DOLOMITES: Domain-Specific Long-Form Methodical Tasks

Chaitanya Malaviya[P]*, Priyanka Agrawal[G], Kuzman Ganchev[G], Pranesh Srinivasan[G], Fantine Huot[G], Jonathan Berant[G], Mark Yatskar[P], Dipanjan Das[G], Mirella Lapata[G], Chris Alberti[G]

[P]University of Pennsylvania, USA [G]Google DeepMind, USA [G]Google DeepMind, UK

cmalaviy@seas.upenn.edu

Abstract

Experts in various fields routinely perform methodical writing tasks to plan, organize, and report their work. From a clinician writing a differential diagnosis for a patient, to a teacher writing a lesson plan for students, these tasks are pervasive, requiring to *methodically* generate structured long-form output for a given input. We develop a typology of methodical tasks structured in the form of a task objective, procedure, input, and output, and introduce DoLoMiTes, a novel benchmark with specifications for 519 such tasks elicited from hundreds of experts from across 25 fields. Our benchmark further contains specific instantiations of methodical tasks with concrete input and output examples (1,857 in total) which we obtain by collecting expert revisions of up to 10 model-generated examples of each task. We use these examples to evaluate contemporary language models, highlighting that automating methodical tasks is a challenging long-form generation problem, as it requires performing complex inferences, while drawing upon the given context as well as domain knowledge. Our dataset is available at https://dolomites-benchmark.github.io/.

1 Introduction

Experts in various fields regularly use writing as a means for planning, organizing, and sharing their work. For instance, a teacher might draft a lesson plan for what they would like to teach in their next class, and a lawyer might draft a patent application for an invention. Experts generally follow a consistent and methodical approach to conduct these writing tasks. In the lesson plan example, a teacher would know the lesson objectives, format, and profile of the class, and would produce a plan with the topics to be covered and activities to improve learning. Importantly, the teacher follows a systematic procedure to write this lesson plan, using their expertise and what they know about the current context (e.g., the class profile).

Across fields, from law to visual arts and engineering, experts accomplish on a regular basis such *methodical* tasks, i.e., writing tasks which loosely follow a standard template for what is usually given as input and what is required from the output. These tasks often follow a structured and consistent procedure as they are performed regularly and tend to be fairly time-consuming, taking from a few hours to several days (see Figure 2). As large language models (LMs) become more capable and widely accessible to a more sophisticated set of users (Owens, 2023; Mollick and Mollick, 2023; Lee et al., 2023; Birhane et al., 2023; Mollick and Mollick, 2023; Frankenreiter and Nyarko, 2022; Demszky et al., 2023; Wang et al., 2023), they hold great potential for assisting experts with methodical writing tasks, increasing their efficiency and allowing them to focus on complex problem-solving activities (Noy and Zhang, 2023).

Given their potential for assisting experts, it would be beneficial to evaluate language models on a realistic set of methodical writing tasks. However, we currently do not have benchmarks that contain a typology of such tasks. The most natural source for such data would be query logs (Nguyen et al., 2016; Kwiatkowski et al., 2019) or chat histories (Zhao et al., 2023). However, these data sources do not specifically reflect domain-specific use cases and do not allow us to study specific use cases in a controlled manner.

In this work, we bridge this gap by eliciting 519 methodical task descriptions (see a few examples in Figure 1) from 266 experts across 25 different fields (Section 3.1). These writing tasks

* Work done at Google DeepMind.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1–29, 2025. https://doi.org/10.1162/tacl_a_00727
Action Editor: Minlie Huang. Submission batch: 5/2024; Revision batch: 8/2024; Published 1/2025.

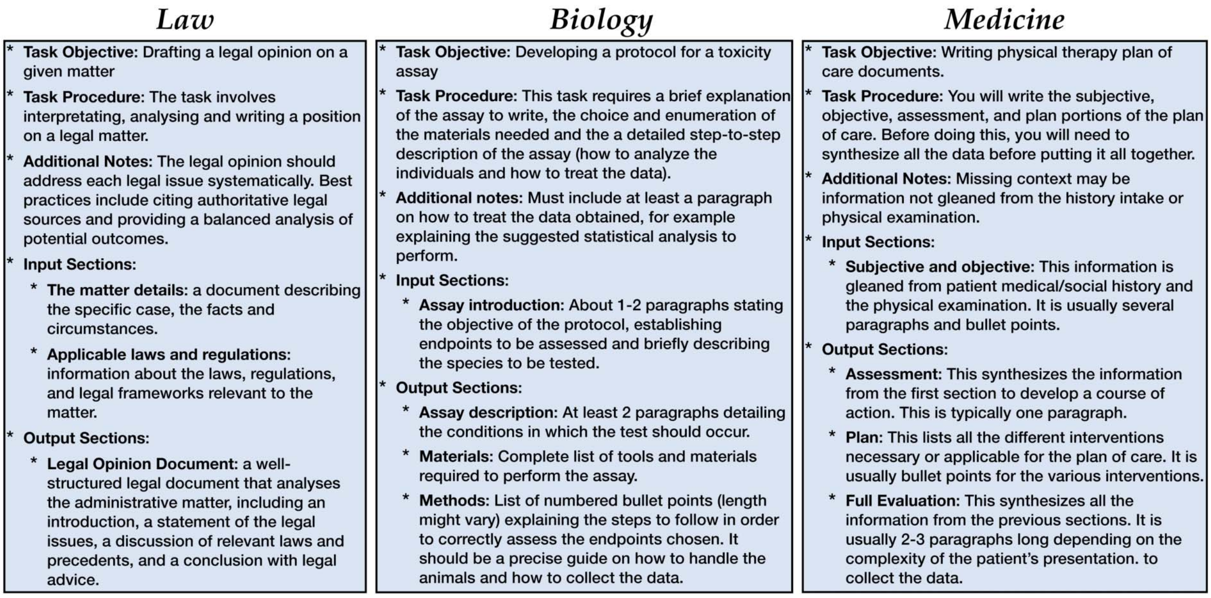

Figure 1: A sample of methodical tasks from law, biology and medicine in DOLOMITES. Each task in DOLOMITES follows a standard template, containing a task objective, task procedure, additional notes about the task, and finally, input sections that are usually expected for the task, and output sections that need to be produced as part of the task. These tasks are instantiated with *examples* that represent plausible inputs and outputs for the task (Section 3.4).

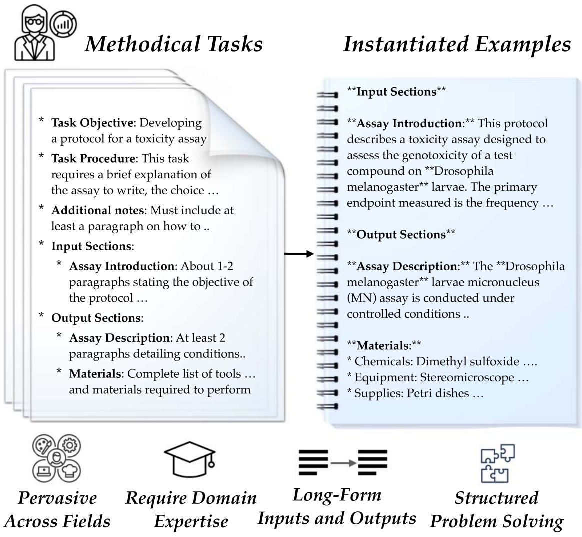

Figure 2: DOLOMITES contains descriptions of 519 methodical tasks elicited from domain experts across various fields. We instantiate these tasks with examples that contain plausible inputs and outputs, formulating a challenging long-form generation problem that requires domain expertise and structured problem-solving.

are formatted in a standard way, with a task procedure, input, and output. Further analysis with an independent group of experts reveals that they are indeed plausible (∼76% of them are likely to be conducted by an expert on a regular basis) and most experts (∼63%) would find it useful if they could use a capable AI model as a writing assistant (Section 3.3). Our tasks serve as the first collection of realistic use cases of experts spanning multiple domains.

To evaluate the ability of existing models to assist experts with these tasks, we collect *examples* (see Figure 2), where we instantiate each task with plausible inputs and outputs (Section 3.4). Examples are created semi-automatically: We first retrieve Web documents that could potentially serve as samples of the task, and then generate an example using a language model based on the retrieved web documents. These examples are then significantly post-edited by the same experts (who contributed the task) for improving adherence to the task description, factual correctness, and level of detail.

We use our benchmark, called DOLOMITES (short for Domain-Specific Long-Form Methodical Tasks), to evaluate current models in their ability to generate accurate and detailed outputs (Section 4). We formulate the modeling problem as long-form generation, where models are provided the task description and the example input and asked to generate the example output. Our experiments reveal that there is significant headroom in improving performance on methodical tasks (Section 5) which are inherently difficult (requiring reasoning skills and domain-knowledge), and in terms of improving automatic evaluation of long-form text. In addition to well-known shortcomings (Schluter, 2017; Krishna et al., 2021), conventional metrics are not designed to capture expert knowledge.

We hope that DOLOMITES can serve as a reference for domain-specific use cases of language models and provide a means for evaluating future models. We release our dataset and code at `https://dolomites-benchmark.github.io`.

2 Problem Formulation

In this section, we first describe the types of writing tasks considered in this work. We refer to these tasks as methodical writing tasks due to two properties that are common to their execution. Firstly, each task requires **structured problem-solving**, where the task follows a specific order where each step logically flows from the previous ones. For instance, in the *Medicine* task in Figure 1, the task requires producing an assessment of the patient, then a plan of care and finally a full evaluation of the patient. Secondly, every task usually follows a **consistent execution** across inputs, where there is a standard specification of the input, the output and the procedure for the task. In the same *Medicine* task, given a patient's subjective and objective data, the task structure and procedure would mostly stay consistent across patients.

To elicit descriptions of tasks from experts, we operationalize our definition of a methodical task into a standard template (see Figure 1 for examples). We require that every task contains a brief *task objective*, a *task procedure* walking a beginner through how this task is conducted, *input and output sections*, which include information that is typically given, and information that needs to be generated. Both input and output sections are formatted in the form of section titles and section descriptions. Finally, we collect *additional notes* about the task, which can include best practices or common mistakes, and missing context that is important when conducting the task.

We further expect our tasks to meet the following criteria: (1) they are **purely textual** and do not involve other modalities in the input or output; (2) they **require domain expertise** and can only be completed by an expert; (3) they **do not require use of specific equipment or software**, with the exception of searching the Web; (4) they are **frequent**, routinely performed by an expert at least once every few months; and (5) **time-consuming**, taking a significant but not indefinite amount of time to complete (e.g., from a half hour to a few days, but not several months).

Aside from task descriptions, our dataset contains specific **instantiations** of methodical tasks (see Figure 2). We create examples by populating descriptions like those shown in Figure 1 with plausible input and output sections (see Section 3.4).

3 DOLOMITES: Data Curation

3.1 Task Collection

In our data curation process, we first collect a typology of realistic tasks that span multiple fields. These tasks are not meant to be exhaustive, but instead represent realistic use cases across fields.

Participants. We recruit 266 participants from the Prolific crowdsourcing platform. We recruit experts from 25 different fields, shown in Table 1, aiming for a broad coverage across disciplines. Participants qualify as experts if they have formal education in the field, and at least 3 years of work experience. Additional details about the participants' backgrounds are provided in Appendix A.

Annotation Task. We ask each annotator to provide descriptions of two writing tasks they routinely perform in their profession subject to the criteria listed in Section 2. For each task, annotators are asked to fill in predefined fields (task objective, procedure, input and output sections, additional notes), the same way as shown in Figure 1. We ask annotators to give thorough descriptions as if they are teaching a novice how to perform each task. Instructions and interface screenshots are in Appendix A.

3.2 Task Analysis

After collecting the initial set of tasks, we filter them manually to ensure they meet the criteria outlined in Section 2, and obtain a total of 519 tasks. We find that there are very few tasks from a field that are highly similar.

Table 1 provides the number of tasks across fields and the task objective of a sample task from each field. Most fields have at least 20 tasks, with some exceptions, where we were not able to

Field	Sample Task Objective
Anthropology (8)	*A survey to examine specific cultural practices, rituals, and societal norms within a cultural group or community*
Architecture (20)	*Developing a construction phasing plan for a building project*
Biology (21)	*Developing a protocol for a toxicity assay*
Business (26)	*Write a section of a non-financial report for a client, focusing on a company's environmental and social activities*
Chemistry (21)	*To write a retrosynthesis scheme/plan for a specific target molecule*
Economics (17)	*Reviewing investment options for advising companies*
Education (23)	*To create a lesson plan for a school class*
Engineering (22)	*To write the instructions for conducting a radioactive experiment.*
Environmental Sci (23)	*Writing the life cycle assessment of a system, product or process*
Geography (20)	*Analyzing the environmental and social impacts of illegal mining activities in a specific region*
History (22)	*Summarize and analyze a specific medieval legal code*
Hospitality (21)	*Adapt existing recipes to cater to various dietary preferences*
Journalism (20)	*Write a news story based on an interview*
Law (38)	*Drafting a petition to challenge a decision*
Linguistics (20)	*Carry out a short literature review of a given problem in linguistics*
Literature (20)	*To write a research proposal for a presentation at a literary research conference*
Mathematics (15)	*Writing an experimental setup suitable for testing a research hypothesis in applied mathematics*
Medicine (24)	*Writing a list of potential radiotherapy regimens for a cancer patient*
Music (23)	*Writing lyrics for a game's soundtrack*
Philosophy (13)	*Provide ethical recommendations for patient/doctor cases*
Physics (21)	*Design specifications for a pump or turbine system*
Political Sci (20)	*Redline a management measure / legislative policy*
Psychology (21)	*Writing a study protocol of a neuroimaging research project*
Sociology (20)	*Analyzing responses from sociological interviews to identify themes relevant to the research question*
Visual Arts (20)	*The objective of this task is to write a catalog entry for an art exhibition*

Table 1: Fields represented in DOLOMITES, with number of tasks in parentheses and a sample task from each field.

recruit as many experts. Across tasks, there are an average of ∼2.78 sections in the input and ∼2.82 sections in the output. Collectively, tasks in DOLOMITES are cognitively demanding and versatile in the types of reasoning they require. For instance, a diagnostic task in medicine requires *inductive reasoning* to go from particular symptoms to a general diagnosis. Whereas, in legal analysis, *deductive reasoning* is required to reason about how laws are interpreted in a specific case and *analogical reasoning* is needed as lawyers compare current cases with precedents. Similarly, in software application design, *abstract* reasoning is important while creativity is necessary for certain tasks in the visual arts. While it is hard to describe the type of reasoning required for all tasks, every methodical task essentially involves **analyzing** the input, **making inferences** based on the input and domain-specific knowledge, and finally, **providing a justification** in writing.

3.3 Task Validation and Societal Implications

We validate our collection of tasks by consulting an *independent* group of experts. Specifically, we collect Likert ratings for each task from three experts on the following axes (the precise description for each item on the scale is provided in Appendix A):

- **Representativeness**: How likely is this task to be conducted by an expert in your field?
- **Complexity**: How would you rate the complexity of this task?
- **Time Required**: How much time is typically required to complete this task?
- **Usefulness**: Would you or other experts find it useful if an AI system were to propose initial outputs for this task (which may be lacking), that can be validated and improved by experts?

The questions above are motivated by prior work showing that AI writing assistants could significantly benefit the productivity of experts (Eloundou et al., 2024; Noy and Zhang, 2023; Dell'Acqua et al., 2023). Beyond productivity, we were additionally interested in expert opinions about the *societal implications* of using language models as writing assistants. Hence, for each task, we elicit answers to the following questions which

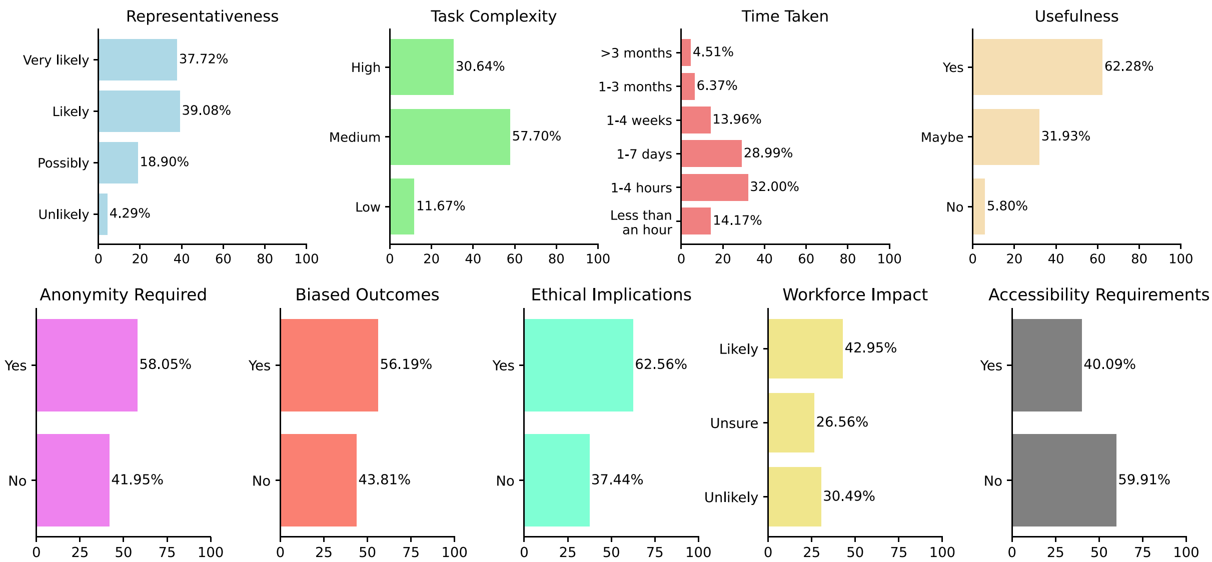

Figure 3: We conducted validation of methodical tasks in the DOLOMITES task collection by consulting an independent group of 3 experts from the field to which the task belongs. Here we show the Likert distributions of their ratings across various axes of importance. The question associated with each axis is listed in Section 3.3.

require a free-text response in addition to a Likert rating.

- **Anonymity Required**: Is it important to ensure anonymity of any individuals or organizations if an AI system is used for conducting this task?
- **Biased Outcomes**: Could relying on automatically generated outputs for this task result in biased or potentially harmful decisions for certain groups of people?
- **Ethical Considerations**: Are there ethical considerations (e.g., privacy, copyright issues) associated with the use of AI systems for this task?
- **Workforce Impact**: Could partial automation of this task have an impact on the workforce in the short term?
- **Accessibility Requirements**: Would the use of AI tools for this task require making exceptional considerations to ensure accessibility?

The main outcomes of our validation study are presented in Figure 3. We find that the tasks collected in DOLOMITES are ecologically valid, i.e., they are likely (∼76%) to be conducted by field experts. Most of them are of medium or high complexity, requiring moderate (a few years of experience) or substantial (several years of experience) expertise. While they are complex tasks, judgments about time taken reveal that most (∼61%) would take an expert from 1 hour to 7 days to complete. This degree of difficulty suggests that it is conceivable for language models to be useful assistants for these tasks. Finally, an overwhelming majority of experts would be interested (∼62%) or open to trying (∼32%) to use a language model that proposes initial outputs for the task.

With regard to the societal implications of language model use, the need for anonymity emerged as a concern for a significant number of tasks (∼58%). In fields like medicine, psychology, and law, experts emphasized the importance of protecting patient/client confidentiality. Similarly, experts felt strongly that proprietary information and trade secrets should be kept private in fields like business. Experts further thought that using language model responses without careful perusal can result in biased outcomes (∼56% of tasks) which could affect marginalized or underrepresented groups. They also raised various ethical concerns relating to copyright issues, privacy issues and stifling of human creativity due to over-reliance on AI.

Many experts (∼43%) recognized that partial automation of writing tasks is likely to impact the workforce in the short term, potentially leading to changes in job roles or skill requirements. At the

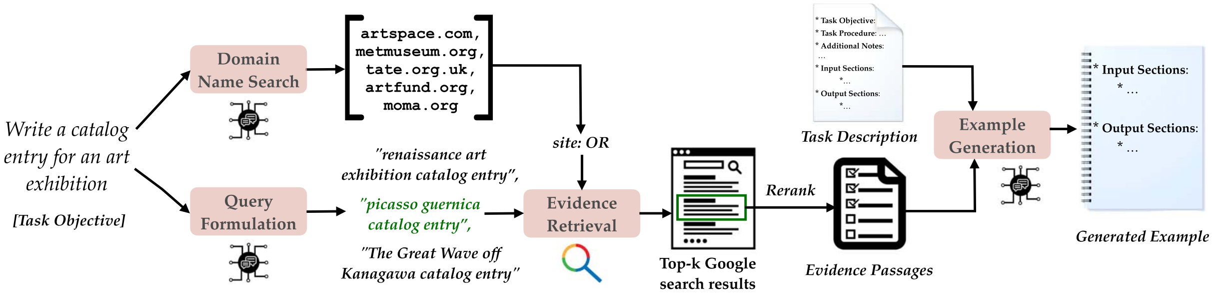

Figure 4: Here, we outline the method for constructing examples of tasks in DOLOMITES. Using the task objective for a task, we first generate more specific queries to search for relevant web documents, where we constrain our search to authoritative domain names for the task. Using a set of retrieved evidence passages and the complete task description, we then generate an example of the task that fits the task structure using a language model. This example is then post-edited by the same expert who provided the task (further described in Section 3.4.2).

same time, they were optimistic that this would improve productivity and bring positive changes to the nature of the work. It is important to ensure that users of all backgrounds and capabilities have equal access to language models. A significant number of tasks (∼40%) were rated as requiring exceptional considerations to be made for ensuring accessibility to all users. Across fields, experts highlighted that while language models as writing assistants can improve productivity, human oversight is important for responsible use of these technologies.

3.4 Example Collection

To evaluate language model capabilities in assisting experts with their tasks, we create examples of input and output sections with concrete details. We adopt a human-in-the-loop methodology, where initial examples are generated by a model, which are then post-edited by the same expert who provided the task. We describe this process below.

3.4.1 Retrieval-Augmented Generation

We believe that samples of methodical tasks are partially available in documents on the Web. Hence, we retrieve relevant documents for each task and generate examples by prompting models with passages from these documents as context. This process is depicted in Figure 4 and explained below.

Query Formulation. Given a task description, we first need to find more specific queries that could potentially result in relevant Web documents. For example, for writing a catalog entry for an art exhibition, search queries like ‘‘*renaissance art exhibition catalog entry*’’ or ‘‘*picasso guernica catalog entry*’’ are likely to result in documents that contain examples of the task. To generate search queries, we prompt Bard (Manyika and Hsiao, 2023) (with 1 exemplar) with the task objective and instruct it to generate more specific queries that can help find Web documents which contain demonstrations of the task (Table 11 in Appendix B). We generate 10 search queries for each task and restrict search to reliable and authoritative sources. These are collected by prompting Bard (with 1 exemplar) with the task objective to generate URLs to domain names which will be useful to find real examples for the given task (Table 12 in Appendix B).

Evidence Collection. Using each search query, we gather the top-10 documents from Google search restricted to relevant domains with the `site:` operator in the query. Documents are then split into passages of 4,000 characters with a 100 character sliding window.

Conditional Generation. Having gathered evidence which may contain task demonstrations, we generate examples by prompting models with this evidence. We explore a *multi-document* setting, where passages are sampled from multiple documents and a *single document setting*, where passages are sampled from a single document as we found that the appropriate choice depends on the task.[1] Passages are reranked using an in-house reranker and the top-5 passages are provided as context, along with the task description, to a large language model (Gemini-Ultra (Team, 2023) and Bard in our case), which is asked to generate an

[1]For instance, a task that requires drafting a legal opinion might benefit from multiple relevant documents whereas a single document might be sufficient for writing the catalog entry for an artwork.

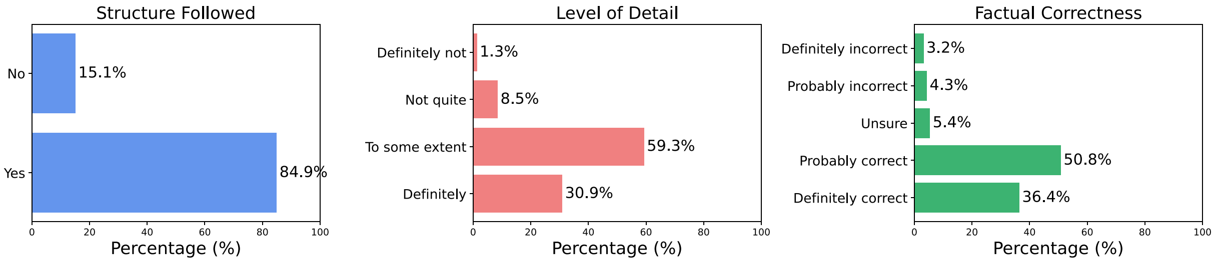

Figure 5: Expert judgments of original examples along three dimensions: **task structure followed** (whether the example includes all the input and output sections from the task description), **level of detail** (whether the example shows a detailed and concrete sample of the task), and **factual correctness** (whether the example is factually correct).

example of the task (Table 13 in Appendix B). Not all information mentioned in the task description is required to be present in the context and the model is allowed to infill content to construct an example. We generate up to 10 examples for each task.

3.4.2 Expert Post-Editing

Even though they are based on retrieved Web documents, model-generated examples could have many issues. They may not adhere to the structure specified in the task description, they may contain factual inconsistencies or inaccuracies, lack in depth, or be vague. To remedy these issues, we present the examples to the same experts who wrote the tasks for post-editing. They are asked to choose the most plausible example out of four variants, and use single or multi-document evidence.

Prior to post-editing, experts are asked to label examples according to three criteria on a Likert scale: adherence to task structure, factual correctness, and level of detail. They are also shown 1) the evidence passages for the example and 2) a critique (generated using Gemini-Ultra with the prompt in Appendix B, Table 14) that may not be comprehensive, to aid them in identifying issues with the example. The critique is provided to make post-editing more efficient. They are required to fix any valid issues they recognize in the example as well as any valid issues identified by the critique.

3.4.3 Example Quality Analysis

Automatic Analysis. Expert judgments of automatically generated examples are shown in Figure 5. We find that the majority of examples already follow the task structure (∼85%) and

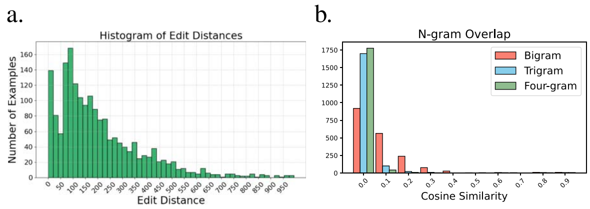

Figure 6: Histogram of (a) word-level edit distance between original and post-edited example and (b) cosine similarities based on n-grams in post-edited examples and evidence used as context.

	Avg Length	**Avg section presence %(↑)**	**Flesch-Kincaid Grade Level (↑)**
`original`	388.98	92.81	11.69
`post-edited`	590.24	97.67	13.46

Table 2: Statistics of the original and post-edited examples in DOLOMITES.

most of them are *probably* or *definitely* correct. However, a large number of examples are lacking in depth and detail. We show a histogram of the word-level edit distance between all tokens in the original and edited example in Figure 6a and relevant statistics of the original and post-edited examples in Table 2. The histogram suggests that on average, there are significant changes made to the original examples during post-editing. Since most experts judge that examples are lacking in depth, the edited examples are expectedly much longer on average. The edited examples also adhere better to the task description (on average, 97.67% sections in the task description are found in the edited examples compared to 92.81% in the original examples). We analyzed a random set of 100 examples and labeled each example with the type(s) of edits it contained. We found broadly the following types of edits: fact addition (88%), fact deletion (20%), fact update (65%), stylistic rewrites (76%), and reorganization

(23%). These edit types are described further in Appendix A. Finally, we compute readability scores of the examples as a noisy approximation of the complexity and level of detail in the text, using the Flesch-Kincaid Grade Level test (Kincaid et al., 1975). Higher readability values indicate that a piece of text requires more formal education and expertise to understand. We find that post-edited examples have higher scores, possibly indicating higher level of technical depth.

Data Contamination Analysis. Examples in DOLOMITES are created by conditioning on passages from Web documents. We do not require these documents to contain complete examples of the task, and models are allowed to infill information to create an example. However, if these documents are seen by models during large-scale pretraining, it might be difficult to conduct a clean evaluation. We examine whether this is the case by computing the similarity of the post-edited examples to the evidence passages provided during generation. Figure 6b plots the cosine similarity between the n-grams found in the post-edited examples and evidence used as context. As can be seen, similarity to retrieved passages is fairly low in most examples, which suggests a low risk of memorization due to pretraining.

In addition, we check directly if text in our examples can be found in large pretraining corpora. We use an open-source toolkit called WIMBD (Elazar et al., 2024) to measure the presence of text in generated examples in two large pretraining corpora: C4 (Raffel et al., 2020) (364 million documents) and Dolma (Soldaini et al., 2024) (5,245 million documents). These corpora are widely used to train large language models (Raffel et al., 2020; Dodge et al., 2021; Chowdhery et al., 2023; Groeneveld et al., 2024). We follow prior work from Chowdhery et al. (2023) and determine an example as contaminated if 70% of all possible 8-grams in the example were seen at least once in the pretraining corpus. Conducting this process for a random sample of 100 examples in DOLOMITES, we find that none of the examples are contaminated.

4 Experiments

4.1 Setup and Models

We create a development-test split for DOLOMITES, with 820 examples in the development (dev) set and 1,037 examples in the test set. There are 172 *seen* tasks with examples in both dev and test and 99 *unseen* tasks with examples only in the test set.

For evaluation, we considered multiple performant models from various companies as well as open-source models. In all cases, we favored instruction-tuned variants because of their better performance on other benchmarks. Specifically, we report experiments with Claude-3 Opus (Anthropic), Command-R-Plus (Cohere, 2024), Gemini-1.5-Pro and Gemini-1.5-0409 (Team, 2024) and Gemini-Pro (Team, 2023), GPT-3.5-Turbo and GPT-4 (OpenAI, 2023), Mixtral-8×7B and Mixtral-8×22B (Jiang et al., 2024) and Mistral-Large (Mistral, 2024), and OLMo-7B-Instruct (Groeneveld et al., 2024). In all cases, we prompt models with the task description, the input sections corresponding to an example, and instruct them to generate the output sections for the example, in a zero-shot manner. Hyperparameters, prompts, and model identifiers are in Appendix B.

4.2 Automatic Evaluation

4.2.1 LM-based Evaluation

LM-based Pairwise Evaluation. We consider two modes of LM-based evaluation: pairwise evaluation and fine-grained absolute evaluation (illustrated in Figure 7). Language models are being increasingly used as evaluators ((Chiang and Lee, 2023), and our primary evaluation also involves using LMs as evaluators. However, recent work points out that LM judgments can be misleading and biased (Shen et al., 2023; Wang et al., 2024; Zheng et al., 2023; Panickssery et al., 2024). We use *multiple* language models as judges to give preferences for a pair of model outputs. While this does not alleviate the problem of biased LM judgments, we believe it is slightly more reliable since we are not biased by a single model's judgments. In all comparisons, we use one of the strongest models, GPT-4, as the base comparison model. We sample outputs on the test set from a candidate model and GPT-4 (randomizing their order in the prompt) and ask the evaluator model to judge which output is better and provide a justification. We consider three models as evaluators: GPT-4, Claude-3 Opus, and Gemini-1.5-0409. The win rate is computed by summing up the number

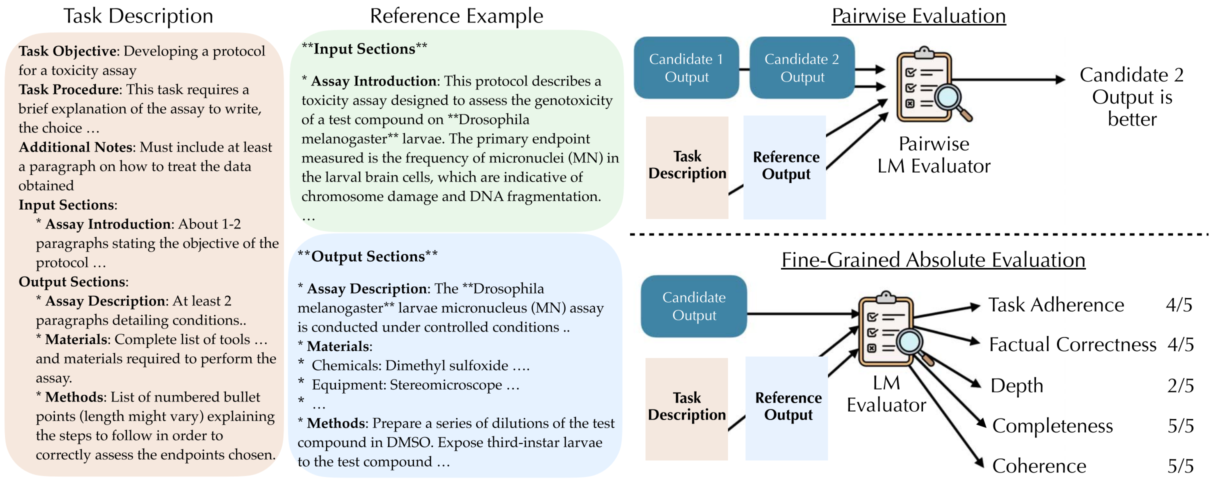

Figure 7: Automatic evaluation protocol for DOLOMITES, involving two modes of evaluation: **pairwise evaluation** of two candidate outputs and **fine-grained absolute evaluation** of a candidate output on our proposed axes.

of wins for a candidate model plus half the number of ties.

LM-based Fine-Grained Evaluation. In addition to preference judgments, we evaluate models on five finer grained aspects of response quality: task adherence, factual correctness, depth, completeness, and coherence.[2] We use GPT-4 to collect absolute ratings (on a scale of 1–5) of individual model responses on each of these aspects.

4.2.2 Other Evaluation Metrics

Round-Trip Factual Consistency. We also measure the extent to which statements in the model output are consistent with statements in the reference output. We compute 1) *forward* entailment considering a reference section as the premise and the corresponding model section as the hypothesis and 2) *reverse* entailment considering a model output section as the premise and the corresponding reference section as the hypothesis. Scores are aggregated over all sections and examples. These metrics loosely capture the notions of precision and recall, we also report the harmonic mean of the two. We use the TRUE model (Honovich et al., 2022) to predict entailment scores (ranging from 0 to 1) and report 95% confidence intervals. Note that this metric has weaknesses, as it assumes that there is a single valid reference for each example, which may not be true for many examples in DOLOMITES.

[2]These aspects of response quality are defined in Table 17.

Conventional Metrics. Prior work has recognized that conventional metrics for text generation are lacking in various ways (Liu et al., 2016; Novikova et al., 2017; Krishna et al., 2021). Nevertheless, for completeness, we report results with ROUGE-L (Lin, 2004) and BLEURT (Sellam et al., 2020). In addition, we report the average output length and average section presence (i.e., the percentage of output sections specified in the task that are present in the generated output, averaged across all examples) as a measure of instruction-following capabilities. Note that the average length of reference outputs is 341.42 tokens.

5 Results and Discussion

Human Evaluation. To evaluate the reliability of automatic metrics, we measure how well the above automatic evaluation measures correlate with human judgments. Specifically, we sample 200 pairs of model outputs, where each pair comes from two randomly chosen models. We (the authors) then label which model output is better (or if they are tied) according to their task adherence, factual correctness, and depth.[3] On these 200 pairs, we also get automatic preference judgments from all the evaluation measures discussed in Section 4.2 (we convert float scores for two outputs into binary judgments). The percentage agreements between human labels (with and without pairs with ties) and all evaluation measures

[3]Human agreement between two annotators was found to be 75% on 100 examples when including ties.

Model	Avg Length	Avg Section Presence %	BLEURT	ROUGE-L	nli (forward)	nli (reverse)	nli (h-mean)
Claude-3 Opus	417.41	91.53	0.4156	0.2395	$0.3584_{[0.34,0.38]}$	$0.3769_{[0.36,0.40]}$	0.3674
Command-R-Plus	440.44	92.92	0.4068	0.2134	$0.3926_{[0.37,0.41]}$	$0.3623_{[0.34,0.38]}$	0.3768
Gemini-1.5-Pro	349.27	95.25	0.4136	0.2371	$\mathbf{0.4065}_{[0.39,0.42]}$	$0.3846_{[0.37,0.40]}$	0.3952
Gemini-1.5-0409	361.87	**95.74**	0.4068	0.2361	$0.3994_{[0.38,0.42]}$	$0.3984_{[0.38,0.42]}$	**0.3989**
Gemini-Pro	269.68	93.61	0.4124	0.2280	$0.3415_{[0.32,0.36]}$	$0.3090_{[0.29,0.33]}$	0.3244
GPT-3.5-Turbo	240.89	88.98	**0.4276**	0.2309	$0.3854_{[0.37,0.40]}$	$0.2949_{[0.28,0.31]}$	0.3341
GPT-4	407.35	95.46	0.4155	0.2271	$0.3934_{[0.38,0.41]}$	$\mathbf{0.3993}_{[0.38,0.42]}$	0.3963
Mistral-Large	327.12	92.61	0.4158	0.2390	$0.3524_{[0.33,0.37]}$	$0.3523_{[0.33,0.37]}$	0.3523
Mixtral-8×22B	339.16	95.16	0.4212	**0.2450**	$0.3951_{[0.38,0.41]}$	$0.3583_{[0.34,0.37]}$	0.3758
Mixtral-8×7B	386.61	88.39	0.4098	0.2266	$0.3290_{[0.31,0.35]}$	$0.3097_{[0.29,0.33]}$	0.3191
OLMo-7B-Instruct	784.22	74.02	0.3905	0.1752	$0.1929_{[0.18,0.21]}$	$0.1721_{[0.16,0.19]}$	0.1819

Table 3: Results on the DOLOMITES test set with standard metrics and factual consistency using NLI models. We report 95% confidence intervals along with the average NLI scores.

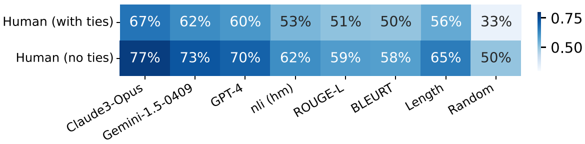

Figure 8: Percentage agreement of automatic evaluation measures with human labels. Pairwise judgments from Claude-3 Opus have the highest correlation with human labels.

Model	GPT-4	Claude-3 Opus	Gemini-1.5-0409
Claude-3 Opus	48.1	52.7	49.6
Command-R-Plus	34.4	45.8	38.7
Gemini-1.5-Pro	41.1	46.1	51.0
Gemini-1.5-0409	42.9	**55.4**	**60.9**
Gemini-Pro	17.6	21.0	22.1
GPT-3.5-Turbo	12.2	11.5	12.4
GPT-4	**50.0**	50.0	50.0
Mistral-Large	27.2	28.8	26.7
Mixtral-8×22B	21.6	25.1	17.6
Mixtral-8×7B	17.8	23.5	15.6
OLMo-7B-Instruct	4.2	5.5	3.3

Table 4: Model win rates (± 3) against GPT-4 on the DOLOMITES test set using three LM-based autoraters (GPT-4, Claude-3 Opus, and Gemini-1.5-PP). GPT-4's win rate is 50% since it is the base comparison model. Note that pairwise judgments from Claude-3 Opus have the highest correlation with human judgments.

are in Figure 8. In summary, we find that **LM-based evaluation measures have the highest correlations with human judgments**, followed by the NLI measures and then ROUGE-L and BLEURT. We note that an evaluator that always picks the longer output also has reasonable agreement rates.

We first report overall statistics of model responses and scores using conventional metrics in Table 3. Based on the average section presence, we note that models are largely effective at generating almost all relevant output sections from the task description. A few models (GPT-4 and Gemini-1.5-0409) excel more at following this instruction. Based on the NLI scores, we note that the nli (reverse) scores are on average lower than nli (forward), which suggests that generated outputs contain statements not entailed by the reference, e.g., because they are inaccurate or irrelevant. We observe that Gemini-1.5-0409 and GPT-4 produce more information that is factually consistent with the reference, while Claude-3 and Command-R-Plus are also performant. Once again, we note that reference-based metrics generally penalize outputs which are valid but different from the reference, and there might not just be a single reference for some of these examples, especially when the task is more subjective.

Pairwise LM Evaluation Results. We show the win rates according to different LM evaluators in Table 4. Based on these win rates, we note that a few models such as Claude-3 Opus, Gemini-1.5-Pro, and Gemini-1.5-0409 prove to be comparable to GPT-4. We also report win rates with a length penalty for longer outputs in Table 7 and the overall rankings do not change.

Fine-Grained LM Evaluation Results. Finally, we show ratings of models according to finer-grained aspects of response quality in Table 5. The overall conclusions are roughly similar, i.e., Gemini-1.5-0409, Claude-3 Opus, and GPT-4 have the highest ratings across axes. Across the axes considered in our rubric, models struggle most with the level of technical depth of the generated text.

Model	Task Adherence	Factual Correctness	Depth	Completeness	Coherence	Average Rating
Claude-3 Opus	4.57	4.73	4.36	4.54	4.84	**4.61**
Command-R-Plus	4.36	4.57	4.17	4.35	4.73	4.44
Gemini-1.5-Pro	4.41	4.70	4.21	4.38	4.82	4.50
Gemini-1.5-0409	4.49	4.71	4.30	4.46	4.83	4.56
Gemini-Pro	3.95	4.24	3.62	3.87	4.43	4.02
GPT-3.5-Turbo	3.90	4.34	3.37	3.73	4.36	3.94
Mistral-Large	4.38	4.59	3.98	4.31	4.72	4.40
Mixtral-8x22B	4.23	4.47	3.83	4.18	4.60	4.26
Mixtral-8x7B	3.99	4.23	3.64	3.94	4.36	4.03
OLMo-7B-Instruct	2.59	3.04	2.62	2.61	2.95	2.76

Table 5: Results on the DOLOMITES test set along fine-grained aspects of response quality. All ratings are performed on a scale of 1-5 by GPT-4-Turbo-Preview and we report average ratings across all examples (all average ratings were found to be statistically different from the best model's average ratings, which is in this case, Claude-3 Opus).

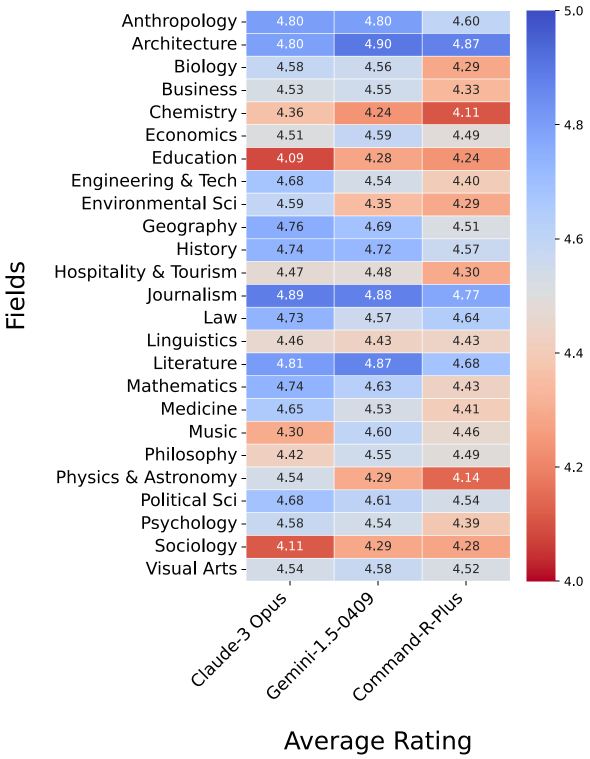

Figure 9: Heatmap of average ratings aggregated by field for Claude-3 Opus, Gemini-1.5-0409, and Command-R-Plus.

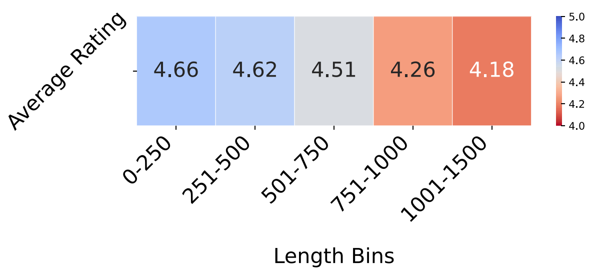

Figure 10: Heatmap of average ratings of Claude-3 Opus aggregated by length (in tokens) of reference output.

Results by Field. Figure 9 illustrates how model performance fluctuates across fields; we show average ratings based on the fine-grained LM evaluation for Claude-3 Opus, Gemini-1.5-0409, and Command-R-Plus. Across models, we find that a few fields have significantly lower average ratings: Education, Sociology, and Chemistry. Tasks from a subset of these fields are sometimes subjective (e.g., *Create a lesson plan to teach STEM educators*, *Write a summary of findings about the conclusions of a sociology book*), which can make them hard to evaluate. In other cases, outputs can be lacking technical depth for tasks which require more domain expertise (e.g., *Drafting a experimental protocol for a chemical synthesis procedure.*). On the other hand, tasks in fields such as Literature, History, and Journalism have higher average ratings, some of which focus on factual reporting and narratives that may be easier to reason about.

Results by Length. Next, we evaluate whether output length is correlated with performance. Specifically, we show average ratings based on the fine-grained LM evaluation for Claude-3 Opus split into bins by length of the reference outputs. Scores stratified by length bins are shown in Figure 10. We find that examples which require longer outputs are significantly harder for models, supported by the fact that examples with longer outputs have lower average scores.

Error Analysis. We analyze generated outputs from 3 high-performing models (Gemini-1.5-0409, GPT-4, and Claude-3 Opus) for examples where average ratings of responses are lowest. Broadly, we observe the following patterns:

- **Lacking depth** (Table 8): Writing technical documents requires depth and focus, which was sometimes found to be lacking. For example, concrete statistical results and method details were absent from a task on *writing up a report on an experimental study in clinical psychology*.
- **Verbosity** (Table 9): A common characteristic of some model outputs was their verbosity. Common patterns included defining jargon when not necessary, and generating many filler statements that do not introduce new information.
- **Missing information** (Table 10): There were a few cases where a single output section required multiple pieces of information, but the model entirely missed producing a subset of them.

6 Related Work

Domain-Specific NLP Benchmarks. The use of language technologies in domain-specific scenarios has the potential to help experts. Prior work has evaluated models through domain-specific benchmarks for standard tasks like QA (Hendrycks et al., 2021; Malaviya et al., 2024) and summarization (Hayashi et al., 2021). Many benchmarks have been proposed for specific fields (Rein et al., 2024; Xia et al., 2024), including law (Shen et al., 2022; Niklaus et al., 2023; Guha et al., 2024) and medicine (Tsatsaronis et al., 2015; Pampari et al., 2018; Jin et al., 2019, 2021; Fleming et al., 2024).

A notable difference between this line of work and DOLOMITES is the task formulation (i.e., QA vs methodical tasks). QA involves addressing a specific information need in response to a query while conducting methodical tasks requires following a structured and consistent procedure, involving multiple steps, to complete a goal-oriented task.

Naturalistic Evaluation. Evaluation that is grounded in realistic use cases, is important for reliable benchmarking (Rolnick et al., 2024). Prior efforts on creating NLP benchmarks which are representative of real user needs use query logs (Nguyen et al., 2016; Kwiatkowski et al., 2019) and chat histories (Lin et al., 2024). While these benchmarks are useful for evaluating responses to generic user queries, they do not allow us to study their abilities in assisting with domain-specific tasks in an isolated manner. For instance, Ouyang et al. (2023) find that user requests in chat histories often involve ‘‘planning’’ and ‘‘design’’, but these are largely ignored in benchmarks. Our work fills this gap by presenting a typology of domain-specific tasks grounded in realistic scenarios.

Language Models as Writing Assistants. Recent work has investigated the potential of language models to act as writing assistants for domain experts (Calderwood et al., 2020; Lee et al., 2022; Gero et al., 2022; Li et al., 2024). While there is favorable evidence that they can improve productivity of experts (Noy and Zhang, 2023), their usage has broader societal consequences, including potential impact on the workforce (Eloundou et al., 2024). We analyze a subset of these societal implications relevant for our methodical tasks in Section 3.

7 Discussion

As part of DOLOMITES, we present various data artifacts that can be used to study the abilities of language models in assisting domain experts with writing tasks. We outline these artifacts and their intended use below.

Task Collection. We present a collection of 519 writing tasks spanning 25 fields that is representative of work undertaken by domain experts on a regular basis. We believe this is the first collection of tasks built with input from domain experts about scenarios in which language models can be useful to them. These tasks can be used to identify applications of LMs in various fields and to cater model development to ecologically valid tasks.

Task Validation Labels. We conduct an independent assessment of the validity of the tasks and the societal implications of using LMs as writing assistants for these tasks. We believe that these validation labels can be useful to study the practical benefits of using LMs as writing assistants for these tasks and to select tasks that are representative of real-world use. Labels concerning societal implications can enable researchers to take into account important considerations such as anonymity of user data, bias in decision making, accessibility requirements, etc, when considering deployment of language models for assisting experts.

Task Examples. Finally, we present examples of tasks that are semi-automatically generated by seeking input from the same experts who provided the tasks. The examples of tasks are meant to represent concrete and plausible instances of the task, so that models can be evaluated on the tasks. To conduct such an evaluation, models are required to generate the output of an example given the task description and the example input. In Section 4.2.1, we present results on DOLOMITES using LM-based evaluators as well as other metrics. We find that LM-based evaluators correlate best with human judgments and propose future work to consider the following modes of evaluation on DOLOMITES:

1. **Pairwise preference judgments for overall response quality**: Pairwise evaluations from models (especially from Claude-3 Opus) have moderately high agreement with human labels (67% with ties, 77% without ties) that are better than other evaluation metrics. These can be used to get an estimate of overall model abilities.
2. **Finer-grained LM evaluation for absolute judgments on given axes**: In Section 5, we present fine-grained evaluation on axes that are important for our tasks. We propose future work to conduct finer-grained evaluations on the same axes to gather better insights about the strengths and weaknesses of models.

We believe that as LM-based evaluators improve, we will be able to more accurately evaluate outputs for examples in DOLOMITES.

8 Conclusion

We introduce DOLOMITES, a benchmark that is closely tied to realistic use cases of domain experts. The generalization of these use cases as methodical tasks provides a way to study capabilities of language models across tasks and domains. We consider a scenario where AI systems can act as tools for experts to amplify their problem-solving capabilities (Engelbart, 2023) and perform their tasks more efficiently. We verify that our tasks are representative across fields and that human oversight is necessary if language models propose initial outputs for these tasks. Evaluation of a broad range of contemporary language models suggests that there is a large room for models to improve on generating outputs for our tasks.

Future directions are many and varied. The tasks in DOLOMITES constitute a mere sample from 25 fields in English language. We hope to further expand the set of tasks to cover a wider range of scenarios and languages. We could also consider tasks that involve modalities other than text in input or output, and multi-turn settings, where models continually improve their outputs through feedback and revision. On the modeling front, we will consider sophisticated generation techniques such as the one proposed by Narayan et al. (2023), that first generate a plan of the output and then fill in different sections, potentially with attributions to sources (Fierro et al., 2024). Our experimental results revealed that automatic evaluation of generated text is particularly challenging. Our data contains a single reference output for an example input and does not model diverse perspectives of experts and the innate subjectivity of tasks (Ganguli et al., 2023). While conventional metrics do not account for this subjectivity, it is unclear if LM-based evaluators innately capture this subjectivity. More research is needed to ensure language model responses are given credit for alternative, but valid responses.

Acknowledgments

We are grateful to the anonymous TACL reviewers and the action editor Minlie Huang for their helpful feedback. We thank the 266 annotators who participated in our data collection studies and the following people for their help with facilitating data collection: Shereen Ashraf, Michelle Chen Huebscher, and Michael Sharman. We also thank the following people for helpful discussions and feedback: Sumit Asthana, Tuhin Chakrabarty, Michael Collins, Sunipa Dev, Jacob Eisenstein, Sebastian Gehrmann, Kalpesh

Krishna, Tom Kwiatkowski, Matthew Lamm, Kenton Lee, Joshua Maynez, Slav Petrov, Hannah Rashkin, David Reitter, Marco Tulio Ribeiro, Elizabeth Sieber, and Kristina Toutanova.

References

Anthropic. The claude 3 model family: Opus, sonnet, haiku. [Accessed on March 4, 2024].

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280. `https://doi.org/10.1038/s42254-023-00581-4`

Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B. Chilton. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+ user2agent@ IUI*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.acl-long.870`

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Cohere. 2024. Introducing Command R+: A scalable LLM built for business. [Accessed on April 16, 2024].

Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013). `https://doi.org/10.2139/ssrn.4573321`

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701. `https://doi.org/10.1038/s44159-023-00241-5`

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.emnlp-main.98`

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. Gpts are gpts: Labor market impact potential of llms. *Science*, 384(6702):1306–1308. `https://doi.org/10.1126/science.adj0998`, PubMed: 38900883

Douglas C. Engelbart. 2023, Augmenting human intellect: A conceptual framework. In *Augmented Education in the Global Age*, pages 13–29. Routledge.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.615

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. 2024. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22021–22030. https://doi.org/10.1609/aaai.v38i20.30205

Jens Frankenreiter and Julian Nyarko. 2022. Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom ed.) Forthcoming*. https://doi.org/10.2139/ssrn.4027030

Deep Ganguli, Nicholas Schiefer, Marina Favaro, and Jack Clark. 2023. Challenges in evaluating AI systems. [Accessed on October 4, 2023].

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019. https://doi.org/10.1145/3532106.3533533

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.841

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36. https://doi.org/10.2139/ssrn.4583531

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225. https://doi.org/10.1162/tacl_a_00362

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.287

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088v1*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14). https://doi.org/10.3390/app11146421

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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 2567–2577. https://doi.org/10.18653/v1/D19-1259

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. https://doi.org/10.21236/ADA006655

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.393

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19. https://doi.org/10.1145/3491102.3502030

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239. https://doi.org/10.1056/NEJMsr2214184, PubMed: 36988602

Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The value, benefits, and concerns of generative ai-powered assistance in writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642625

Bill Yuchen Lin, Khyathi Chandu, Faeze Brahman, Yuntian Deng, Abhilasha Ravichander, Valentina Pyatkin, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016*

Conference on Empirical Methods in Natural Language Processing, pages 2122–2132, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1230

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2024.naacl-long.167

James Manyika and Sissie Hsiao. 2023. An overview of bard: An early experiment with generative ai. *AI. Google Static Documents*, 2.

Mistral. 2024. Au large. Accessed on March 26, 2024.

Ethan R. Mollick and Lilach Mollick. 2023. Using ai to implement effective teaching strategies in classrooms: Five strategies, including prompts. *Including Prompts (March 17, 2023)*. https://doi.org/10.2139/ssrn.4391243

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996. https://doi.org/10.1162/tacl_a_00583

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.200

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1238

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192. https://doi.org/10.1126/science.adh2586, PubMed: 37440646

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774v6.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.146

Brian Owens. 2023. How nature readers are using chatgpt. *Nature*. https://doi.org/10.1038/d41586-023-00500-8

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1258

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076v1*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien

Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L. Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, and Adam White. 2024. Position: Application-driven innovation in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42707–42718. PMLR.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2007

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.278

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.840

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805v4*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530v4*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):1–28. https://doi.org/10.1186/s12859-015-0564-6, PubMed: 25925131

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60. https://doi.org/10.1038/s41586-023-06221-2, PubMed: 37532811

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.acl-long.511`

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. FOFO: A benchmark to evaluate LLMs' format-following capability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 680–699, Bangkok, Thailand. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2024.acl-long.40`

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2023. (inthe) wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Annotation Details

Participants. We recruited 266 participants for our study from Prolific. Participants were required to be fluent in English, and came from 25 different countries, across Africa, Europe, and North and South America. In terms of their background, participants were required to have an undergraduate degree and 3 years of work experience in their respective field. These requirements were first enforced through Prolific's audience filters, followed by a screening where participants were asked to self-report their educational qualifications and work experience. They each provided two tasks, so for each field, we recruited half the number of the participants as the number of tasks reported in Table 1. Lastly, they were required to have at least 50 prior approved submissions and an approval rate of over 99%. Participants were informed that their provided data will be used to evaluate large language models in realistic scenarios. We obtained prior consent from all annotators before recruiting them for all studies.

Setup. Annotators were paid $20 per hour for their work. For task collection, we allocated 40 minutes to write two tasks, and for task validation, we allocated 15 minutes per task. For post-editing examples, we allocated 20 minutes per example.

Edit Types. In a random sample of 100 examples, we found the following types of edits were made to the examples:

- Fact Addition (88%): Addition of new statements to the example.
- Fact Deletion (20%): Removing statements from the example.
- Fact Update (65%): Updating existing statements with further elaboration of details, or adding of new numbers or references.
- Stylistic Rewrites (76%): Simplification, paraphrasing, or improving grammar, spelling or tone of text.
- Reorganization (23%): Restructuring of sentences, paragraphs, or sections in the example, which may be done to fit the task description.

Annotation Interface Screenshots. We show screenshots of the annotation interfaces presented to annotators for task validation and example post-editing in Figures 11 and 12, respectively.

B Experimental Details

Models. The specific identifiers for the models evaluated in this work are given in Table 6. Open-source models were obtained from the HuggingFace model hub, while proprietary models were obtained through the organizations' official APIs.

Generation Configurations. In all generation tasks, we set the temperature for generation to be 0.1. For both example generation and model evaluation, we sampled a maximum of 4,096 tokens (or the maximum sequence length of the model).

Prompts. We provide the prompts used for various components of our work. The prompts used for example creation are given in Tables 11–13. The prompt used to generate the critique shown to annotators is shown in Table 14. The prompt used to generate outputs from candidate models is shown in Table 15. Finally, the prompt used for generating LM-based judgments for evaluation are shown in Tables 16 and 17.

Model Name	Identifier
`Claude-3 Opus`	`claude-3-opus-20240229`
`Command-R-Plus`	`command-r-plus`
`Gemini-1.5-Pro`[4]	`gemini-1.5-pro-latest`
`Gemini-1.5-0409`[5]	`gemini-1.5-pro-preview-0409`
`Gemini-Pro`	`gemini-pro`
`GPT-3.5-Turbo`	`gpt-3.5-turbo`
`GPT-4`	`gpt-4-turbo-preview`
`Mistral-Large`	`mistral-large-latest`
`Mixtral-8×22B`	`Mixtral-8×22B-Instruct-v0.1`
`Mixtral-8×7B`	`Mixtral-8×7B-v0.1`
`OLMo-7B-Instruct`	`OLMo-7B-Instruct`

Table 6: List of models used in our experiments and their identifiers.

Model	GPT-4	Claude-3 Opus	Gemini-1.5-0409
`Claude-3 Opus`	47.6	52.1	49.1
`Command-R-Plus`	33.4	44.4	37.5
`Gemini-1.5-Pro`	43.2	48.4	53.6
`Gemini-1.5-0409`	44.6	**57.6**	**63.4**
`Gemini-Pro`	19.8	23.6	25.0
`GPT-3.5-Turbo`	14.1	13.3	14.4
`GPT-4`	**50.0**	50.0	50.0
`Mistral-Large`	29.4	31.1	28.8
`Mixtral-8×22B`	22.9	26.6	18.7
`Mixtral-8×7B`	18.2	24.0	15.9
`OLMo-7B-Instruct`	2.7	3.5	3.0

Table 7: Model win rate (± 3) against GPT-4 on the DOLOMITES benchmark using three LM-based autoraters (GPT-4, Claude-3 Opus, and Gemini-1.5-PP), with a *length penalty*.

[4]Accessed from `https://aistudio.google.com/app`.

[5]Accessed from `https://console.cloud.google.com/vertex-ai/generative`.

Pattern: Lack of Detail

TASK DESCRIPTION

Task Objective: Designing an observation plan for different celestial bodies and objects using an infrared telescope

Task Procedure: Observing space can be quite a complicated task to achieve, space contains a lot of different celestial bodies and a variety of objects. Some of these objects emits "special" kind of electromagnetic radiation - radiation that we humans cannot see with our eyes. So in this task we're focusing on the invisible glow these objects emit in the infrared part of the spectrum. Our task is to basically decide which celestial bodies (star, planet, galaxies, nebulas and more) we want to study and investigate considering and taking into account their unique infrared features. We also need to plan what kind of telescope is going to be used in order to successfully achieve that mission and literally "see" what we want to see.

Additional Notes: Planning it correctly can save a lot of time and frustrations, taking in account different information like I presented and learning from mistakes can lead to a successful observation.

Input Sections:

* **Scientific/main Goal AND target object**: 1 paragraph, 3-4 sentences. To begin with the planning, we mostly need to understand what is our main objective - so we have to outline the scientific goals we aim to gather from the observation such as understanding atmospheric composition of different bodies, stars/planets life cycle and formation process and more.

* **Kind of telescope and wavelength range**: 2 paragraphs, 7-8 sentences. The user should provide detailed information regarding what instrument is being used, which includes technical specifications such as focal length of the telescope and eyepieces, apertures, focal ratios, the type of telescope and also what kind of additional items are being used like filters and cameras to detect that special electromagnetic spectrum. Providing what wavelength the observation is going to be in can surely help, infrared radiation varies in different nano-metrica (NIR, MIR, FIR), each wavelength is good for a certain type of observation.

Output Sections:

* **Observing conditions**: 1 paragraph, 3-4 sentences. We need to know what conditions the observation is taking place and what location, mostly because this plays a real key factor in terms of how the observation would be carried out. Providing information about things like light pollution (Bortle scale), density of air, outside temperature and more.

* **Instrument sensitivity**: 1 paragraph, 2-3 sentences. We need to know what are the specifications of sensitivity of the instruments being used. Sensors or cameras can detect so many details in comparison to the human eye, as well as filters of eyepieces.

EXAMPLE INPUT

* **Scientific/main Goal AND target object**: Our scientific goal is to study the formation and evolution of young stars within the Orion Nebula. The Orion Nebula is a vast region of star formation, and by observing it in the infrared, we can penetrate the dust and gas that obscure these young stars in visible light.

* **Kind of telescope and wavelength range**: We will use the Stratospheric Observatory for Infrared Astronomy (SOFIA), a 2.7-meter infrared telescope mounted on a Boeing 747SP aircraft. SOFIA is capable of observing in the mid-infrared and far-infrared wavelengths, from 5.4 to 25.5 microns. We will specifically target the mid-infrared range (8-13 microns), which is optimal for detecting the warm dust and gas associated with young stellar objects. Additionally, we will employ a medium-resolution spectrograph to obtain detailed information about the composition and physical properties of the observed targets.

REFERENCE OUTPUT

* **Observing conditions**: The observations will be conducted from an altitude of approximately 45,000 feet, above the majority of Earth's water vapor, which absorbs infrared radiation. The flight will take place during a moonless night to minimize light pollution. The observing site will be chosen to have a low Bortle scale rating, indicating minimal light pollution from artificial sources.

* **Instrument sensitivity**: SOFIA's mid-infrared camera, FORCAST, has a sensitivity of approximately 0.1 Jy for a point source in a 10-minute integration time. The spectrograph has a spectral resolution of R 1000, allowing us to resolve individual spectral lines and determine the chemical composition of the target objects.

Gemini-1.5-0409 OUTPUT

* **Observing conditions**: SOFIA operates at altitudes of 38,000-45,000 feet, above 99% of the Earth's water vapor, which allows for clear observations in the infrared spectrum. The observations will be conducted during nighttime flights to minimize background radiation from the Sun. Additionally, SOFIA's mobility allows it to choose observing locations with minimal light pollution, typically over remote ocean areas or sparsely populated regions, corresponding to Bortle Class 1 or 2.

* **Instrument sensitivity**: SOFIA's mid-infrared instruments, such as the FORCAST camera and the FLITECAM spectrograph, are highly sensitive to faint infrared emissions. They can detect temperature variations as small as a few degrees Kelvin and resolve spatial details down to a few arcseconds. This sensitivity enables the detection and characterization of young stellar objects and their surrounding environments within the Orion Nebula.

Table 8: Sample response showcasing lack of detail in the generated output.

Survey on the Use of AI Systems for Expert Tasks

Goal: We are a group of researchers and we are interested in studying the use of AI systems for assisting experts from different fields. With your help, we would like to study the practical and societal implications of using AI systems as tools for assisting experts from different fields. Thank you for participating!

Instructions: In this study, you will be shown the description of a textual task from your field written by an expert from your field. Each task has input sections and output sections, and any system would need to produce the content in the output sections given the content in the input sections. We would like you to first read this task description closely, and then answer the questions that follow below.

We would like you to imagine a scenario where an expert uses an AI system to get an initial output of this task from an AI system, which they can then improve and validate. Essentially, they would be using an AI system as a tool, the same way a mathematician might use a calculator as a tool. The questions listed below ask about the practical and societal implications of using AI systems this way, where they are used in the loop as experts conduct their tasks.

Notes:

- For all questions under societal implications, you are required to write 2-5 sentences elaborating upon your choice.
- Please try your best to provide high-quality labels. If you end up taking more time than allocated, we would be happy to bonus you.

Task Description

Task Objective: The task is to figure out which regional varieties of spoken languages are used by certain speakers in a region

Task Procedure: The persons who are chosen for the task are required to read some sentences in the pronounciation they usually speak when communicating with members of the family or friends. Those sentences are transcribed by an expert.

Additional Notes: Common mistakes: Is there any code-mixing or code-switching? Is the design of the exploration a bias for the people involved?

Input Sections**:

- Interview: 2 paragraphs & Answers to different questions in terms of the biography: age, gender, religion, date and place of birth, family members, date and place of birth of mother and father, residences.
- Exploration in regional variant: Transcribed readings and/or writings of a prepared set of sentences in the domestic variant of language.
- Exploration in standard language: Transcribed readings and/or writings of the same sentences as in section 2 in the standard language.

Output Sections**:

- Names of regional variants: The names of regional variants spoken by the person or group of people, identified through their transcribed readings and writings.

Representativeness: How likely is this task to be conducted by an expert in your field?

- Very likely: This is a very common task that most or all experts in my field work on.
- Likely: This is a common task that some experts in my field would work on.
- Possibly: This is not a common task but it is conceivable that some experts in my field would work on this.
- Unlikely: This does not seem like a task that any expert from my field would work on.

Complexity: How would you rate the complexity of this task?

- High: This is a highly technical task that requires immense field expertise (several years of experience).
- Medium: This is a technical task that requires a moderate amount of field expertise (a few years of experience).
- Low: This is not a very technical task and requires very little expertise in my field.

Time Required: How much time is typically required to complete this task?

- Less than an hour
- 1-4 hours
- 1-7 days
- 1-4 weeks
- 1-3 months
- >3 months

Usefulness: Would you or other experts find it useful if an AI system could be used to propose initial outputs for this task (which may be lacking), that can be validated and improved by experts?

- Yes: I would be interested in using this system as it could improve my efficiency.
- Maybe: It is not immediately obvious that this system would be helpful for the given task, but I am open to trying it and deciding based on that.
- No: I would not be interested in using this system because it would not improve my efficiency.

Societal Implications

Anonymity Required: Is it important to ensure anonymity of any individuals or organizations mentioned if an AI system is used for conducting this task? This may be the case if there is sensitive information in the input that should not be stored or accidentally leaked by an AI system.

- Yes: It is imperative, as there might be sensitive information about individuals or organizations in the input for this task. (Please elaborate in the text box below)
- No: It is not necessarily a concern for this task.

Please elaborate on your choice in brief.

Biased Outcomes: Could relying on automatically generated outputs for this task result in biased or potentially harmful decisions for certain groups of people?

- Yes: AI systems can potentially exhibit bias in the outputs of this task, which can negatively affect people from specific groups or demographics. (Please elaborate in the text box below)
- No: I don't think that automatically generated outputs for this task could lead to biased outcomes for people from any specific group or demographic.

Please elaborate on your choice in brief.

Ethical Considerations: Are there ethical considerations associated with the use of AI systems for this task? This can include privacy issues, moral issues, copyright issues or any other issues.

- Yes: There is at least one ethical issue which is important to consider. (Please elaborate in the text box below)
- No: I don't foresee any ethical concerns related to the use of AI systems for this task.

Please elaborate on your choice in brief.

Workforce Impact: Could partial automation of this task potentially have an impact on the workforce in the short term?

- Likely: Partial automation is likely to affect the workforce in my field in the short term.
- Unsure: Unable to make an informed judgment about this task.
- Unlikely: Partial automation is unlikely to affect the workforce in my field in the short term.

Please elaborate on your choice in brief.

Accessibility Requirements: Does the use of AI systems for this task require making exceptional considerations for ensuring accessibility to all users? For instance, a task that requires producing visual outputs might pose challenges for people with visual or motor disabilities.

- Yes: It is critical to pay special attention to adapting the output for this task to users with special needs. (Please elaborate in the text box below)
- No: Beyond general considerations in making technology accessible to all users, I cannot identify any exceptional considerations for users with special needs.

Please elaborate on your choice in brief.

Figure 11: Interface shown to annotators for task validation.

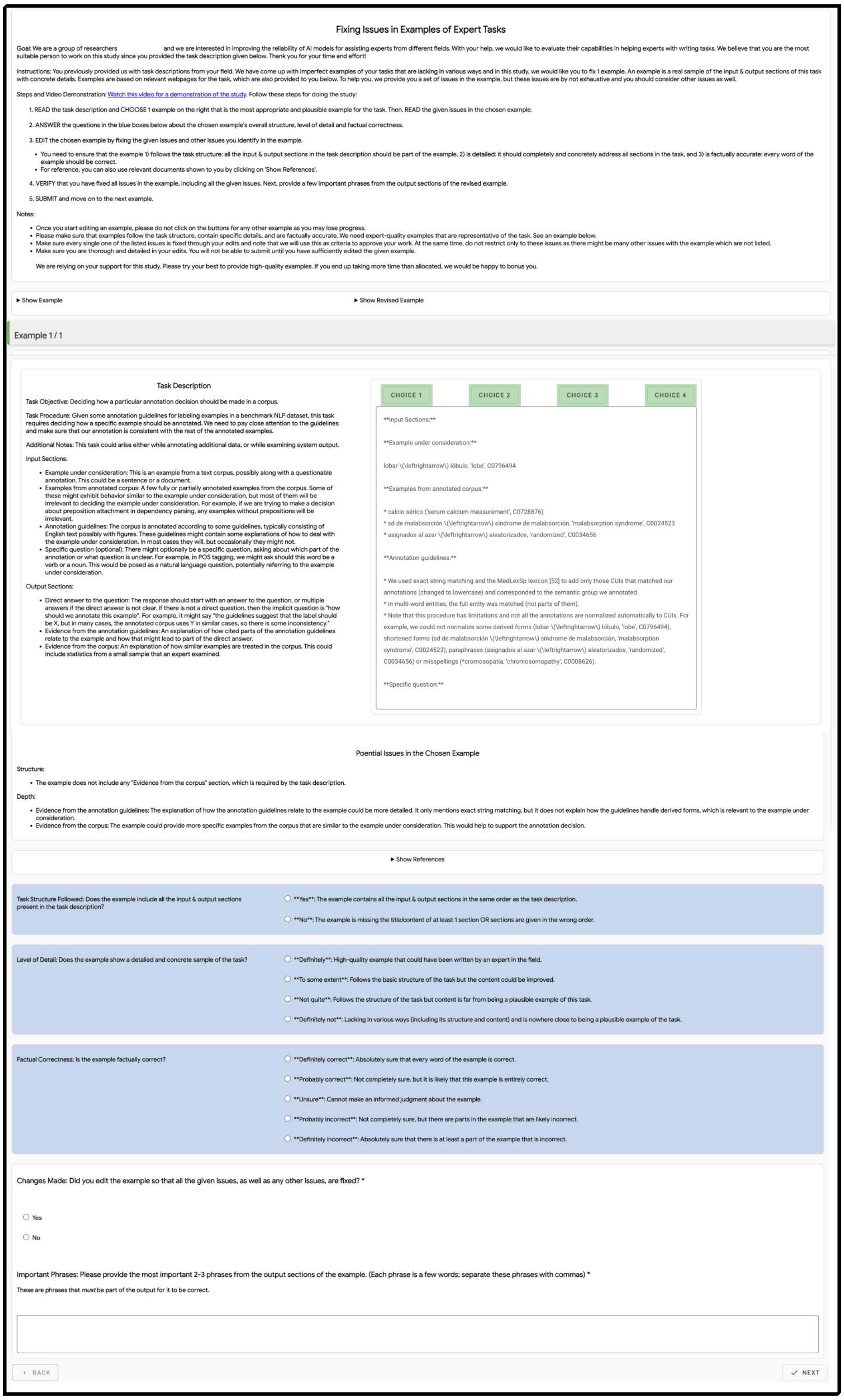

Figure 12: Interface shown to annotators for example post-editing.

Pattern: Verbosity

TASK DESCRIPTION

Task Objective: Write a review of a novel.

Task Procedure: You need to write an extensive review of a novel, you need to summarize and give basic facts of the writer and novel, but also give value judgement.

Additional Notes: nan

Input Sections:

* **Title**: only one or several words, this is the title of the book that you will review.

* **Writer**: only a name, this is the author's name.

* **Pages**: only a number, the number of pages in the novel.

* **Introduction**: 1 paragraph, a general introduction to both the author and their work (the title that will be reviewed in the article)

* **Summary**: 1 paragraph, a specific summary of the book which gives the reader an insight into the story without spoiling major events or the ending; it needs to stay exciting/enticing.

Output Sections:

* **Analysis**: 2-3 paragraphs, varying analyses of several aspects of the story, this includes content analysis, but also linguistic and semantical analysis; what is meant by certain passages? What are the cultural/societal connotations? In what way does the writer use language and form?

* **Verdict**: 1 paragraph, what do we think about the things we just analysed? Is it of high quality or low quality? And in what ways? Why?

* **Conclusion**: 1 sentence, summarize everything said above and give a final review in only one sentence.

EXAMPLE INPUT

* **Title**: Dead in Long Beach, California

* **Writer**: Venita Blackburn

* **Pages**: 157

* **Introduction**: A rising literary star, Venita Blackburn, tackles the complexities of grief and unconventional coping mechanisms in her captivating novel, "Dead in Long Beach, California." This poignant tale explores the depths of human sorrow through the story of a young woman grappling with loss. Blackburn delves into the lengths individuals go to find solace, weaving a narrative that is both heartbreaking and thought-provoking.

* **Summary**: The novel follows the journey of a young woman who, after the sudden death of her brother, stumbles upon an unexpected and ethically dubious method of coping. Posing as her deceased sibling, she immerses herself in his online world, seeking comfort in the digital memories and connections he left behind. As she delves deeper into her brother's virtual existence, she uncovers hidden secrets and confronts the consequences of her actions.

REFERENCE OUTPUT

* **Analysis**: Blackburn's novel is a profound exploration of the grieving process. Through the protagonist's unconventional coping mechanism, she examines the ways in which technology can both connect and isolate us in times of loss. The author skillfully intertwines the protagonist's emotional turmoil with the digital landscape, creating a thought-provoking commentary on the nature of identity and the boundaries between the living and the dead.

Furthermore, Blackburn's prose is both lyrical and incisive. For instance, her description of the protagonist staring at her brother's empty online profile, "a hollow echo of a life once lived," perfectly captures the profound sense of emptiness and longing she experiences. The novel's exploration of ethical dilemmas, such as the protagonist's manipulation of online identities, is handled with sensitivity and nuance. It leaves readers to ponder the complexities of right and wrong in the face of overwhelming grief, sparking valuable discussions about the boundaries of acceptable coping mechanisms.

* **Verdict**: "Dead in Long Beach, California" is a powerful and deeply moving novel. Blackburn's ability to capture the raw emotions of grief and the complexities of our digital lives is commendable. Her exploration of unconventional coping mechanisms and the ethical implications they raise is both thought-provoking and timely.

* **Conclusion**: Venita Blackburn's "Dead in Long Beach, California" is a haunting and unforgettable tale that lingers long after the last page is turned. It serves as a powerful reminder of the enduring strength of human connection and the profound bond that transcends even death.

GPT-4 OUTPUT

* **Analysis:** Venita Blackburn's "Dead in Long Beach, California" is a masterful exploration of grief and identity in the digital age. Through her concise yet powerful prose, Blackburn challenges the reader to consider the boundaries between the self and the other, the living and the dead. The novel's central theme of coping with loss through digital impersonation raises poignant questions about the authenticity of online relationships and the ethics of assuming another's identity posthumously. Blackburn's use of language is both evocative and precise, capturing the protagonist's turmoil and desperation with striking clarity. The narrative structure, alternating between the protagonist's real-life struggles and her online escapades, effectively mirrors the fragmented nature of grief, enhancing the reader's empathy for the protagonist's plight. Furthermore, Blackburn subtly incorporates societal critiques, particularly concerning the internet's role in shaping our perceptions of death and mourning. The novel's setting in Long Beach, California, with its juxtaposition of urbanity and the vastness of the Pacific, further underscores the protagonist's feelings of isolation and her quest for connection in an increasingly disconnected world.

* **Verdict**: "Dead in Long Beach, California" is a compelling and thought-provoking novel that showcases Venita Blackburn's talent for delving into complex emotional landscapes with nuance and sensitivity. The quality of Blackburn's writing is evident in her ability to weave a deeply personal story with broader societal implications, making the novel both intimate and universally relevant. The ethical dilemmas and emotional depth explored in the book are handled with a deftness that invites readers to reflect on their own experiences of loss and the ways in which technology intersects with our most human experiences. Blackburn's linguistic prowess and her skillful narrative construction make this novel a high-quality literary work that is both engaging and enlightening. It is a poignant reminder of the power of fiction to explore the nuances of the human condition.

* **Conclusion**: Venita Blackburn's "Dead in Long Beach, California" is a beautifully crafted novel that offers a profound and nuanced exploration of grief in the digital age, making it a must-read for those who appreciate literature that delves into the complexities of the human heart.

Table 9: Sample response showcasing verbosity in the generated output (note the *Analysis* and *Verdict* sections).

Pattern: Missing Information

<u>TASK DESCRIPTION</u>

Task Objective: Writing a funding application for a medical research project.

Task Procedure: To write a funding application for a research project, you need to start with a clear and concise idea for the research project, identify matching funding sources, understand the funder's priorities. Then, write a detailed project description and prepare a specific budget. Complete the application form provided by the funder and follow all guidelines. Review the application, submit it, and await feedback.

Additional Notes: It is important to clearly define the research idea, understand the funder's priorities, write a detailed project description, prepare a specific budget, and review before submitting. Common mistakes include overlooking eligibility, using jargon, poor structure of the application, and an overwhelming page layout.

Input Sections:

* **Project Idea**: 1 paragraph. A clear and concise idea for the research project is essential.

* **Potential Funding Sources**: 1 paragraph. Estimated cost and identification of potential funding sources that align with the research topic.

Output Sections:

* **Funding application**: 2 paragraphs. The funding application will include a summary, an introductory section including background and motivation, detailed project description, and detailed budget.

* **Timeline**: In addition, a detailed timeline with key milestones is essential to demonstrate preparedness and careful planning for the project.

<u>EXAMPLE INPUT</u>

* **Project Idea**: We aim to develop a novel risk prediction model for asthma exacerbations in children aged 1-6 years. By integrating clinical data, environmental factors, and genetic markers, we will create a comprehensive tool to identify children at high risk of exacerbations, enabling early intervention and improved disease management.

* **Potential Funding Sources**: The National Heart, Lung, and Blood Institute (NHLBI) offers funding opportunities for research on respiratory diseases, including asthma. We estimate that the project will require a budget of $250,000 over five years.

<u>REFERENCE OUTPUT</u>

* **Funding Application**:

Summary: This application proposes a comprehensive study to develop a risk prediction model for asthma exacerbations in children. The model will integrate clinical, environmental, and genetic data to identify high-risk individuals, facilitating early intervention and improved disease management.

Introduction: Asthma is a common chronic disease that affects millions of children worldwide. Exacerbations can be life-threatening and lead to significant healthcare costs. Current prediction methods are inadequate, and there is an urgent need for more accurate tools. This project will address this gap by developing a novel risk prediction model that combines multiple factors to provide a personalized assessment of exacerbation risk.

Project Description: We will recruit 500 children aged 1-6 years with asthma from the Aspen Medical Group and St. Paul Children's Hospital. Data collection will include clinical assessments, environmental monitoring, and genetic testing. Baseline clinical assessments will include use of short-acting beta agonists, previous asthma-related emergency department attendances and hospitalizations, history of lower respiratory tract infections, environmental tobacco exposure. Environmental monitoring will be performed using indoor air-quality sensors. Genetic testing will include the analysis of specific genes linked to asthma. Outcome events will include asthma-related emergency hospitalization. We will use survival models to estimate the hazard ratio associated with key predictors. The model will be validated using an independent cohort of children.

Budget: The total budget for the project is $250,000 over five years. This includes costs for participant recruitment, data collection, analysis, and dissemination. The budget includes:
* $150,000: Staff salaries
* $50,000: Patient recruitment data, collection and analysis
* $30,000: Equipment and supplies
* $20,000: Travel expenses

* **Timeline**:
* Year 1: Participant recruitment and data collection: This include three months for participant recruitment and nine months for data collection.
* Year 2: Model development and validation: This includes nine months for model development and three months for validation.
* Year 3: Model refinement, implementation, and real-world testing.
* Year 4: Dissemination of results and exploration of clinical applications: This include three months for preparing and publishing findings, and nine months for exploring clinical applications.
* Year 5: Project evaluation and future directions: The entire year will be dedicated to project evaluation and planning for future research.

<u>GPT-4 OUTPUT</u>

* **Funding Application**: Our project proposes the development of an innovative risk prediction model for asthma exacerbations in children aged 1-6 years, a critical period for asthma management and intervention. By leveraging a multidisciplinary approach that combines clinical data, environmental factors, and genetic markers, our model aims to revolutionize how healthcare providers identify children at high risk of asthma exacerbations. This early identification will facilitate timely interventions, potentially reducing the frequency and severity of exacerbations and improving overall disease management and patient outcomes.

The project will unfold in several phases, beginning with the collection and analysis of clinical data from pediatric asthma patients. Concurrently, we will conduct a comprehensive review of environmental factors known to influence asthma exacerbations, such as air quality and allergen exposure, and perform genetic analyses to identify markers associated with increased risk. This integrated data will form the basis of our predictive model, which will be rigorously tested and validated in clinical settings. To support this ambitious project, we are seeking funding of $250,000 from the National Heart, Lung, and Blood Institute (NHLBI), an organization that prioritizes research on respiratory diseases, including asthma. This budget will cover costs associated with data collection, analysis, personnel, and dissemination of findings, ensuring a thorough and impactful study.

* **Timeline**:

Year 1:
- Q1-Q2: Project initiation, including team assembly and training, and development of data collection protocols.
- Q3-Q4: Begin data collection on clinical data and environmental factors, and initiate genetic marker identification.

Year 2:
- Q1-Q2: Continue data collection and start preliminary data analysis.
- Q3-Q4: Development of the initial model based on collected data and initial testing.

Year 3:
- Q1-Q2: Refinement of the model based on initial testing feedback and integration of additional data as necessary.
- Q3-Q4: Begin large-scale validation of the model in clinical settings.

Year 4:
- Q1-Q4: Continued validation and refinement of the model, including adjustments based on ongoing feedback from clinical applications.

Year 5:
- Q1-Q2: Finalization of the model and preparation of materials for dissemination.
- Q3-Q4: Dissemination of findings through publications, presentations at conferences, and direct outreach to healthcare providers.

This timeline demonstrates our commitment to careful planning and preparedness, ensuring that each phase of the project builds upon the last towards the successful development and implementation of our novel asthma risk prediction model.

Table 10: Sample response showcasing missing information in the generated output (note the *Funding Application* section).

Query Formulation Prompt

```
Generate 10 search queries for finding specific examples of the given task from the specified
field. The search queries should be brief and request documents in more specific contexts than the
given task. We would like the documents to contain real examples of the task. List the queries and
nothing else.

FIELD: Visual Arts (Graphic design)
TASK: The objective of this task is to write a catalog entry for an art exhibition.
QUERIES: 1) Example of catalog entry for art exhibition
2) Catalog entry art exhibition Dali
3) Notable art catalog entries 2023
4) memorable art catalog entries 2000s
5) catalog entry for jackson pollock painting
6) frida kahlo painting catalog entry
7) picasso guernica catalog entry
8) da vinci mona lisa catalog entry
9) The Great Wave off Kanagawa catalog entry
10) renaissance art exhibition catalog entry

FIELD: [FIELD]
TASK: [TASK]
QUERIES:
```

Table 11: Prompt used for generating specific search queries for a task.

Domain Name Prompt

```
List 20-30 URLs to domain names which will be useful to find real examples for the given task.
These websites should be reliable, trustworthy and authoritative sources for an expert in the
field. They should be ranked by their likely usefulness.

FIELD: Engineering and Technology (NLP research)
TASK:Summarizing related work on an NLP subproblem.
URLs: 1) arxiv.org
2) aclweb.org
3) ldc.upenn.edu
4) nlp.stanford.edu
5) aclanthology.org
6) towardsdatascience.com
7) semanticscholar.org
8) openreview.net
9) medium.com
10) nature.com
11) transacl.org
12) cambridge.org
13) iclr.cc
14) aaai.org
15) academic.microsoft.com
16) nips.cc
17) onlinelibrary.wiley.com
18) link.springer.com
19) naacl.org
20) plos.org

FIELD: [FIELD]
TASK: [TASK]
URLs:
```

Table 12: Prompt used for searching for authoritative domain names for a task.

Example Generation Prompt

```
You are given a description of a task from the field of [FIELD] by an expert. Generate a concrete
example of all the Input Sections and Output Sections listed for the given TASK DESCRIPTION. The
example should resemble a real example that is written by an expert in the field, and should be
highly technical and detailed.

Further instructions:
- You are also given CONTEXT in the form of Passages from web documents and will need to generate
an example based on this CONTEXT. Make sure to generate the example based on the provided CONTEXT.
If the CONTEXT is insufficient, you can say "The context is insufficient".
- Make sure the length of each section matches the required length and the section headers are
exactly the same.
- The example should be highly detailed, and not be generic and vague.

====CONTEXT====

[CONTEXT]

====TASK DESCRIPTION====

[TASK DESCRIPTION]

====EXAMPLE====
```

Table 13: Prompt used for generating initial examples for a task.

Critique Generation Prompt

```
You are an expert in the field of [FIELD]. You are given a task description of a writing task from
your field and an imperfect example for this task, where an example is a concrete sample of the
task. You need to describe what is lacking in the example for the task. You are given a list of
properties based on which you should critique the example:

* Inconsistencies: Are there any inconsistencies in the information provided across the input and
output?
* Factual Inaccuracies: Are there any factual inaccuracies in the information presented in the
input or how the output is inferred?
* Structure: Are there any issues with how closely the example follows the instructions specified
in the task description? This includes information requested in the task but missing in the
example, or mismatch in the length required for a section.
* Depth: How could the example benefit from more detail? Note that the example should resemble
what an expert might write and so it should not be vague with details.

====TASK DESCRIPTION====

[TASK DESCRIPTION]

====EXAMPLE====

[EXAMPLE]

====Critique====
```

Table 14: Prompt used for generating critiques for model-generated examples.

Output Generation Prompt

```
You need to perform a writing task from the field of [FIELD]. You are given (1) a task description
which contains input and output sections, and (2) an example input for this task, which is a
sample of the input sections of the task with concrete details. You need to generate the output
sections for the given example input.

- Make sure the length of each output section matches the required length and the section headers
are exactly the same.
- Make sure the output follows the structure of the output sections in the task description, is
factually accurate and detailed.

====TASK DESCRIPTION====

[TASK DESCRIPTION]

====EXAMPLE INPUT====

[EXAMPLE INPUT]

====EXAMPLE OUTPUT====
```

Table 15: Prompt used for generating outputs from candidate models for evaluation.

LM-based Overall Preference Evaluation Prompt

```
You are an expert in the field of [FIELD]. You are given a task description of a writing task from
your field. For this task description, you are given an input example, which is a concrete sample
of the input sections of this task, as well as the reference output, which is the gold standard
output for this input. You will be given two candidate outputs for the input example and you need
to judge which output is better by comparing it to the reference output.

First, you should say "**output 1**" if output 1 is better, "**output 2**" if output 2 is better
and "**same**", if the two outputs are equivalent in quality (note the stars). Then you should
explain why you picked this output.

**Important: Keep in mind that longer outputs are not necessarily better quality outputs. Being
concise is a good quality for outputs.**

====TASK DESCRIPTION====

[TASK DESCRIPTION]

====INPUT EXAMPLE====

[EXAMPLE INPUT]

====REFERENCE OUTPUT====

[REFERENCE OUTPUT]

====EXAMPLE OUTPUT 1====

[EXAMPLE OUTPUT 1]

====EXAMPLE OUTPUT 2====

[EXAMPLE OUTPUT 2]

====Decision====
```

Table 16: Prompt used for generating LM-based judgments.

LM-based Fine-Grained Evaluation Prompt

```
You are an expert in the field of [FIELD]. You are given a task description of a writing task from
your field. For this task description, you are given an input example, which is a concrete sample
of the input sections of this task, as well as the reference output, which is the gold standard
output for this input. You will be given a candidate output for this input example. You need to
evaluate the output by comparing it to the reference output. We will also give you a rubric, which
should guide your evaluation. You need to rate the output on the rubric on a scale of 1-5.

Here is the rubric based on which you should evaluate the outputs:
* Adherence to Task Structure: The output should closely follow the instructions specified in the
output sections of the task description. The information requested in the task description should
be present in the output, and the sections should be the correct length.
* Factual Accuracy: There should not be any factual inaccuracies or inconsistencies in the output.
* Depth: The output text should be technically detailed and thorough, so that it resembles how an
expert might conduct the task.
* Completeness: The output text should be complete and contain all the information requested in
the task description.
* Coherence: The output text should flow logically and be easily understandable.

You should produce the final output as a dictionary in precisely this format: "**output:
"adherence": _, "accuracy": _, "depth": _, "completeness": _, "coherence": _**", where you should
fill in the spaces with ratings. Make note of the ** required to enclose the output dictionary.

====TASK DESCRIPTION====

[TASK DESCRIPTION]

====INPUT EXAMPLE====

[EXAMPLE INPUT]

====REFERENCE OUTPUT====

[REFERENCE OUTPUT]

====EXAMPLE OUTPUT====

[EXAMPLE OUTPUT]

====Output====
```

Table 17: Prompt used for generating LM-based judgments.