Appendix

A PAW and MM-PAW prompt template

A.1 MM-PAW Template

You are a friendly, expert, and helpful agent helping a content creator write coherent sections to create a document on `article_name`.

You will be given the heading of the section you are supposed to write, and the title of the document under which this section should occur. Additionally, you will be given some initial context, and reference sentences to use generate the section.

First, come up with a plan with various topics to be discussed to write a section on `section_name`. Then, write a section using the generated plan by filling it with the reference sentences in more than `min_num_words` and less than `max_num_words` words. Do not use your own knowledge and only rely on reference sentences. Give image descriptions that are suitable for the section. Only output the final section content and image description.

Section heading: `section_name`

Document title: `article_name`

Initial context: `init_context`

Reference sentences: `references`

Output format:

{

"Plan": ["Key topic 1", "Key topic 2", "Key topic 3"],

"Section content": "section generation output"

"Image descriptions": ["Image decription 1", "Image description 2", "Image description 3"]

}

Output only a valid JSON from now on

A.2 PAW Template

You are a friendly, expert, and helpful agent helping a content creator write coherent sections to create a document on `article_name`.

You will be given the heading of the section you are supposed to write, and the title of the document under which this section should occur. Additionally, you will be given some initial context, and reference sentences to use generate the section.

First, come up with a plan with various topics to be discussed to write a section on `section_name`. Then, write a section using the generated plan by filling it with the reference sentences in more than `min_num_words` and less than `max_num_words` words. Do not use your own knowledge and only rely on reference sentences. Only output the final section content.

Section heading: `section_name`

Document title: `article_name`

Initial context: `init_context`

Reference sentences: `references`

Output format:

{

"Plan": ["Key topic 1", "Key topic 2", "Key topic 3"],

"Section content": "section generation output"

}

Output only a valid JSON from now on

B Baseline prompt template

B.1 Baseline Template

You are a friendly, expert, and helpful agent helping a content creator write coherent sections to create a document on article$_{n}ame$.

You will be given the heading of the section you are supposed to write, and the title of the document under which this section should occur. Additionally, you will be given some initial context, and reference sentences to use generate the section.

Your goal is to come up with a section based on the given inputs in more than `min_num_words` and less than `max_num_words` words. Do not use your own knowledge and only rely on reference sentences.

Section heading: `section_name`
Document title: `article_name`
Initial context: `init_context`
Reference sentences: `references`

C G-Eval Prompt Templates

C.1 Coverage

You are an expert evaluator of text generation quality.

You will be given three sections: two of them generated by two AI models, and the third one is a reference section.

Your task is to rate the quality of the model-generated section texts using the given reference text.

Evaluation Criteria:

Coverage: Compare each model-generated text with the reference text to check their coverage. Outputs with high coverage cover most important aspects discussed in the reference text.

Evaluation Steps:

1. List the key topics or subjects addressed in the reference text.
2. Examine each model-generated text to identify whether it addresses the key topics from the reference.
3. Compare the content of the model-generated texts with the reference text.
4. Look for instances where the model-generated text addresses or omits important topics.
5. After addressing the above factors, score the output text on a scale of 1 (low quality) to 5 (high quality).

Output Format: The output form should be a list of scores `[model_1_score, model_2_score]`.

Reference Text: `{reference_text}`

Model-Generated Texts:

Text generated using Model 1: `{model1_output}`

Text generated using Model 2: `{model2_output}`

Evaluation Form (List of Scores ONLY):

C.2 Groundedness

You are an expert evaluator of text generation quality.

You will be given two sections that are automatically generated by AI models, and reference sentences used to generate the sections.

Your task is to rate the quality of the model-generated section texts using the given reference text.

Evaluation Criteria:

Grounding: This refers to the extent to which the content produced by a model is substantiated and supported by the information presented in the reference sentences.

Evaluation Steps:

1. Examine each model-generated section to identify the specific claims, statements, or information it presents.
2. Determine whether each element in the model-generated section is directly supported by corresponding information in the reference sentences.
3. Penalize if portions of the model-generated section lack direct support from the reference sentences.
4. Reward portions of the model-generated section that align well with and are directly supported by the reference sentences.
5. After addressing the above factors, score the output text on a scale of 1 (low grounding) to 5 (high grounding).

Output Format: The output form should be a list of scores `[model_1_score, model_2_score]`.

Reference Text: `{reference_text}`

Model-Generated Texts:

Text generated using Model 1: `{model1_output}`

Text generated using Model 2: `{model2_output}`

Evaluation Form (List of Scores ONLY):

C.3 Overall Structure

You are an expert evaluator of text generation quality.You will be given three sections: two of them generated by two AI models, and the third one is a reference section. Your task is to rate the quality of the model-generated section texts using the given reference text.

Evaluation Criteria:

Coverage: Compare each model-generated text with the reference text to check their coverage. Outputs with high coverage cover most important aspects discussed in the reference text.

Fluency: Assess the grammar, syntax, and naturalness in the model-generated texts. Ensure that the sentences are well-formed and coherent.

Style consistency: Assess the tone and style of the model-generated texts. It should mirror the tone and style of the reference text.

Evaluation Steps:

1. List the crucial aspects or topics discussed in the reference text and examine each model-generated text to identify the coverage of key aspects from the reference text.
2. Assess the overall coherence and natural flow of sentences in the model-generated texts. Check for varied sentence structures and ensure that they contribute to a smooth reading experience.
3. Evaluate whether the tone and style of the model-generated texts mirror those of the reference text.
4. After addressing the above factors, score the output text on a scale of 1 (low quality) to 5 (high quality).

Output Format: The output form should be a list of scores `[model_1_score, model_2_score]`.

Reference Text: `{reference_text}`

Model-Generated Texts:

Text generated using Model 1: `{model1_output}`

Text generated using Model 2: `{model2_output}`

Evaluation Form (List of Scores ONLY):

D Standard Deviation of experiments

Method	Overall RL F1 Score	SD
BL GPT-4	23.33	1.45
PAW	23.80	1.27
MM-PAW	24.04	0.98
BL Claude (Haiku)	23.43	1.32
PAW	25.31	1.79
MM-PAW	25.09	1.12
BL GPT-3.5	19.90	1.14
PAW	22.73	1.67
MM-PAW	21.52	0.83
BL LLaMa 2 (70B)	16.16	1.58
PAW	19.81	1.43
MM-PAW	18.55	0.97
BL LLaMa 2 (13B)	11.41	1.03
PAW	16.59	1.62
MM-PAW	16.38	1.54
BL LLaMa 2 (7B)	9.41	1.47
PAW	12.90	1.78
MM-PAW	13.13	1.13
BL Mistral (8x7B)	14.22	1.35
PAW	21.29	1.27
MM-PAW	22.26	1.69
BL Mistral (7B)	9.57	1.64
PAW	16.40	1.11
MM-PAW	15.97	0.87

Table 5: Standard Deviations of overall RL F1 scores for each model and variant

E Example Outputs (PAW)

E.1 Example 1

The topics that are present in the ground truth but are either missing in PAW or Baseline output are highlighted in red.

Ground Truth

Virginia has several nicknames, the oldest of which is the "Old Dominion. " King Charles II of England first referred to "our auntient dominion of Virginia" in 1660, the year of his restoration, perhaps because Virginia was home to many of his supporters during the English Civil War. These supporters were called Cavaliers, and the nickname "The Cavalier State" was popularized after the American Civil War. Students at the University of Virginia began using The Cavalier Song as their school fight song in 1925, and the school's sports teams were named Cavaliers after the song. Virginia has also been called the "Mother of Presidents", as eight Virginians have served as President of the United States, including four of the first five. The state's motto, Sic Semper Tyrannis, translates from Latin as "Thus Always to Tyrants", and is used on the state seal, which is then used on the flag. While the seal was designed in 1776, and the flag was first used in the 1830s, both were made official in 1930. The majority of the other symbols were made official in the late 20th century. The Virginia reel is among the square dances classified as the state dance. In 1940, "Carry Me Back to Old Virginny" was named the state song, but it was retired in 1997 due to its references to slavery. In March 2015, Virginia's government named "Our Great Virginia", which uses the tune of "Oh Shenandoah", as the traditional state song and "Sweet Virginia Breeze" as the popular state song.

Baseline Output

Virginia is home to many state symbols. Virginia's state bird is the Northern cardinal, which was declared the state bird in 1950. The state's state flower is the American dogwood, which was designated in 1918. Virginia's state beverage is milk, while its state fish is the brook trout. Virginia's state seal is based on a design described in the convention of 1776 and represents Virtus, the genius of the commonwealth, dressed like an Amazon, resting on a spear with one hand, and holding a sword in other, and treading on tyranny, represented by a man prostrate, a crown fallen from his head, a broken chain in his left hand and a scourge in his right. The exergon has the word Virginia over Virtus's head and underneath the motto "Sic Semper Tyrannis".

PAW Output

Plan

1. State Seal
2. State Motto
3. State Flag
4. State Nicknames
5. State Songs
6. State Animals

Write

Virginia is home to a variety of state symbols that reflect its rich history. One of the most recognizable symbols is the state seal which features Virtus, the genius of the Commonwealth, dressed as an Amazon, standing atop tyranny, represented by a fallen man holding a broken chain and a scourge.

The seal also includes the state motto "Sic Semper Tyrannis" which means "Thus Always to Tyrants". This seal is found on the state flag as well. Virginia also shares the designation of "Commonwealth" with only three other states. The state nickname, "Old Dominion," comes from Virginia's historical significance as the first permanent English colony in the new world and thus representing old antiquity or a long-standing Custom. Additionally, Virginia has two state songs, a traditional song, and a popular song that encapsulate the sentiments of the state. Finally, Virginia also recognizes several state animals including the Virginia Big-Eared Bat and the northeastern tiger salamander, and state plants such as the dogwood and American dogwood. Virginia's state beverage is milk, while its state fish is the brook trout. All these symbols represent the diverse history, culture, and natural beauty of Virginia.

E.2 Example 2

Ground Truth

The Vatican City State budget, covering the Vatican Museums and post office, is sustained by the sale of stamps, coins, medals, and tourist mementos, as well as admission fees to museums and publication sales. Lay workers enjoy comparable incomes and living standards to their counterparts in Rome. Other industries include printing, mosaic production, and staff uniform manufacturing. Vatican City issues its own coins and stamps but uses the euro as its currency since January 1, 1999, by agreement with the European Union. The Vatican had its own lira currency before adopting the euro. The city-state, employing nearly 2,000 people, reported a surplus of 6.7 million euros in 2007 but faced a deficit of over 15 million euros in 2008. In 2012, Vatican City was listed among countries of concern for money laundering. In 2014, the Vatican established a secretariat for the economy to oversee economic, financial, and administrative activities.

Baseline Output

Vatican City's economy relies on stamp, coin, and memento sales, along with museum admission fees and publication sales. The per capita GDP is estimated at $21,198, ranking it as the 18th wealthiest nation. The Vatican's economy is volatile due to investments in artifacts, gold, and heavy dependence on tourism. The Nominal GDP is unpublished, but the Vatican receives subsidies from Italy, uses the Euro, and engages in international banking and financial activities. Tourists can enjoy the cultural experience the Vatican has to offer.

PAW Output

Plan

1. Sources of income for the Vatican
2. Per capita GDP of Vatican City and its ranking in the world
3. Volatility of the Vatican's economy
4. Budget of Vatican City and its expenditures
5. Vatican City's reliance on Italy for subsidies and free water supply
6. Recent financial reforms implemented by the Vatican

Write

Vatican City's economy relies on stamp, coin, and memento sales, along with museum admission fees and publication sales. Other industries that contribute to the economy include printing, mosaic production, and staff uniform manufacturing. The per capita GDP is estimated at $21,198, ranking it as the 18th wealthiest nation. The Vatican's economy is volatile due to investments in artifacts, gold, and heavy dependence on tourism. The Nominal GDP is unpublished, but the Vatican receives subsidies from Italy, uses the Euro, and engages in international banking and financial activities. In 2014, the Vatican established a secretariat for the economy to oversee economic, financial, and administrative activities.

F Example Outputs (MM-PAW)

F.1 Example 1

Intent: Vertebrate - First Vertebrates

Text: Vertebrates originated during the Cambrian explosion, which saw a rise in organism diversity. The earliest known vertebrates belongs to the Chengjiang biota and lived about 518 million years ago. These include Haikouichthys, Myllokunmingia, Zhongjianichthys, and probably Haikouella. Unlike the other fauna that dominated the Cambrian, these groups had the basic vertebrate body plan: a notochord, rudimentary vertebrae, and a well-defined head and tail. All of these early vertebrates lacked jaws in the common sense and relied on filter feeding close to the seabed. A vertebrate group of uncertain phylogeny, small eel-like conodonts, are known from microfossils of their paired tooth segments from the late Cambrian to the end of the Triassic. Refer Figure 4 and 5.

Plan

1. Cambrian explosion lead to rise in organism diversity
2. Discuss the earliest known vertebrates - *Myllokunmingia*, *Haikouichthys ercaicunensis*
3. Mention vertebrate body structure
4. Discuss jawless vertebrates and their eating habits

Image Descriptions

1. Illustration of the *Myllokunmingia*, the earliest known vertebrate
2. Reconstruction of the *Haikouichthys ercaicunensis*, another early vertebrate

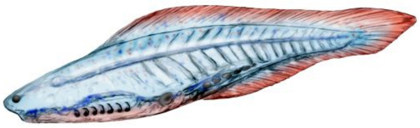
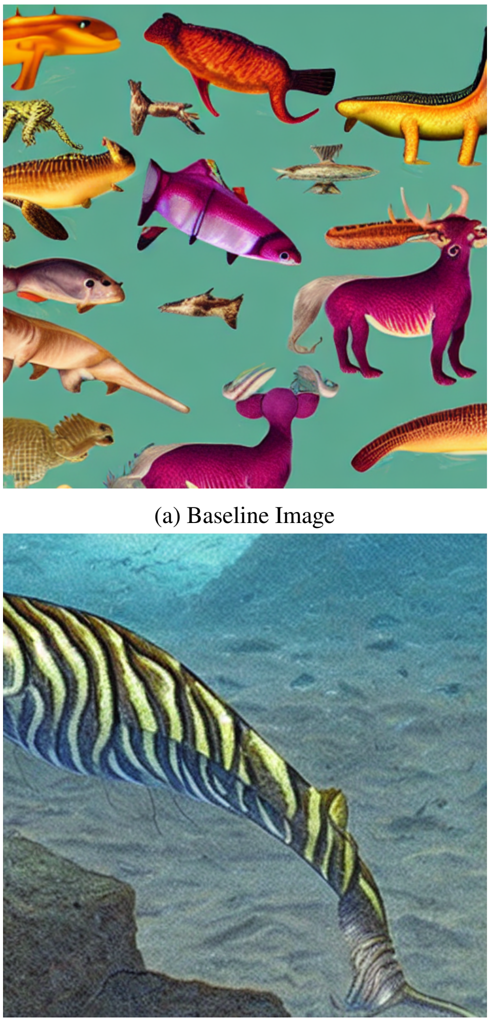


Figure 4: Ground truth image.

F.2 Example 2

Intent: Vikings - Trade

Text: The Vikings established and engaged in extensive trading networks throughout the known world and had a profound influence on the economic development of Europe and Scandinavia. Other than in such trading centres as Ribe Hedeby in Denmark, Scandinavia was unfamiliar with the use of coinage, therefore its economy was based on bullion, that is, the purity and weight of precious metals used in exchange. Silver was the precious metal most commonly used, although gold was also used. Traders carried small portable scales, enabling them to measure weight precisely, which allowed an accurate medium of exchange, even lacking a regular coinage. Refer Figure 6 and 7.



(a) Baseline Image

(b) MM-PAW Image

Figure 5: Images generated for section on Vertebrates - First Vertebrates

Plan

1. Introduction to Viking trade
2. Viking trading networks and their reach
3. Commodities traded by the Vikings (silver, gold, and other goods)
4. Methods of trade (bullion, scales, lack of coinage)

Image Descriptions

1. A Viking longship sailing on the water
2. A Viking market with traders selling products
3. Excavations showing artifacts such as jewelry and coins

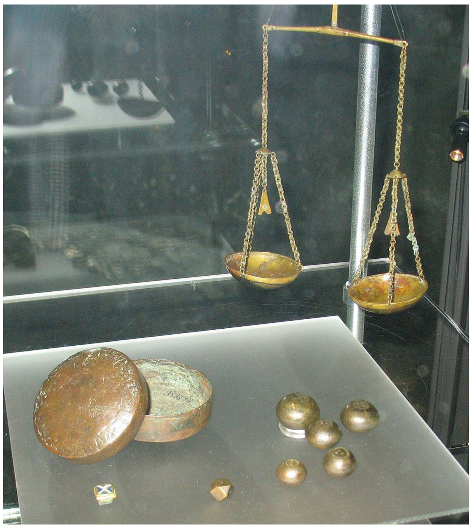


Figure 6: Ground truth image for ’Vikings - Trade’.

G Human Evaluation Details

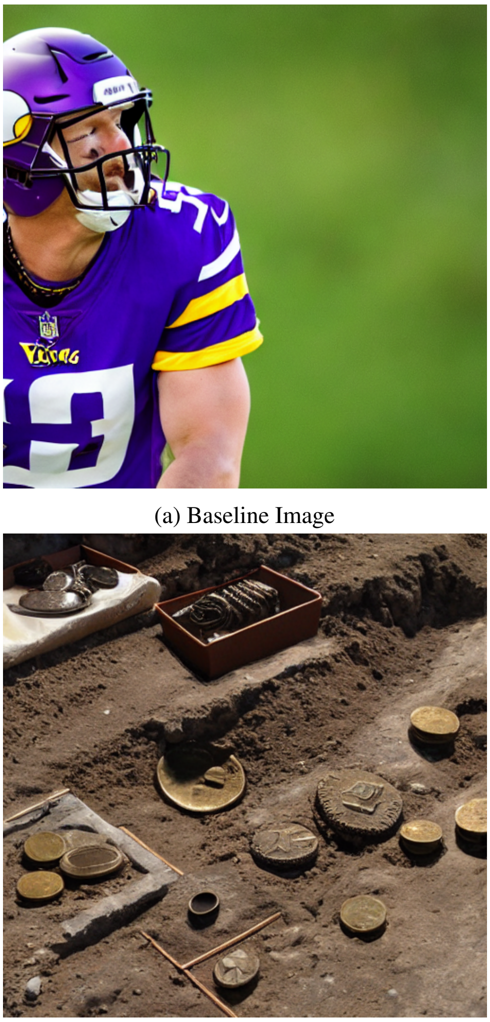
To assess the quality of generated outputs concerning alignment with intent and coverage, we conducted human evaluations using annotations from three annotators sharing a similar background (Indian origin, above undergraduate studies) and proficiency in English. Volunteers were found via word of mouth.

For the evaluation of Plan-And-Write (PAW), annotators were presented with 20 examples, each featuring a section title, outputs from our model and a GPT-based baseline (in a random order), along with ground truth references. Annotators were instructed to compare model outputs based on relevance to intent, coverage, and overall structure. No specific guidelines were given, allowing annotators to form their own perspectives on coverage and well-formed content. The survey comprised two parts with 10 questions each, taking an average of 27 minutes for completion.

Questions included:

1. Which output is more aligned/relevant to the given intent?
2. Which output has greater coverage of the topics mentioned in the ground truth?
3. Which output has the most well-formed content generation?

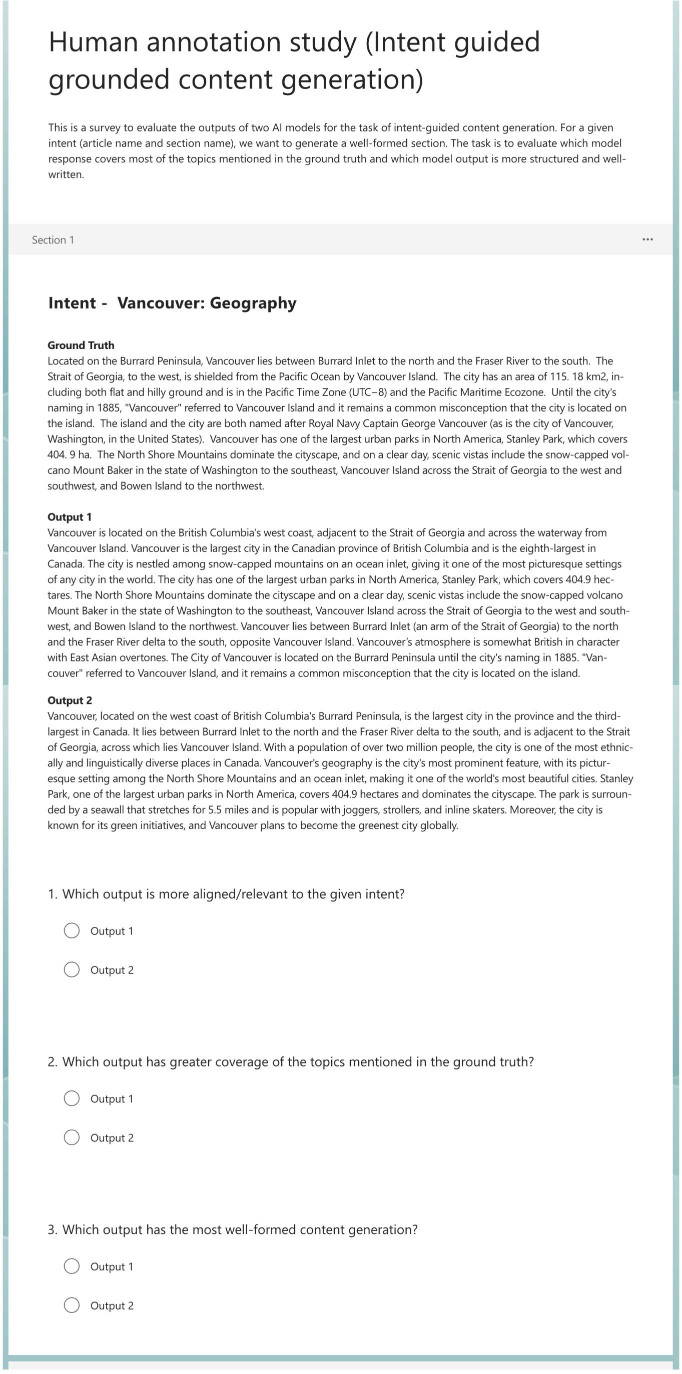
In the evaluation of Multimodal Plan-And-Write (MM-PAW), annotators were presented with 20 examples, each featuring a section title, ground truth text, and images from the baseline and MM-PAW. Annotators were asked a single question regarding the relevance of images to the given section, with the exclusion of ground truth images to mitigate potential biases. This approach aimed to specifically evaluate the effectiveness of multimodal content generation in MM-PAW. The survey took an average of 7.5 minutes for completion of 20 questions.



(a) Baseline Image

(b) MM-PAW Image

Figure 7: Images generated for ’Vikings - Trade’.



Human annotation study (Intent guided grounded content generation)

This is a survey to evaluate the outputs of two AI models for the task of intent-guided content generation. For a given intent (article name and section name), we want to generate a well-formed section. The task is to evaluate which model response covers most of the topics mentioned in the ground truth and which model output is more structured and well-written.

Section 1 ···

Intent - Vancouver: Geography

Ground Truth
Located on the Burrard Peninsula, Vancouver lies between Burrard Inlet to the north and the Fraser River to the south. The Strait of Georgia, to the west, is shielded from the Pacific Ocean by Vancouver Island. The city has an area of 115. 18 km2, including both flat and hilly ground and is in the Pacific Time Zone (UTC−8) and the Pacific Maritime Ecozone. Until the city's naming in 1885, "Vancouver" referred to Vancouver Island and it remains a common misconception that the city is located on the island. The island and the city are both named after Royal Navy Captain George Vancouver (as is the city of Vancouver, Washington, in the United States). Vancouver has one of the largest urban parks in North America, Stanley Park, which covers 404. 9 ha. The North Shore Mountains dominate the cityscape, and on a clear day, scenic vistas include the snow-capped volcano Mount Baker in the state of Washington to the southeast, Vancouver Island across the Strait of Georgia to the west and southwest, and Bowen Island to the northwest.

Output 1
Vancouver is located on the British Columbia's west coast, adjacent to the Strait of Georgia and across the waterway from Vancouver Island. Vancouver is the largest city in the Canadian province of British Columbia and is the eighth-largest in Canada. The city is nestled among snow-capped mountains on an ocean inlet, giving it one of the most picturesque settings of any city in the world. The city has one of the largest urban parks in North America, Stanley Park, which covers 404.9 hectares. The North Shore Mountains dominate the cityscape and on a clear day, scenic vistas include the snow-capped volcano Mount Baker in the state of Washington to the southeast, Vancouver Island across the Strait of Georgia to the west and southwest, and Bowen Island to the northwest. Vancouver lies between Burrard Inlet (an arm of the Strait of Georgia) to the north and the Fraser River delta to the south, opposite Vancouver Island. Vancouver's atmosphere is somewhat British in character with East Asian overtones. The City of Vancouver is located on the Burrard Peninsula until the city's naming in 1885. "Vancouver" referred to Vancouver Island, and it remains a common misconception that the city is located on the island.

Output 2
Vancouver, located on the west coast of British Columbia's Burrard Peninsula, is the largest city in the province and the third-largest in Canada. It lies between Burrard Inlet to the north and the Fraser River delta to the south, and is adjacent to the Strait of Georgia, across which lies Vancouver Island. With a population of over two million people, the city is one of the most ethnically and linguistically diverse places in Canada. Vancouver's geography is the city's most prominent feature, with its picturesque setting among the North Shore Mountains and an ocean inlet, making it one of the world's most beautiful cities. Stanley Park, one of the largest urban parks in North America, covers 404.9 hectares and dominates the cityscape. The park is surrounded by a seawall that stretches for 5.5 miles and is popular with joggers, strollers, and inline skaters. Moreover, the city is known for its green initiatives, and Vancouver plans to become the greenest city globally.

1. Which output is more aligned/relevant to the given intent?

○ Output 1

○ Output 2

2. Which output has greater coverage of the topics mentioned in the ground truth?

○ Output 1

○ Output 2

3. Which output has the most well-formed content generation?

○ Output 1

○ Output 2

Figure 8: Screenshot of Microsoft Forms used for survey.