

Zooming in on Zero-Shot Intent-Guided and Grounded Document Generation using LLMs

Pritika Ramu¹ Pranshu Gaur^{2†} Rishita Emandi^{3†}
Himanshu Maheshwari^{4‡} Danish Javed^{5†} Aparna Garimella¹

¹ Adobe Research, Bangalore, India
^{2,3,5} Indian Institute of Technology, {Kanpur, Madras, Delhi}

⁴ Microsoft, India
{pramu,garimell}@adobe.com

Abstract

Repurposing existing content on-the-fly to suit author’s goals for creating initial drafts is crucial for document creation. We introduce the task of intent-guided and grounded document generation: given a user-specified intent (*e.g.*, section title) and a few reference documents, the goal is to generate section-level multimodal documents spanning text and images, grounded on the given references, in a zero-shot setting. We present a data curation strategy to obtain general-domain samples from Wikipedia, and collect 1,000 Wikipedia sections consisting of textual and image content along with appropriate intent specifications and references. We propose a simple yet effective planning-based prompting strategy *Multimodal Plan-And-Write (MM-PAW)*, to prompt LLMs to generate an intermediate plan with text and image descriptions, to guide the subsequent generation. We compare the performances of MM-PAW and a text-only variant of it with those of zero-shot Chain-of-Thought (CoT) using recent close and open-domain LLMs. Both of them lead to significantly better performances in terms of content relevance, structure, and groundedness to the references, more so in the smaller models (upto 12.5 points ↑ in Rouge 1-F1) than in the larger ones (upto 4 points ↑ R1-F1). They are particularly effective in improving relatively smaller models’ performances, to be on par or higher than those of their larger counterparts for this task.

1 Introduction

Recent advances in generative models (Brown et al., 2020; Ramesh et al., 2021; Blattmann et al., 2022; Touvron et al., 2023) have enabled the creation of high-quality textual and visual content through natural language prompts. Techniques like Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al.,

[†] Work done while interning at Adobe Research

[‡] Work done while working at Adobe Research

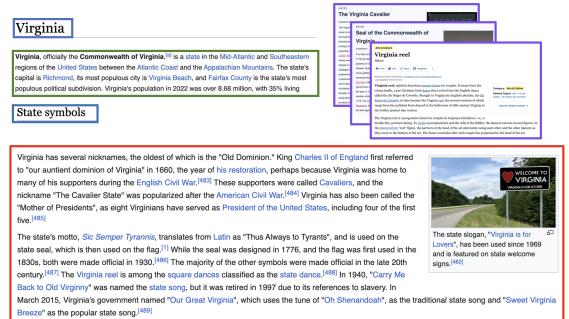


Figure 1: Example¹ of intent-guided and grounded document generation; Input is **intent** (Wikipedia article name and section name), **initial context** and **reference articles**. Output is **multimodal content**.

2023) have improved LLMs’ performance across NLP tasks, including question answering (Tafjord et al., 2022; Yoran et al., 2023), reasoning (Wang et al., 2023a), summarization (Wang et al., 2023b), and conversation generation (Lee et al., 2023).

Document creation can be a creative process; while the content itself may or may not always be unique, the goal or *intent* of each document can be very specific to the user’s needs. It typically involves reusing and piecing together portions of content from multiple sources to create a rich *first draft* based on the intent, and then iteratively edit it until it reaches a suitable final stage. Figure 1 illustrates this scenario of creating a Wikipedia section; the author aims to create a first draft for a specific section using a few reference articles. In such scenarios, zero-shot generation of first draft can provide a strong starting point, and save the time and effort of content creators creating general-domain documents such as marketing blogs, reports, etc.

In this paper, we **introduce** *intent-guided and grounded long document generation* in zero-shot setting, with three constraints: (*i*) documents are to be generated from the given reference documents

¹Example obtained from Wikipedia (Virginia). Reference articles depicted: Virginia Cavalier, Seal of Commonwealth, Virginia Reel

and an intent specified by the user; (*ii*) documents can be multimodal in nature with text and image content; and (*iii*) the generation is to be on-the-fly for any given intent with a few source documents and no additional training data. We present a data curation strategy to obtain general-domain Wikipedia samples, and **curate** an evaluation set comprising of 1,000 sections along with the corresponding intents and external references using XML parsing and Bing Search.

Grounding has been a well-known paradigm in natural language generation wherein some source content is used to condition the generation (Narayan et al., 2018; Prabhumoye et al., 2019). However, most grounded generation works focused on short texts (Prabhumoye et al., 2019), whereas our focus will be on long documents ranging over several sentences. Further, most document generation works are limited to text-only generation; while text-to-image models (Ramesh et al., 2021; Blattmann et al., 2022) like Dall-E generate high-quality images from textual prompts, automatically determining the appropriate textual and visual composition of a document based on an intent and references remains underexplored.

Inspired by the superior performances of LLMs in zero-shot settings (Wang et al., 2023a; Saha et al., 2024), we **propose** a zero-shot prompting strategy that infuses *content planning* as an intermediate step in the generation task. Our pipeline comprises of a retriever module to retrieve the relevant content from the given references based on the intent, followed by an LLM prompting module to plan and synthesize the output. Specifically, we propose Multimodal Plan-And-Write (MM-PAW) prompting, to generate multimodal plans comprising of text topics and image-specific descriptions, based on intent and retrieved content, and condition the text generation on the generated plan. Appropriate images are generated using image descriptions using text-to-image models.

We compare MM-PAW and a text-only variant of it (PAW) (for multimodal and text-only section generation respectively) with Zero-Shot CoT using 8 close and open-source LLMs. We note improvements using our prompting variants in terms of content relevance and coverage while maintaining groundedness. Specifically, they improve the smaller models’ performances to be on par with or higher than those of their larger counterparts, indicating the effectiveness of our approach in utilizing smaller models to perform the task comparable to

the larger ones. To our knowledge, this is the first study on grounded multimodal document generation using LLMs.

2 Related Work

Grounded document generation. Grounded text generation has been receiving increasing attention (Prabhumoye et al., 2019, 2021; Iv et al., 2022; Brahman et al., 2022), as it leads to generation of more contentful outputs while not running into the risk of hallucinating irrelevant or factually incorrect concepts. Prabhumoye et al. (2019) introduced the task of grounded content transfer, to infuse content from an external source to generate a follow-up sentence of an existing document. Iv et al. (2022) addressed the task of updating existing textual content based on new evidence, so as to make the given input text consistent with new information. Brahman et al. (2022) addressed the task of generating a factual description about an entity given a set of guiding keys and grounding passages. Another popular task following this paradigm is abstractive summarization (Narayan et al., 2018) in which the generation should capture the most salient information from a given source. We aim to generate longer texts going beyond single sentence additions, and take as input only reference documents for grounding and a user-provided intent, without any additional form of guidance. Further, we aim to generate Wikipedia-style documents composed of text and images. We believe this scenario is closer to real-world document creation, and an instant first draft kickstarts the creation process. Further, unlike in the summarization task, our the input references contain lot more noise which is be filtered out based on the given intent to generate the output.

Plan-based generation. Content planning has been a widely studied topic in natural language generation tasks (Kang and Hovy, 2020; Goldfarb-Tarrant et al., 2020; Jansen, 2020; Chen et al., 2021), as they assist in enforcing coherence, structure, and logical consistency for longer text generation. Kang and Hovy (2020) addressed paragraph completion by first predicting key words for the missing content, and using them to generate the sentences. Chen et al. (2021) focussed on planning a sequence of events using event graphs to guide a story generator. Narayan et al. (2021) use ordered sequences of entities to ground the summary generation. More recently, planning-based approaches

Dataset	Source document(s) length (words / sentences)	Target length (words / sentences)	% Novel n-grams (in source not in the target)			
			Unigrams	Bigrams	Trigrams	4-grams
CNN	760.50 / 33.98	45.70 / 3.59	65.76	93.48	96.82	97.99
DailyMail	653.33 / 29.33	54.65 / 3.86	66.89	94.23	97.71	98.14
Our Dataset	22,922.21 / 876.79	357.75 / 15.44	93.67	97.13	98.14	98.45

Table 1: Statistics of our dataset in comparison with those of a few existing summarization datasets (average stats).

to better prompt large language models have been gaining attention (Kang and Hovy, 2020; Hu et al., 2022; Li et al., 2022). Wang et al. (2023a) proposed zero-shot plan-and-solve prompting for multi-step reasoning tasks. Wang et al. (2023b) used planning in summarization using LLMs, by first prompting them to answer a few elemental questions and using them to generate the summaries step by step. We extend the concept of planning to prompt LLMs in a zero-shot manner to generate *multimodal plans* providing cues on the preferred textual and visual composition of output, and ground the subsequent generation on them.

3 Task Setup & Dataset

Writer’s block is a major challenge for content creators, which can affect their productivity and creativity while creating new content. However, document creation seldom starts from scratch, and obtain rough first drafts and revising them can enhance the writing abilities of creators (Lamott, 1995). We study the task of automatically providing a rich multimodal first draft that aligns to author’s goals, while reusing relevant information from across different related sources, which they can further iterate upon to create their final version. We study this task in a zero-shot setup without any fine-tuning, and investigate the capabilities of LLMs to generate content on any given topic provided a few references to it.

There do not exist datasets tailored for our task. We find Wikipedia as the most suited source due to the following reasons: (i) We can view the various section titles as intents, and the citations can act as the external references to create a given section; and (ii) Wikipedia articles have text and image content, where the images contain content related to specific concepts in the text. Wikipedia is increasing being used as a source for various tasks (Qian et al., 2023); however, they do not provide multimodal articles with images along with text.

Data Scraping. We obtain samples by scraping articles from Wikidump.² We use Pywikibot Python library to parse the Wikipedia pages. “Text” is an attribute of “Page” that provides the text content of a Wikipedia page in wiki markup format. Sections are demarcated by “==” tags before and after the section heading; we use this information to extract headings (as intents) and corresponding textual content for each section. Reference links used in the section are found within `<ref>` tags in the wiki markup format. Images present within a section are indicated by their file names in the format `[[File:*`image file name*`|...]] or `[[Image:*`image file name*`|...]]. They are downloaded by identifying their corresponding URLs in the HTML version of Wikipedia articles using BeautifulSoup. This process helps us to curate multimodal sections including text and images, along with the intents and reference links. Some of the images are not grounded to any topic in the corresponding text in a few sections, as it is common in Wikipedia articles. To ensure that images are grounded to some concepts in the text, we calculate the CLIPScore (Hessel et al., 2021) between each sentence in the section and the corresponding section image(s), and filter out sections that have image relevance score below a threshold (manually set at 0.31 using a small validation set).

It is worth noting that the accessibility of every extracted reference link (citation) is not guaranteed (503 error). Also, there is no assurance that web scrapers are permitted to gather content from these sources (403 error). Many references are in the form of PDFs (from Google Books, journals, etc.), videos, audios or inaccessible links (404 errors). Due to this, several links are discarded, due to which the corresponding source content to generate the sections would be missing. To overcome this issue, we use the Bing Search API³ to curate reference articles. Each sentence in a section is

²<https://dumps.wikimedia.org>

³<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

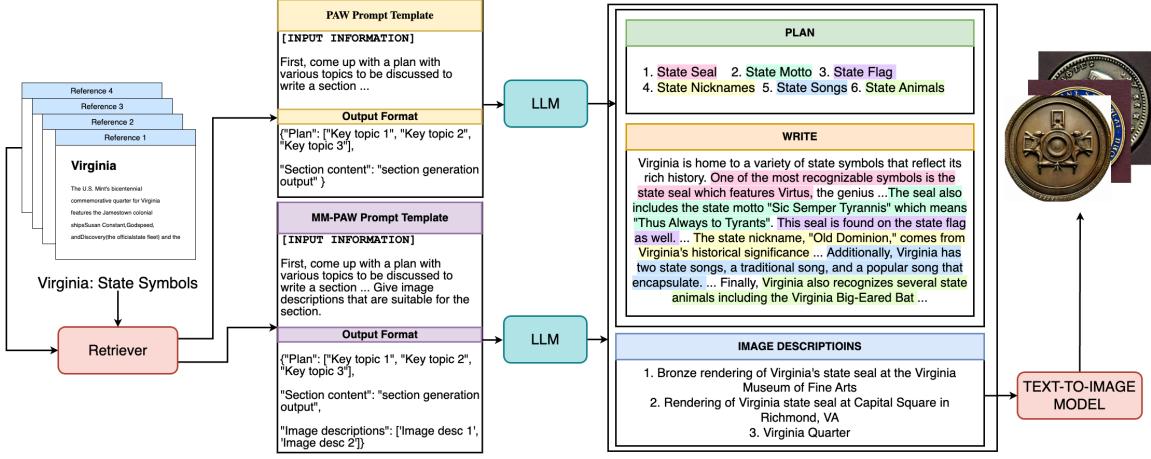


Figure 2: Intent-guided document generation pipeline: Sentences are retrieved based on intent and reference articles. The MM-PAW prompt is filled and sent to an LLM for document content generation.

used as a search query, and allowing us to retrieve relevant web pages for the entirety of the section. We parse content exclusively from pages permitting bot scraping. We curate 1,000 multimodal sections with intents and references as our evaluation set, respecting copyright and intellectual property rights. The content obtained from these websites belongs to the respective owners or authors. The resulting sections cover a wide range of topics, including Science, History, Government, Art, Health, Technology, Culture, Education, Sports, Economy, among others.

Table 1 presents a few statistics on our dataset. On an average, there exist 7.36 reference articles for each section. The average word count for the references put together is as high as 23K compared to just 357.75 words in the generated sections. This vast discrepancy in length indicates that the sections are not merely condensed versions of the references but rather selective extractions from them, and that the references also contain a lot of noise which is to be filtered out when creating the sections. This is further seen in the high percentage of novel n-grams in the references compared to the target sections in our dataset, indicating that a large amount of the content is not used to create the section. On the contrary, summarization typically requires a more proportional reduction in content length, where the summary still encompasses all key elements of the original text.

4 Method

Our pipeline follows the retrieve-and-generate paradigm (Lazaridou et al., 2022; Qian et al., 2023) and consists of two stages, namely intent-guided

content extraction and document generation (Figure 2). In the first stage, we perform query-based sentence retrieval to extract relevant sentences from the reference articles, using the given intent (section title) as the query. We use SBERT embeddings (Reimers and Gurevych, 2019) to encode reference sentences and employ FAISS (Johnson et al., 2019) to perform fast semantic search by indexing these embeddings. We compute the similarity of the intent with the indexed sentences, and the top- k sentences are selected. In the second stage, we incorporate the intent and the retrieved sentences in our zero-shot prompt template namely Multimodal Plan-And-Write (MM-PAW) to prompt an LLM. The order of retrieved sentences in the prompt is in order of semantic similarity (cosine similarity) with the given intent.

Multimodal Plan-And-Write. Planning is a very effective paradigm in generation to first obtain a high-level overview of the content to be generated, and ground the subsequent generation on the inferred intermediate plan. While LLMs by themselves can generate high-quality text, we probe them to come up with text-based multimodal plans to provide cues on the topics to be discussed in the text and descriptions for any images that can visually illustrate specific concrete concepts in the text. Specifically, we prompt the LLM to generate such multimodal plan based on the intent and given reference sentences, and use it to ground the text generation for the section. We also provide a desired length specification for the output section, based on the ground truth section length ($0.8n <$ desired length $< 1.2n$ where n is the number of

tokens in the ground truth section), for a fair comparison. The textual content is generated by the LLM conditioned on the text plan, while we use the image description(s) to prompt a text-to-image model (Blattmann et al., 2022) to get the accompanying image(s), as opposed to using the retrieved sentences or generated text, which will exceed their context limit, or just the intent which will be too generic. The prompt format looks like below:

```

Instruction:
Intent:
Retrieved sentences:
Output (json):
{
    "Text plan": <Key topics to be
    present in the text>,
    "Text output": Section text
    with <min> and <max length>,
    "Image plan": Description(s) of
    image(s) to accompany the text
}

```

To generate text-only sections, we use Plan-And-Write, a variant of MM-PAW that does not generate image descriptions, and only generates the text plan followed by textual section content. The PAW and MM-PAW prompt templates are provided in Appendix A.

5 Experiments

We conduct our experiments using two close-source and two open-source family of LLMs, namely Claude (claude-3-Haiku) (Anthropic, 2024), GPT (gpt-4, gpt-35-turbo) (Brown et al., 2020), LLaMa (fine-tuned chat 70B, 13B, 7B models) (Touvron et al., 2023), and Mistral (7B, 8x7B) (Jiang et al., 2023). We use NVIDIA A100 GPUs to perform inference with the LLaMa and Mistral variants. For intent-based sentence retrieval, we set $k = 150$ using fast semantic search for all the experiments, so as to accommodate for the context length limits of LLaMa and Mistral models.⁴ We use the Stable-Diffusion-v1-5 checkpoint (Blattmann et al., 2022) to generate images. In order to have a fair comparison, a length constraint is enforced in the prompt template so as to ensure that the generation and the ground truth are of similar lengths. The expected range of words to be produced is defined as [0.8, 1.2] times the number of words in the ground truth. Results are averaged

⁴We note that 150 sentences approximate to 3K tokens on an average across the reference articles.

across 5 runs with different seeds. Standard deviation of the runs are provided in Appendix D.

Baselines. The instructions to the LLMs are minimal in the baseline setup. The LLMs are prompted to generate coherent section text using the intent and retrieved sentences along with the length specification. The intent itself used as the text prompt to generate images using the text-to-image model. The baseline prompt is provided in Appendix B.

Evaluation Metrics. We evaluate the different variants on five dimensions namely, text relevance, text coverage, text groundedness with respective to the references, text structure, and image relevance. We use a mixture of traditional metrics and LLM-based one for each of these aspects. We use Rouge precision as an approximation to text relevance, Rouge recall to approximate the coverage of the resulting text output, and Rouge F1 as overall measure, and use the ground truth sections as references (Lin, 2004). We also use G-Eval (Liu et al., 2023b), a GPT-4-based evaluation measure, to assess the overall relevance and coverage aspects with reference to the ground truth on a scale of 1-5. For groundedness, we aim measure the extent to which the reference sentences support the generated text. For this, we use a Natural Language Inference (NLI) model RoBERTa Large (Liu et al., 2019) which is fine-tuned on the Multi-Genre NLI corpus (Williams et al., 2018). We compute the average number of sentences in the generated text that are entailed by at least one reference sentence using the model. In addition, we use a G-Eval variant to assess this on a scale of 1-5 given all the reference and generated sentences. For structure, we use G-Eval to assess the fluency and coherence of the generated text on a scale of 1-5. All the G-Eval prompts are presented in Appendix C. For image relevance, we use ClipScore (Hessel et al., 2021) to compute the cosine similarity between the generated and ground truth images. In the case of more than one generated or ground truth image, we take the maximum similarity scores for each of them and provide an average across them. Additionally, we report human ratings to verify our approach.

6 Results & Discussion

Table 2 presents a comparison of the results of both of our prompting variants against the baselines. For most of the models, we note that PAW and MM-PAW lead to increased performances in terms

	TXT. REL.			COVERAGE			OVERALL				GROUNDING		STRUCTURE	IMG. REL.	
	PRECISION			RECALL			F1								
METHOD	R1	R2	RL	R1	R2	RL	R1	R2	RL	G-EVAL	NLI	G-EVAL	G-EVAL	CLIPSCORE	
BL GPT-4	50.49	17.82	26.59	35.26	13.79	19.24	41.52	15.55	23.33	3.37	10.35	4.83	3.13	60.82	
PAW	51.38	18.85	27.23	41.45	15.83	21.14	45.88	17.21	23.80	4.02	11.47	4.76	3.67	-	
MM-PAW	55.78	20.17	29.28	39.62	16.52	20.39	46.33	18.16	24.04	4.36	10.75	4.72	3.67	69.95	
BL Claude (Haiku)	52.95	18.25	27.73	37.33	14.79	20.28	43.79	16.34	23.43	3.87	10.85	5.33	3.63	60.82	
PAW	53.93	19.34	28.38	43.64	16.38	22.84	48.24	17.74	25.31	4.52	11.97	4.76	3.78	-	
MM-PAW	56.38	21.74	30.36	40.84	17.72	21.38	47.37	19.53	25.09	4.86	11.75	4.74	3.70	70.45	
BL GPT-3.5	47.81	16.00	23.37	34.02	12.44	17.32	39.75	14.00	19.90	2.87	9.75	4.33	2.63	60.82	
PAW	47.99	16.99	24.90	41.69	14.90	20.90	44.62	15.88	22.73	3.52	10.47	4.74	3.28	-	
MM-PAW	50.87	18.36	26.64	35.72	12.14	18.05	41.97	14.62	21.52	3.36	9.75	4.67	3.26	69.45	
BL LLaMa 2 (70B)	34.68	7.82	22.14	24.70	7.34	12.72	28.85	7.57	16.16	2.12	8.98	4.12	1.97	60.82	
PAW	36.62	10.78	18.82	41.00	12.73	20.91	38.69	11.67	19.81	3.24	10.45	4.74	3.16	-	
MM-PAW	37.98	11.13	22.67	31.35	9.62	16.21	34.35	10.32	18.55	3.16	9.33	4.33	3.11	65.52	
BL LLaMa 2 (13B)	28.81	5.13	14.04	19.69	5.94	9.61	23.39	5.51	11.41	1.97	6.34	3.54	1.62	60.82	
PAW	33.57	8.18	16.02	38.11	9.93	17.21	35.70	8.97	16.59	2.78	8.02	3.63	2.99	-	
MM-PAW	34.83	8.98	19.88	29.14	8.12	13.93	31.73	8.53	16.38	3.07	7.98	3.56	2.98	64.32	
BL LLaMa 2 (7B)	24.19	4.18	11.91	13.71	4.33	7.78	17.50	4.25	9.41	1.83	6.01	2.99	1.55	60.82	
PAW	28.13	4.77	14.12	21.26	5.85	11.88	24.22	5.25	12.90	2.56	7.66	3.12	2.13	-	
MM-PAW	29.81	6.92	17.42	20.13	5.29	10.53	24.03	5.99	13.13	2.96	7.54	3.03	2.10	62.19	
BL Mixtral (8x7B)	35.92	8.12	24.88	26.09	8.88	14.22	30.23	8.48	18.10	2.23	9.01	4.12	1.98	60.82	
PAW	38.29	11.18	27.97	41.47	12.98	21.29	39.82	12.01	24.18	3.37	10.47	4.76	3.23	-	
MM-PAW	38.33	11.19	29.91	31.98	9.55	17.73	34.87	10.31	22.26	3.26	9.58	4.56	3.23	66.67	
BL Mistral (7B)	28.75	5.07	13.86	18.99	5.87	9.57	22.87	5.44	11.32	1.97	6.27	3.54	1.57	60.82	
PAW	33.37	7.96	15.93	37.68	9.44	16.89	35.39	8.64	16.40	2.67	7.86	3.57	2.87	-	
MM-PAW	34.76	7.58	19.01	28.28	8.03	13.77	31.19	7.80	15.97	3.08	7.96	3.54	2.78	63.84	

Table 2: PAW and MM-PAW results. R1, R2, RL depict ROUGE-1, ROUGE-2, ROUGE-L respectively.

of the overall text quality (Rouge F1 and G-Eval overall). These improvements are more notable in smaller models such as Mistral 7B, LLaMa 2 7B, and LLaMa 2 13B (upto \uparrow 12.52 R1-F1) compared to those in the larger ones (upto \uparrow 4.8 R1-F1). Further, we note that a given smaller model’s performance using our prompting variants approximates or increases over that of its larger counterpart. That is, PAW-LLaMa 2 7B has higher Rouge F1 scores compared to those of BL LLaMa 2 13B; similarly, PAW-LLaMa 2 13B has higher Rouge F1 scores compared those of BL LLaMa 2 7B; and PAW-GPT-3.5 has higher scores compared to both BL GPT-4 and BL Claude. This indicates that using our prompting variant is able to improve the generation quality of a relatively smaller LLM with lower performance over a larger one which may have higher latency and/ or cost implications.

On an average, the improvements of our variants over the baselines in terms of text coverage (recall) are higher than those for relevance (precision). Given the retrieved sentences as input, we believe the baseline models’ selection of relevant details may not result in a good coverage of relevant topics. This challenge arises from the complex and under-specified dependency between a short intent (the section heading) and retrieved reference sentences, making it more challenging for language models to accurately capture, as highlighted in (Li et al.,

2016; Fan et al., 2018). Our proposed approach formulates a high-level topic-based plan first, providing the model with an intermediate overview of the references’ diverse aspects, thereby increasing coverage.

Interestingly, between PAW and MM-PAW, we note that the former has higher coverage and overall scores for text generation, while MM-PAW has slightly higher relevance values. We speculate that including image-specific details in the multimodal plan may have made the topics more “accurate” and aligned with the themes in the image descriptions, thereby increasing relevance (and reducing coverage). We believe further investigation would help in understanding the interplay between the text and image quality while generating multimodal content and text-only subset of it.

In terms of groundedness and structure, our proposed variants result in improved scores for both NLI and G-Eval in most cases, indicating our outputs are more grounded to the references compared to baselines. Finally, the images generated using the multimodal plans result in more relevant outputs as seen in the ClipScore compared to using only the intent or high-level section details with the baseline approaches. We note that the improvements over the baselines are higher in larger model variants (such as Claude, GPT-4, and GPT-3.5) compared to the smaller ones, indicating their su-

GT	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 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..., small eel-like conodonts, are known from microfossils of their paired tooth segments from the late Cambrian to the end of the Triassic.	
Baseline	Vertebrates emerged during the Cambrian explosion. These primitive vertebrates likely possessed simple skeletal structures ... The feeding habits of early vertebrates were ... The earliest known vertebrates lacked conventional jaws, relying on filter feeding near the seabed. Although limited fossil evidence makes it difficult to precisely determine their physical characteristics,...	
MM-PAW	Plan: ['Cambrian explosion lead to rise in organism diversity', 'Discuss the earliest known vertebrates - Myllokunmingia, Haikouichthys ercaicunensis', 'Mention vertebrate body structure', 'Discuss jawless vertebrates and their eating habits'] Image Descriptions: ['Illustration of the Myllokunmingia, the earliest known vertebrate'] Vertebrates appeared during the Cambrian explosion... These species displayed vertebrate traits with a notochord, rudimentary vertebrae, and a distinct head and tail. They lacked conventional jaws and primarily engaged in filter feeding near the seabed. While their exact forms remain elusive due to sparse fossil evidence, these early... Conodonts, eel-like vertebrates are evidenced by microfossils of their tooth segments.	

Table 3: Sample output of MM-PAW and the GPT baseline on “Vertebrate - First Vertebrate”. The textual content that is relevant to the groundtruth are highlighted in blue. Our generated image is more similar to the ground truth one.

perior ability to plan for content beyond textual modality.

We conduct an ablation study comparing the performances of the models as the length of the text generation increases (Figure 3). We note that the improvements of our variant over the baseline are intact with increasing length. Further, we note that the baselines’ performances degrade slightly with the increasing lengths, whereas models with our prompting variant in general remain robust to length variations.

Tables 3 and 4 show two qualitative examples along with their generated plans; the textual content has higher topical coverage and the image by our approach is more relevant. Please refer to Appendix E and F for more examples.

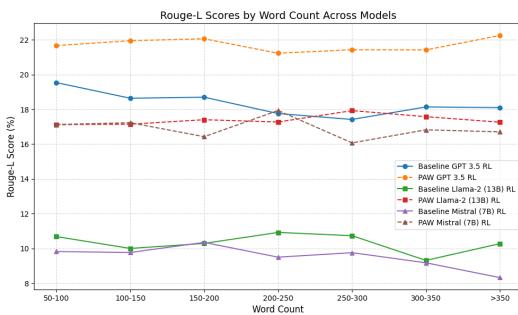


Figure 3: Comparison of Rouge-L scores of Baseline and PAW across varying word count.

Human evaluation. We conduct surveys to assess the alignment with the intent, topical coverage, and well-formedness of the generated text with re-

spect to the ground truth. We ask 3 annotators of similar backgrounds (Indian origin, above undergraduate studies). They are provided with 20 examples each, along with the section titles, ground truth text and image content, and the outputs from the GPT-based baseline and the MM-PAW variant.

In terms of the alignment of the text generation with the given intent (section title), our outputs are marked better than the baseline in 85% cases; for topical coverage, 90% our outputs are rated better than the baseline outputs, and for well-formedness, 80% our outputs are rated better. The inter-rater reliability for the three surveys are 0.28, 0.39, 0.30 respectively (Krippendorff, 1970), indicating a decent agreement among the workers particularly in subjective tasks such as text generation, and comparable to those in other subjective tasks (Hossain et al., 2019; Garimella et al., 2020, 2021). For image relevance with respect to the ground truth images, 85% MM-PAW-based generations are rated better than the baseline images ($\alpha = 0.79$), supporting the effectiveness of our approaches for multimodal document generation based on given intent and references (full details in Appendix G).

7 Conclusions

We introduced the task of intent-guided and grounded long document generation, and planning-based prompting strategies to come up with multimodal plans while generating documents. We curate a dataset for evaluating our approaches with respect to baseline LLMs by scraping multimodal

GT	The Vikings established and engaged in extensive trading networks ... 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.	
Baseline	The Vikings established extensive trading networks that significantly influenced the economy of Europe and Scandinavia. This also facilitated the exchange of goods, ideas and technologies even into Asia and Middle East. In trading centers like Ribe and Hedeby in Denmark, the economy relied on bullion, with silver being the most commonly used metal. Barter system is another form of trade for exchanging goods and services directly without the use ...	
MM-PAW	Plan: ['Introduction to Viking trade', 'Viking trading networks and their reach', 'Commodities traded by the Vikings (silver, gold, and other goods)', 'Methods of trade (bullion, scales, lack of coinage)'] Image Descriptions: ['Excavations showing artifacts such as jewelry and coins'] Viking trades were instrumental ... Extensive trading networks significantly influenced the economic development of Europe and Scandinavia. ... In trading centers like Ribe and Hedeby in Denmark, coinage was uncommon. The economy relied on bullion where precious metals are measured by purity and weight. Silver was the most commonly ... The Vikings traded a variety of goods including furs, textiles, weapons, spices ... Traders carried portable scales to precisely measure weight, ensuring accurate exchanges even without regular coinage.	

Table 4: Sample output of MM-PAW and the GPT baseline on “Vikings - Trade”. The textual content that is relevant to the groundtruth are highlighted in blue. Our generated image is more similar to the ground truth one.

sections from Wikipedia. Our plan-based prompting significantly improves the topical coverage in the outputs, particularly in longer document generation. The multimodal plan enhances the content relevance for text and image outputs. We note that our variants are able to bring smaller models’ performances closer to their larger counterparts, or even surpass them by a significant margin. An AI assistant generating high-coverage outputs along with images based on a given intent can be particularly useful in providing a good first drafts in the creating documents.

Real-world documents span over several more modalities, such as tables, charts, infographics, etc. such planning strategies can be extended to provide cues on which modalities will appropriate to generate the content in, and select content for each of those modalities. We believe our work can provide a starting point for further explorations into grounded multimodal document generation.

8 Limitations and Future Work

While our plan-based prompting strategies increased the topical coverage, we note that sometimes may also include redundancy. While we provided initial insights into why this may happen, we believe studies are needed to examine this further.

It is known that Wikipedia data must be in the seen samples while pre-training these LLMs; we believe because we are comparing our variants with the base LLMs, this should not impact the improve-

ments brought about by our prompting variants.

Although our suggested methods show encouraging results in grounded and intent-guided document development, they also provide new directions for future study. As input, our current approach simply considers textual material. Given the recent progress made in multimodal understanding (Liu et al., 2023a), it is worthwhile to investigate the ways in which authors use various modalities, including tables, images, or videos, while creating documents. Moreover, while MM-PAW presents multimodal plans by combining visual descriptions with written plans, it is worthwhile to investigate the ways in which a plan might be extended to other modalities such as charts and tables. Furthermore, a trade-off between coverage (recall) and precision in document production algorithms is revealed by our comparison of PAW and MM-PAW. We need to explore flexible strategies to optimise this trade-off in accordance with user needs or desires.

References

- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.
- Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. 2022. [Retrieval-augmented diffusion models](#).
- Faeze Brahman, Baolin Peng, Michel Galley, Sudha Rao, Bill Dolan, Snigdha Chaturvedi, and Jianfeng Gao. 2022. [Grounded keys-to-text generation: Towards factual open-ended generation](#). In *Findings of the Association for Computational Linguistics: FTX-22*.

- EMNLP 2022*, pages 7397–7413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language models are few-shot learners.** In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. **GraphPlan: Story generation by planning with event graph.** In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 377–386, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. **Hierarchical neural story generation.** In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. **He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation.** In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, Online. Association for Computational Linguistics.
- Aparna Garimella, Carmen Banea, Nabil Hossain, and Rada Mihalcea. 2020. **“judge me by my size (noun), do you?” YodaLib: A demographic-aware humor generation framework.** In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2814–2825, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. **Content planning for neural story generation with aristotelian rescoreing.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. **CLIPScore: A reference-free evaluation metric for image captioning.** In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nabil Hossain, John Krumm, and Michael Gamon. 2019. **“President vows to cut <taxes> hair”: Dataset and analysis of creative text editing for humorous headlines.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. **PLANET: Dynamic content planning in autoregressive transformers for long-form text generation.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2305, Dublin, Ireland. Association for Computational Linguistics.
- Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. **FRUIT: Faithfully reflecting updated information in text.** In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3670–3686, Seattle, United States. Association for Computational Linguistics.
- Peter Jansen. 2020. **Visually-grounded planning without vision: Language models infer detailed plans from high-level instructions.** In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4412–4417, Online. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. **Mistral 7b.**
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Dongyeop Kang and Eduard Hovy. 2020. **Plan ahead: Self-supervised text planning for paragraph completion task.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6533–6543, Online. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.
- Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

- Anne Lamott. 1995. *Bird by bird: Some instructions on writing and life*. Vintage.
- Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *ArXiv*, abs/2203.05115.
- Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554, Toronto, Canada. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Qintong Li, Piji Li, Wei Bi, Zhaochun Ren, Yuxuan Lai, and Lingpeng Kong. 2022. Event transition planning for open-ended text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3412–3426, Dublin, Ireland. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *NeurIPS*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.
- Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online. Association for Computational Linguistics.
- Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. On zero-shot counterspeech generation by LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454, Torino, Italia. ELRA and ICCL.
- Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madiam Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. [Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. [Answering questions by meta-reasoning over multiple chains of thought](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5942–5966, Singapore. Association for Computational Linguistics.

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"

}

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_{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.

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

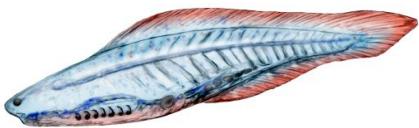


Figure 4: Ground truth image.

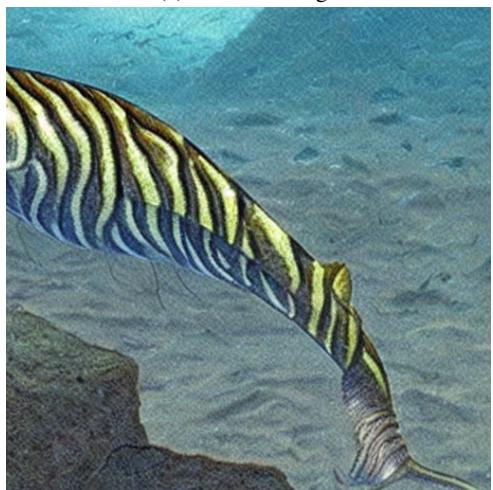
F.2 Example 2

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.



(a) Baseline Image



(b) MM-PAW Image

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

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.

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 Ecosystem. 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.