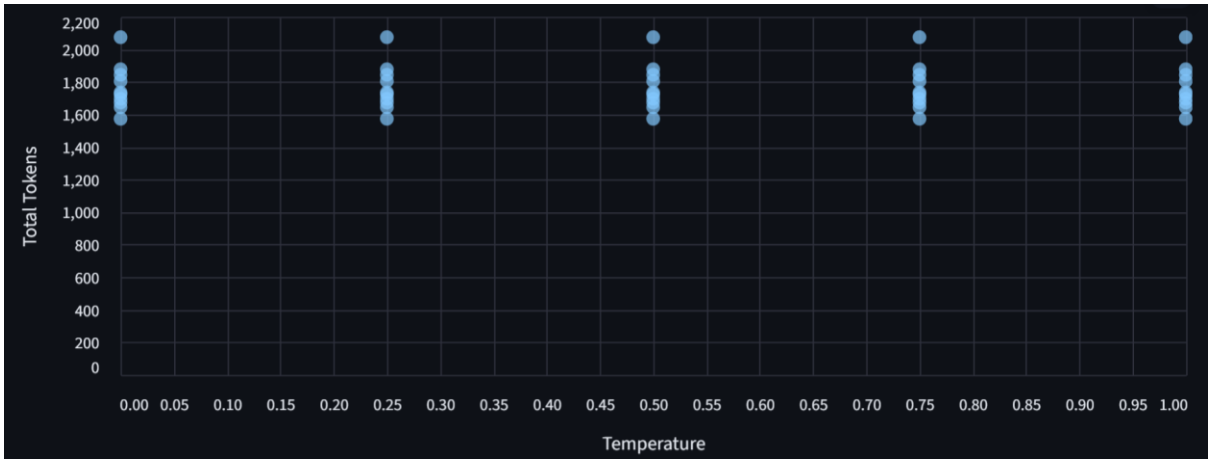
Temperature	Avg Tokens	Std Tokens	Min Tokens	Max Tokens	Avg ROUGE-	Avg ROUGE-2	Avg ROUGE-L	Avg BERT-F1	Avg BLEU
0	1764.9	142.79	1574	2075	0.1648	0.0251	0.1155	0.4224	0.0000000000000007
0.25	1764.9	142.79	1574	2075	0.1528	0.0251	0.1121	0.4204	0.0000000000000005
0.5	1764.9	142.79	1574	2075	0.1603	0.0248	0.1183	0.4212	0.0000000000000006
0.75	1764.9	142.79	1574	2075	0.1614	0.0229	0.1176	0.422	0.000000000000002
1	1764.9	142.79	1574	2075	0.1536	0.0234	0.1144	0.4199	0.0000000000000005

From the above results we can observe that the average, standard deviation, minimum, and maximum token usage are consistent across all temperature values (0.0,0.25,0.5,0.75,1) suggesting that temperature has no noticeable effect on token usage. But minor differences exist in the evaluation metrics (ROUGE, BERT-F1, BLEU), indicating that temperature has a negligible effect on output quality. This suggests that temperature slightly affects the diversity and wording of the generated text, but not enough to impact length or token count.

Temperature vs. Total Tokens

From the graph below, we can observe that for each temperature value (0, 0.25, 0.5, 0.75, 1), the distribution of total tokens is visually identical since the points are clustered at the same ranges. There are no upward or downward

trends in the graph. The consistent spread and density of the points confirm that the temperature adjustments do not influence token count.



Recommendation

Since token usage varies very little between temperature settings, the need for predictable behaviour should be the main consideration. Lower temperatures guarantee more consistent and repeatable answers without affecting token consumption by reducing randomness in model outputs. Consequently, it is advised to use a lower temperature in order to satisfy the determinism criteria and to avoid any scenario of the model hallucinating while keeping the output length constant.

Appendix

System Prompt

```
    You are an executive of a company.
    You are given a set of product reviews from the same cluster.

    Task:
    1. Generate a short title (max 8 words) suited for the main theme of the
reviews of the cluster.
    2. Write a concise, actionable description (max 50 words) summarizing the main
idea of this cluster.
    3. Ensure the title and description are clear and understandable to executives
who want to decide quickly what to address next.
```

```
    4. Focus on insight, not just summary—Use plain, professional English that is
easily understandable by non-technical stakeholders.

    Context:
    - The reviews are from a specific cluster.
    - Executives must be able to scan the title and description to grasp the main
takeaway.
    - The title and description should be concise and informative. Avoid vague
language—be specific and impactful.

    Inputs:
        Cluster ID: {cluster_id}
        Reviews:
        {cluster_texts}

    Output format:
        Title: <title>
        Description: <description>
```