

461 A Example PyTorch Implementation of Gist Masking

462 See Listing [A.1](#) for a sample annotated implementation of gist masking. This PyTorch implementation
463 relies on basic NumPy-style tensor operations and can thus be adapted easily to a framework like
464 JAX.

465 B Data, Training, Evaluation, and Compute Details

466 Code and data are available in the supplementary material and will be made public upon paper
467 acceptance via GitHub. Single gist token model checkpoints (positive control, gist, negative control)
468 will also be made public via Hugging Face.

469 **Data.** For LLaMA-7B, we used a maximum sequence length of 512 tokens during training and
470 evaluation, except with the Human validation split, where the maximum length was increased to
471 768 (the Human instructions are longer). Examples longer than this length are truncated from the
472 end. For FLAN-T5-XXL, we set a maximum input length (task t + input x) of 128 and a maximum
473 output length of 256, except again for the Human split, where the maximum input and output lengths
474 were both set to 384. For both models, we set a maximum generation length of 512 tokens. These
475 lengths were chosen such that $< 1\%$ of examples across the board were truncated during training and
476 evaluation for both models.

477 **Training.** Full hyperparameters for training runs are located in Table [A.1](#). These parameters were
478 adapted from previous published work finetuning LLaMA/FLAN-T5. For LLaMA-7B, parameters
479 are identical to those used in training Alpaca Taori et al. [\[31\]](#). For FLAN-T5-XXL, parameters are
480 identical to those used in training Tk-INSTRUCT [\[37\]](#), except with a $5e-5$ learning rate, as used in the
481 Tk-INSTRUCT GitHub repository [\[3\]](#) rather than the $1e-5$ learning rate in the paper.

482 LLaMA-7B was trained for 3000 steps, while FLAN-T5-XXL was trained for 16000 steps. Since
483 there are about 130k examples in Alpaca+, given the batch sizes in Table [A.1](#) this corresponds to
484 about ~ 3 epochs and ~ 2 epochs of training, respectively. These numbers, again, are identical to Taori
485 et al. [\[31\]](#) and Wang et al. [\[36\]](#). We note that the training time is relatively flexible; for example, we
486 did not see substantial gains training beyond 1 epoch for FLAN-T5-XXL.

487 **Evaluation.** During evaluation and benchmarking, we simply greedily decoded the most likely
488 sequence. We saw limited gains from beam search with beam size $B = 4$.

489 **Compute.** Experiments were run on a cluster machine with 4xA100-SXM4-80GB NVIDIA GPUs,
490 480GB RAM, and 16 CPUs, using PyTorch 2.0 [\[24\]](#), Hugging Face Transformers [\[41\]](#), and DeepSpeed
491 [\[29\]](#). Training runs take about ~ 7 hours to complete for LLaMA-7B and ~ 25 hours for FLAN-T5-
492 XXL. Benchmarking results were obtained on the same machine, but using just 1 of the A100
493 GPUs.

494 C Details of ChatGPT Evaluation

495 We used the ChatGPT API, specifically the chatgpt-3.5-turbo engine, to run our ChatGPT evalua-
496 tion experiments over a period of 2 weeks between March 27 and April 7, 2023.

497 The full prompt given to ChatGPT is located in Listing [A.2](#), and contains 4 examples of desired
498 output from ChatGPT, including preferring factually accurate responses (Example 1), preferring
499 responses that follow the instruction, even if imperfect (Examples 2 and 3), and examples of models
500 being equally wrong (Examples 4). For the two models under comparison, we randomized the order
501 of presentation of each model as either Assistant A or Assistant B, to avoid order effects.

502 ChatGPT was instructed to only respond in JSON format, outputting first a reason key followed by a
503 choice key, to encourage chain-of-thought reasoning [\[39\]](#). On rare occasions ($< 0.25\%$ of the time),
504 ChatGPT would output a response that did not conform to the requested JSON format (e.g. it would

https://github.com/yizhongw/Tk-Instruct/blob/1ab6fad/scripts/train_tk_instruct.sh

```

1 import torch
2
3
4 def reverse_cumsum(x: torch.Tensor) -> torch.Tensor:
5     """Cumulative sum from right to left.
6
7     See https://github.com/pytorch/pytorch/issues/33520.
8     """
9     return x + torch.sum(x, dim=-1, keepdims=True) - torch.cumsum(x, dim=-1)
10
11
12 def make_mask_pre_first_gist(inputs: torch.Tensor, gist_token: int, dtype=torch.int64) -> torch.Tensor:
13     """Returns a mask where all tokens prior to the first gist token are masked out.
14
15     Args:
16         inputs: a Tensor of input tokens where the last dimension is the sequence length.
17         gist_token: the integer id of the gist token.
18         dtype: the dtype of the mask, default int64.
19     Returns:
20         The requested mask.
21     """
22     return ((inputs == gist_token).cumsum(-1) >= 1).type(dtype)
23
24
25 def make_mask_post_last_gist(inputs: torch.Tensor, gist_token: int, dtype=torch.int64) -> torch.Tensor:
26     """Returns a mask where all tokens after the last gist token are masked out.
27
28     Computes the same as mask_pre_first_gist_token, but reverses the sequence before and after the cumsum.
29
30     Args:
31         inputs: a Tensor of input tokens where the last dimension is the sequence length.
32         gist_token: the integer id of the gist token.
33         dtype: the dtype of the mask, default int64.
34     Returns:
35         The requested mask.
36     """
37     return (reverse_cumsum(inputs == gist_token) >= 1).type(dtype)
38
39
40 def make_gist_mask(inputs: torch.Tensor, gist_token: int, pad_token: int, dtype=torch.int64) -> torch.Tensor:
41     """Creates a gist mask from supplied inputs and gist/pad tokens.
42
43     Tokens after the last gist cannot attend to tokens prior to the first gist. Additionally, tokens *before*
44     the last gist cannot attend to tokens *after* the last gist.
45
46     The gist mask is broadcasted to 4D (with a singleton dim 1) for compatibility with multi-headed attention
47     (where dim 1 is the head dimension).
48
49     Args:
50         inputs: a Tensor of shape (batch_size, seq_len) input tokens.
51         gist_token: the integer id of the gist token.
52         pad_token: the integer id of the pad token. inputs == pad_token are masked out.
53         dtype: the dtype of the mask, default int64.
54     Returns:
55         The requested mask of shape (batch_size, 1, seq_len, seq_len)
56     """
57     # Attention mask for tokens before the last gist token.
58     pre_gist_mask = make_mask_post_last_gist(inputs, gist_token, dtype=torch.bool)[: , None, None]
59     # Attention mask for tokens after the last gist token.
60     post_gist_mask = make_mask_pre_first_gist(inputs, gist_token, dtype=torch.bool)[: , None, None]
61     # Construct time masks by permuting to time dimension.
62     pre_gist_time_mask = pre_gist_mask.permute((0, 1, 3, 2))
63
64     mask = torch.where(pre_gist_time_mask, pre_gist_mask, post_gist_mask)
65     mask = mask & (inputs != pad_token)[: , None, None] # Mask out pad tokens.
66
67     return mask.type(dtype)

```

Listing A.1: Sample PyTorch implementation of gist masking.

Table A.1: **Hyperparameters for training runs.**

	LLaMA-7B	FLAN-T5-XXL
num steps	3000	16000
num train epochs	≈ 3	≈ 2
batch size	128	16
learning rate	2e-5	5e-5
warmup ratio	0.03	0
precision	bf16	bf16
optimizer	AdamW	AdamW
Deepspeed		
# GPUs (A100 80GB)	4	4
ZeRO stage	3	3
subgroup size	1e9	1e9
max live params	1e9	1e9
max reuse distance	1e9	1e9

Table A.2: **Pairwise Cohen’s κ** between human annotators (H1, H2, H3) and ChatGPT.

(a) LLaMA-7B					(b) FLAN-T5-XXL				
	H1	H2	H3	Average		H1	H2	H3	Average
H1	–	.21	.33	.27	H1	–	.33	.34	.34
H2	.21	–	.19	.20	H2	.33	–	.33	.33
H3	.33	.19	–	.26	H3	.34	.33	–	.33
ChatGPT	.38	.22	.26	.29	ChatGPT	.35	.18	.34	.29

505 just give an unstructured paragraph). In these cases we manually went through and converted these
 506 responses to JSON, without altering ChatGPT’s reasoning.

507 In total, we collected ~22.5k judgments from ChatGPT for an estimated cost of \$29.28. The full
 508 outputs for each model across the Alpaca+ validation splits, as well as ChatGPT’s responses and
 509 choices, are available in the code link above.

510 D Additional Human Evaluation Details and Results

511 D.1 Experimental Details

512 For each of the 100 examples randomly selected from the Human validation split, we recruited 3 US
 513 or UK-based, English-fluent annotators from Prolific, an online crowdsourcing platform. Experiments
 514 were IRB approved under a generic human experiments IRB given to the authors.

515 The annotation interface given to Prolific crowdworkers is located in Figure [A.1](#). To verify task
 516 comprehension, participants were shown two simple examples before the main body of the task
 517 (Figure [A.2](#)), and were required to answer correctly before proceeding. We compensated participants
 518 USD \$14.35/hour for an estimated cost (including Prolific fees) of USD \$141.64.

519 D.2 Additional Results

520 See Table [A.2](#) for a breakdown of Cohen’s κ between human annotators and ChatGPT. We used a
 521 weighted version of Cohen’s κ with linear weights, since the response scale is ordinal (e.g. “tie” is a
 522 closer judgment to “pos control win” than “gist win”).

523 E Exact Match Results

524 See Figure [A.3](#) for a plot of exact match rates for the gist and positive control models (as measured
 525 by exact string match).

Given a user instruction and two AI assistant responses, your job is to select the response that is more helpful, and give reasons why. Judge responses holistically, paying special attention to whether the response (1) correctly follows the provided instruction and (2) is factually accurate. If both responses are equal, you may call a tie, but please use this response sparingly.

Example 1:

```
Instruction: Given the name of a city, tell me the name of the state it is in.
Input: Los Angeles
Assistant A: California
Assistant B: Wyoming
```

Your output should be:

```
{"reason": "Los Angeles is in California. Only Assistant A is correct.", "choice": "A"}
```

Example 2:

```
Instruction: Give me some examples of fun things to do with the kids on a weekend.
Assistant A: For example, what should I do with my kids on a weekend? What if I'm in a city? What if I'm on the beach?
Assistant B: You could try going to the zoo or a museum. If you want to stay at home, you could play board games or bake cookies.
```

Your output should be:

```
{"reason": "Assistant A doesn't seem to be following the instruction. Assistant B provides helpful information with examples.", "choice": "B"}
```

Example 3:

```
Instruction: Write me a python function that prints "Hello, World".
Assistant A: def hello_world(): print("Hello!")
Assistant B: "Hello, World"
```

Your output should be:

```
{"reason": "Assistant B just wrote 'Hello, World'. Assistant A actually wrote a Python function, even if it doesn't exactly print the right thing, so overall Assistant A is better.", "choice": "A"}
```

Example 4:

```
Instruction: Translate the following sentence from English to French.
Input: I like cats.
Assistant A: Me gustan los gatos.
Assistant B: 我喜欢猫.
```

Your output should be:

```
{"reason": "Both assistants got the language wrong.", "choice": "tie"}
```

Your response should only be in the JSON format above; THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE. Write the "reason" key before writing the "choice" key, so that you think step-by-step before making your decision. KEEP YOUR REASONING BRIEF.

Listing A.2: Full prompt given to ChatGPT for evaluation. This prompt populates the system field in the ChatGPT API; the actual example to be evaluated is formatted like the examples in the prompt above, then given as the sole input in the user field.

526 F Additional FLOPs details

527 The FLOPs required for a Transformer forward pass with varying KV cache lengths can be estimated
528 by modifying existing equations to account for self-attention back to the KV cache. As an example,
529 we modify the FLOPs equations used for computing FLOPs in the Chinchilla paper (Appendix F in

Judging AI Assistants

Progress: 12/50

User Instruction: Write a good Twitter Bio. Try to cover all the provided main points in your generated bio.

User Input: Name: Jane Main points: - Architect - Interested in Tech Written in first person Use casual tone

Assistant A: Jane is an architect who is interested in tech. She loves to solve problems and create beautiful things.

Assistant B: I'm Jane, an architect who loves to explore the latest tech trends. I'm always looking for new ways to make our lives easier and more convenient. Follow me for the latest tech news, tips, and inspiration!

Which Assistant is more helpful?

A

B

TIE

I DON'T KNOW

Remember:

When making your judgment, please consider whether the response

1. attempts to follow the provided instruction, and
2. is factually accurate.

If you can't decide, e.g. because the responses are the same or equally good or bad, you may call a tie, but please use this response sparingly.

Finally, if you do not understand the question being asked: you may still be able to make an educated guess about which Assistant is better, for example if one assistant is clearly not following the directions. If so, please do! Otherwise, click the "I don't know" button. There is no need to look up additional information or spend a huge amount of time on an example; just click "I don't know."

Figure A.1: Annotation interface given to Prolific crowdworkers.

User Instruction: Give me some examples of fun things to do with the kids on a weekend.

Assistant A: For example, what should I do with my kids on a weekend? What if I'm in a city? What if I'm on the beach?

Assistant B: You could try going to the zoo or a museum. If you want to stay at home, you could play board games or paint photographs.

User Instruction: Find me a synonym of the following word.

User Input: speedy

Assistant A: quick

Assistant B: lazy

Figure A.2: Example items given to humans before the start of the task.

530 [14]). Let $\text{seq_len_with_past} = \text{seq_len} + \text{kv_cache_len}$. Then the modified Transformer FLOPs
531 equations are:

532 Embeddings

533 • $2 \times \text{seq_len} \times \text{vocab_size} \times \text{d_model}$

534 Attention (Single Layer)

535 • *Key, query, and value projections:* $2 \times 3 \times \text{seq_len} \times \text{d_model} \times (\text{key_size} \times \text{num_heads})$

536 • *Key and query logits:* $2 \times \text{seq_len} \times \text{seq_len_with_past} \times (\text{key_size} \times \text{num_heads})$

537 • *Softmax:* $3 \times \text{num_heads} \times \text{seq_len} \times \text{seq_len_with_past}$

538 • *Softmax @ query reductions:* $2 \times \text{seq_len} \times \text{seq_len_with_past} \times (\text{key_size} \times \text{num_heads})$

539 • *Final linear:* $2 \times \text{seq_len} \times (\text{key_size} \times \text{num_heads}) \times \text{d_model}$

540 Dense Block

541 • $2 \times \text{seq_len} \times (\text{d_model} \times \text{ffw_size} + \text{d_model} \times \text{ffw_size})$

542 Final Logits

543 • $2 \times \text{seq_len} \times \text{d_model} \times \text{vocab_size}$

544 Total Forward Pass FLOPs

545 • $\text{embeddings} + \text{num_layers} \times (\text{attention_single_layer} + \text{dense_block}) + \text{final_logits}$

546 It can be seen that only 3 operations in each attention layer depend on the KV cache size, and
547 they take up a relatively insignificant amount of FLOPs. As an illustrative example, Figure A.4
548 shows the relative FLOPs contributions within a single layer of attention for LLaMA-7B, assuming a
549 2000-length KV cache and a single input token. Operations dependent on the KV cache constitute at
550 most ~10% of the total attention layer FLOPs; the rest are used in KQV projections and dense layers
551 for processing the single new input token.

552 Given a KV cache compression rate of 26, as observed in our Human validation split, the Chinchilla
553 equations predict a relative improvement of Gist caching of 0.12%. This is extremely close to the
554 0.11% improvement actually observed in Table 3. These results show that optimizing the KV cache
555 size does not actually lead to huge compute speedups during Transformer inference, at least for
556 relatively small prompt lengths. Nevertheless, there are clear memory and storage benefits to be
557 gained from prompt compression, as discussed in Section 6.

558 G Additional Examples of Model Outputs

559 See Listing A.3 for additional examples of instructions, inputs, model completions (for LLaMA-7B
560 and FLAN-T5-XXL across Pos, Gist, TF-IDF, and Neg conditions), human judgments, and ChatGPT
561 judgments (+reasons). Examples are not cherry-picked; we enumerate the first 10 examples of the
562 Alpaca+ Human validation split.

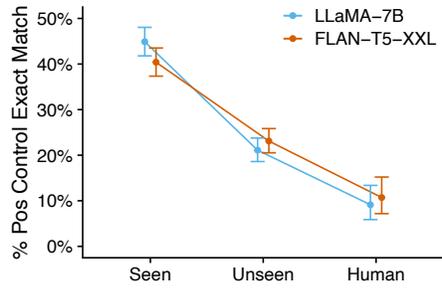


Figure A.3: **Exact match rates.** Rate at which the 1 token gist models give the same output as the positive control (exact string match). Error bars are 95% exact binomial confidence intervals.

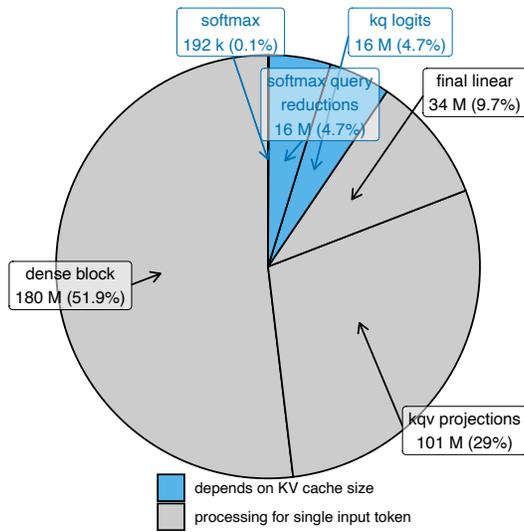


Figure A.4: **FLOPs for each operation involved in a single layer of self attention with a 2000-length KV cache**, according to the Chinchilla estimates for LLaMA-7B. At most 9.6% of FLOPs can be optimized away by reducing the size of the KV cache.

563 Example 1

564 **Instruction:** Write a good Twitter Bio. Try to cover all the provided main points in your generated bio. <G1>

565 **Input:** Name: Jane

566 Main points:

567 - Architect

568 - Interested in Tech

569 Written in first person

570 Use casual tone

571 **Reference Output:** I like to think of myself as a kind of architect.

572 I'm interested in the tech side of things, and I love coming up with new ideas for how to make things better.

573 **TF-IDF Word:** bio

574

575 LLaMA-7B

576 **Pos:** I'm Jane, an architect and tech enthusiast.

577 I'm interested in the latest trends and always looking to learn more.

578 I'm here to share my thoughts and experiences with the world.

579 **Gist:** I'm Jane, an architect and tech enthusiast.

580 I love to explore new places and learn about different cultures.

581 I'm always looking for fun things to do and new experiences.

582 I'm an avid reader and enjoy learning about the world around me.

583 I'm also a big fan of tech and love to stay up to date on the latest trends.

584 I'm looking forward to seeing what the world has to offer!

585 **TF-IDF:** Jane is an architect who is interested in technology.

586 She likes to stay up to date on the latest trends and is always looking for new ways to apply them in her work.

587 Jane is a creative problem solver who approaches each project with an innovative eye.

588 **Neg:** Jane is an architect who is interested in technology.

589 She likes to stay up to date on the latest trends and is always looking for new ways to incorporate them into her work.

590 She's an avid reader of tech news and loves to explore the latest gadgets.

591 Jane is also a big fan of the arts and enjoys visiting museums and galleries when she gets the chance.

592 **Human choice (Pos vs Gist):** Pos Tie Gist

593 **ChatGPT choice (Pos vs Gist):** Pos

594 **ChatGPT reason (Pos vs Gist):** Assistant A covers all the main points provided in the instruction, while Assistant B goes off-topic

595 and doesn't mention the interest in tech as much. Assistant A also has a more professional tone, which is more appropriate

596 for a Twitter bio. (Assistant A: Pos, Assistant B: Gist)

597

598 FLAN-T5-XXL

599 **Pos:** I'm Jane, an architect who loves to explore the latest tech trends.

600 I'm always looking for new ways to make our lives easier and more convenient.

601 Follow me for the latest tech news, tips, and inspiration!

602 **Gist:** Jane is an architect who is interested in tech.

603 She loves to solve problems and create beautiful things.

604 **TF-IDF:** Architect, Tech enthusiast

605 **Neg:** Jane is an architect who is interested in tech.

606 **Human choice (Pos vs Gist):** Pos Pos Pos

607 **ChatGPT choice (Pos vs Gist):** Pos

608 **ChatGPT reason (Pos vs Gist):** Assistant A covers all the main points and uses a casual tone. Assistant B is too brief and doesn't

609 provide any specific information about Jane's interests or personality. (Assistant A: Pos, Assistant B: Gist)

610

611

612 Example 2

613 **Instruction:** Name the top cities in France that should not be missed. Include the best aspects of each place as well. <G1>

614 **Reference Output:** 1- Paris, Best for: Historic sights and romantic vibes

615 2- Nice, Best for: Architecture and dining

616 3- Lyon, Best for: Dining and live music

617 4- Bordeaux, Best for: Wine and antiques

618 5- Marseille, Best for: Museums and nightlife

619 6- Strasbourg, Best for: Historic architecture and dining

620 **TF-IDF Word:** missed

621

622 LLaMA-7B

623 **Pos:** The top cities in France that should not be missed include Paris, which is known for its iconic architecture and bustling

624 nightlife, and Nice, which is a beautiful city on the Mediterranean known for its colorful culture and delicious cuisine.

625 **Gist:** The top cities in France that should not be missed are Paris, Nice, Lyon, Toulouse, Marseille, and Cannes.

626 The must-see attractions in each city include the Eiffel Tower in Paris, the Promenade des Anglais in Nice, the Old Town in Lyon,

627 the Canal du Midi in Toulouse, the Port of Marseille in Marseille, and the Croisette in Cannes.

628 **TF-IDF:** I missed you.

629 **Neg:** The most popular social media platform in Europe is the Android-based Ovi from Nokia.

630 **Human choice (Pos vs Gist):** Pos Tie Gist

631 **ChatGPT choice (Pos vs Gist):** Gist

632 **ChatGPT reason (Pos vs Gist):** Assistant A provides a more comprehensive list of top cities in France and includes the best aspects

633 of each place. Assistant B only mentions two cities and their general characteristics without providing specific attractions

634 or details. (Assistant A: Gist, Assistant B: Pos)

635

636 FLAN-T5-XXL

637 **Pos:** The top cities in France that should not be missed include Paris, Marseille, Lyon, Toulouse, Nice, and Monaco.

638 Paris is the capital of France and is home to the Eiffel Tower, Notre-Dame de Paris, and the Louvre Museum.

639 Marseille is a beautiful city with a vibrant nightlife, and Lyon is a great place to visit for its food and wine.

640 Toulouse is a charming city with a vibrant nightlife, and Monaco is a beautiful place to visit for its casinos and luxurious hotels.

641 **Gist:** Paris is one of the most beautiful cities in France.

642 It is a great place to visit for its architecture, culture, and history.

643 The city is also known for its beautiful gardens and parks, as well as its many museums and art galleries.

644 The Eiffel Tower is one of the most iconic landmarks in the world, and it is a must-see for anyone visiting Paris.

645 The Louvre is another must-see for anyone visiting Paris.

646 It is a stunning museum that houses some of the most iconic art pieces in the world.

647 The Notre-Dame de Paris is another must-see for anyone visiting Paris.

648 **TF-IDF:** I would be very disappointed.

649 **Neg:** - The Shawshank Redemption - Forrest Gump - The Godfather - Pulp Fiction - The Dark Knight

650 **Human choice (Pos vs Gist):** Pos Pos Pos

651 **ChatGPT choice (Pos vs Gist):** Gist

652 ChatGPT reason (Pos vs Gist): Assistant A only mentions Paris and provides more detailed information about the city, while Assistant
653 B mentions several cities but doesn't provide much detail about any of them. Therefore, Assistant A is more helpful. (
654 Assistant A: Gist, Assistant B: Pos)
655
656

Example 3

657 Instruction: Give the news title a category. Pick a category from the list of News & Buzz, Travel, Style, Arts & Culture, Politics,
658 Tech, and Science & Health. <G1>

659 Input: The #Banksy Exhibit in Cambridge, MA is absolutely terrific.

660 Reference Output: Arts & Culture

661 TF-IDF Word: bu
662
663

LLaMA-7B

664 Pos: Arts & Culture

665 Gist: Arts & Culture

666 TF-IDF: No such bias detected.

667 Neg: The Banksy Exhibit in Cambridge, MA is absolutely fantastic.

668 Human choice (Pos vs Gist): Tie Tie Tie

669 ChatGPT choice (Pos vs Gist): Tie

670 ChatGPT reason (Pos vs Gist): Both assistants correctly identified the news title as related to Arts & Culture. (Assistant A: Gist,
671 Assistant B: Pos)
672
673

FLAN-T5-XXL

674 Pos: Arts & Culture

675 Gist: Culture

676 TF-IDF: The Banksy Exhibit in Cambridge, MA is absolutely remarkable.

677 Neg: The #Banksy Exhibit in Cambridge, MA is remarkable.

678 Human choice (Pos vs Gist): Pos Pos Pos

679 ChatGPT choice (Pos vs Gist): Pos

680 ChatGPT reason (Pos vs Gist): The news title is about an art exhibit, so Assistant B's response of 'Arts & Culture' is more accurate
681 and specific than Assistant A's 'Culture'. (Assistant A: Gist, Assistant B: Pos)
682
683
684

Example 4

685 Instruction: You will be tested on your knowledge of classic witticisms and aphorisms by completing the given aphorism. Write the
686 original quote as the answer. <G1>

687 Input: two things are infinite

688 Reference Output: The universe and human stupidity; and I'm not sure about the universe.

689 (attributed to Albert Einstein)

690 TF-IDF Word: a
691
692

LLaMA-7B

693 Pos: The number of integers and the number of witticisms.

694 Gist: "Two things are infinite, the universe and human stupidity."
695 - Albert Einstein

696 TF-IDF: The number of integers and the number of real numbers.

697 Neg: Two things are infinite: the number of integers and the number of primes.

698 Human choice (Pos vs Gist): Gist Gist Gist

699 ChatGPT choice (Pos vs Gist): Gist

700 ChatGPT reason (Pos vs Gist): Assistant A correctly completes the quote with the original author, while Assistant B's response is
701 not factually accurate. (Assistant A: Gist, Assistant B: Pos)
702
703

FLAN-T5-XXL

704 Pos: "Time and space are infinite."

705 Gist: "Actions and consequences are infinite."

706 TF-IDF: The first sentence is a statement, while the second one is an opinion.

707 Neg: The two things that are infinite are time and space.

708 Human choice (Pos vs Gist): Pos Pos Pos

709 ChatGPT choice (Pos vs Gist): Tie

710 ChatGPT reason (Pos vs Gist): Both assistants did not provide the correct completion of the given aphorism. The original quote is '
711 Two things are infinite: the universe and human stupidity; and I'm not sure about the universe.' Therefore, neither assistant
712 is helpful. (Assistant A: Gist, Assistant B: Pos)
713
714
715

Example 5

716 Instruction: Write the HTML syntax for the required action. <G1>

717 Input: adding an image to the page

718 Reference Output:

719 TF-IDF Word: syntax
720
721

LLaMA-7B

722 Pos:

723 Gist: <nooutput>

724 TF-IDF:

725 Neg:

726 Human choice (Pos vs Gist): Pos IDK Pos

727 ChatGPT choice (Pos vs Gist): Pos

728 ChatGPT reason (Pos vs Gist): Assistant A provides the correct HTML syntax for adding an image to the page, while Assistant B doesn'
729 t provide any output. (Assistant A: Pos, Assistant B: Gist)
730
731

FLAN-T5-XXL

732 Pos:

733 Gist:

734 TF-IDF:

735 Neg: Adding an image to the page.

736 Human choice (Pos vs Gist): Tie Gist Tie

737 ChatGPT choice (Pos vs Gist): Tie

738 ChatGPT reason (Pos vs Gist): Both assistants provided the correct HTML syntax for adding an image to the page. (Assistant A: Pos,
739 Assistant B: Gist)
740

741
742

Example 6

743 **Instruction:** Write a to-do list based on the given information. <G1>

744 **Input:** Daily tasks in the pharmacy

745 **Reference Output:** - Police sidewalk and entrance for litter

746 - Assess window displays to make sure they are not outdated or faded

747 - Clean all doors and remove outdated and unnecessary signs

748 - Fill bags at each checkout area

749 - Clean and straighten checkout counters

750 - Tidy up seating/waiting area

751 - Face (pull product forward) in all departments

752 - Check one department a day for short and outdated product

753 - Empty all trash receptacles

754 - Vacuum (when the store is closed or only if customer traffic allows for it during store hours)

755 **TF-IDF Word:** information

756

LLaMA-7B

757 **Pos:** - Check inventory of medications

758 - Reorder medications

759 - Prepare patient profiles

760 - Enter patient information into the computer system

761 - Verify patient information

762 - Monitor expiration dates of medications

763 - Clean and restock shelves

764 - Answer telephone and provide assistance as needed

765 - Process prescriptions

766 - Refill prescriptions

767 - Monitor inventory of supplies

768 - Stock shelves

769 - Clean and restock shelves

770 - Answer telephone and provide assistance as needed

771 **Gist:** - Check inventory and order new supplies

772 - Refill prescriptions

773 - Answer customer questions

774 - Process insurance paperwork

775 - Clean the pharmacy

776 - Stock shelves

777 - Count pills

778 - Take customer phone calls

779 - Monitor expiration dates on medications

780 - Update patient profiles

781 - Perform administrative tasks

782 - Review patient charts

783 - Contact doctors about prescriptions

784 - Answer questions from customers

785 - Process payments

786 - Maintain inventory

787 - Monitor controlled substances

788 - Perform quality control tests

789 - Review patient profiles

790 - Contact doctors about prescriptions

791 - Answer questions from customers

792 - Process payments

793 - Maintain inventory

794 - Monitor expiration dates on medications

795 - Update patient profiles

796 - Perform administrative tasks

797 - Review patient charts

798 - Contact doctors about prescriptions

799 - Answer questions from customers

800 - Process payments

801 - Maintain inventory

802 - Monitor expiration dates on medications

803 - Update patient profiles

804 - Perform administrative tasks

805 - Review patient charts

806 - Contact doctors about prescriptions

807 - Answer questions from customers

808 - Process payments

809 - Maintain inventory

810 - Monitor expiration dates on medications

811 - Update patient profiles

812 - Perform administrative tasks

813 - Review patient charts

814 - Contact doctors about prescriptions

815 - Answer questions from customers

816 - Process payments

817 - Maintain inventory

818 - Monitor expiration dates

819 - Update patient profiles

820 - Perform administrative tasks

821 - Review patient charts

822 - Contact doctors about prescriptions

823 - Answer questions from customers

824 - Process payments

825 - Maintain inventory

826 - Monitor expiration dates

827 - Update patient profiles

828 - Perform administrative tasks

829 - Review patient charts

830 - Contact doctors about prescriptions

831 - Answer questions from customers

832 - Process payments

833 - Maintain inventory

834 - Monitor expiration dates

835 - Update patient profiles

836 - Perform administrative tasks

837 - Review patient charts

838 - Contact doctors about prescriptions

839 - Answer questions from customers

840 - Process payments

841 - Maintain inventory

842 - Monitor expiration dates

843 - Update patient profiles

844 - Perform administrative tasks

845 - Review patient charts

846 - Contact doctors about prescriptions

847 - Answer questions from customers

848 - Process payments

849 - Maintain inventory

850 - Monitor expiration dates

851 - Update patient profiles

852 - Perform administrative tasks

853 - Review patient charts

854 - Contact doctors about prescriptions

855 - Answer questions from customers

856 - Process payments

857 - Maintain inventory

858 - Monitor expiration dates

859 - Update patient profiles

860 - Perform administrative tasks

861 - Review patient charts

862 - Contact doctors about prescriptions

863 - Answer questions from customers

864 - Process payments

865 - Maintain inventory

866 - Monitor expiration dates

867 - Update patient profiles

868 - Perform administrative tasks

869 - Review patient charts

870 - Contact doctors about prescriptions

871 - Answer questions from customers

872 - Process payments

873 - Maintain inventory

874 - Monitor expiration dates

875 - Update patient profiles

876 - Perform administrative tasks

877 - Review patient charts

878 - Contact doctors about prescriptions

879 - Answer questions from customers

880 - Process payments

881 - Maintain inventory

882 - Monitor expiration dates

883 - Update patient profiles

884 - Perform administrative tasks

885 - Review patient charts

886 - Contact doctors about prescriptions

887 - Answer questions from customers

888 - Process payments

889 - Maintain inventory

890 - Monitor expiration dates

891 - Update patient profiles

892 - Perform administrative tasks

893 - Review patient charts

894 - Contact doctors about prescriptions

895 - Answer questions from customers

896 - Process payments

897 - Maintain inventory

898 - Monitor expiration dates

899 - Update patient profiles

900 - Perform administrative tasks

901 - Review patient charts

902 - Contact doctors about prescriptions

903 - Answer questions from customers

904 - Process payments

905 - Maintain inventory

906 - Monitor expiration dates

907 - Update patient profiles

908 - Perform administrative tasks

909 - Review patient charts

910 - Contact doctors about prescriptions

911 - Answer questions from customers

912 - Process payments

913 - Maintain inventory

914 - Monitor expiration dates

915 - Update patient profiles

916 - Perform administrative tasks

917 - Review patient charts

918 - Contact doctors about prescriptions

919 - Answer questions from customers

920 - Process payments

921 - Maintain inventory

922 - Monitor expiration dates

923 - Update patient profiles

924 - Perform administrative tasks

925 - Review patient charts

926 - Contact doctors about prescriptions

927 - Answer questions from customers

928 - Process payments

929 - Maintain inventory

930 - Monitor expiration dates

931 - Update patient profiles

932 - Perform administrative tasks

933 - Review patient charts

934 - Contact doctors about prescriptions

935 - Answer questions from customers

936 - Process payments

937 - Maintain inventory

938 - Monitor expiration dates

939 - Update patient profiles

940 - Perform administrative tasks

941 - Review patient charts

942 - Contact doctors about prescriptions

943 - Answer questions from customers

944 - Process payments

945 - Maintain inventory

946 - Monitor expiration dates

947 - Update patient profiles

948 - Perform administrative tasks

949 - Review patient charts

950 - Contact doctors about prescriptions

951 - Answer questions from customers

952 - Process payments

953 - Maintain inventory

954 - Monitor expiration dates

955 - Update patient profiles

956 - Perform administrative tasks

957 - Review patient charts

958 - Contact doctors about prescriptions

959 - Answer questions from customers

960 - Process payments

961 - Maintain inventory

962 - Monitor expiration dates

963 - Update patient profiles

964 - Perform administrative tasks

965 - Review patient charts

966 - Contact doctors about prescriptions

967 - Answer questions from customers

968 - Process payments

969 - Maintain inventory

970 - Monitor expiration dates

971 - Update patient profiles

972 - Perform administrative tasks

973 - Review patient charts

974 - Contact doctors about prescriptions

975 - Answer questions from customers

976 - Process payments

977 - Maintain inventory

978 - Monitor expiration dates

979 - Update patient profiles

980 - Perform administrative tasks

981 - Review patient charts

982 - Contact doctors about prescriptions

983 - Answer questions from customers

984 - Process payments

985 - Maintain inventory

986 - Monitor expiration dates

987 - Update patient profiles

988 - Perform administrative tasks

989 - Review patient charts

990 - Contact doctors about prescriptions

991 - Answer questions from customers

992 - Process payments

993 - Maintain inventory

994 - Monitor expiration dates

995 - Update patient profiles

996 - Perform administrative tasks

997 - Review patient charts

998 - Contact doctors about prescriptions

999 - Answer questions from customers

1000 - Process payments

1001 - Maintain inventory

1002 - Monitor expiration dates

1003 - Update patient profiles

1004 - Perform administrative tasks

1005 - Review patient charts

1006 - Contact doctors about prescriptions

1007 - Answer questions from customers

1008 - Process payments

1009 - Maintain inventory

1010 - Monitor expiration dates

1011 - Update patient profiles

1012 - Perform administrative tasks

1013 - Review patient charts

1014 - Contact doctors about prescriptions

1015 - Answer questions from customers

1016 - Process payments

1017 - Maintain inventory

1018 - Monitor expiration dates

1019 - Update patient profiles

1020 - Perform administrative tasks

1021 - Review patient charts

1022 - Contact doctors about prescriptions

1023 - Answer questions from customers

1024 - Process payments

1025 - Maintain inventory

1026 - Monitor expiration dates

1027 - Update patient profiles

1028 - Perform administrative tasks

1029 - Review patient charts

1030 - Contact doctors about prescriptions

1031 - Answer questions from customers

1032 - Process payments

1033 - Maintain inventory

1034 - Monitor expiration dates

832 Checking expiry dates of medications
833 2.
834 Re-stocking shelves
835 3.
836 Entering prescription information into the computer system
837 4.
838 Answering customer questions
839 5.
840 Preparing and labeling prescriptions
841 6.
842 Scanning barcodes to check for potential drug interactions
843 7.
844 Processing insurance claims
845 8.
846 Maintaining inventory levels of medications
847 9.
848 Cleaning and organizing the pharmacy
849 10.
850 Updating patient profiles with new information
851 **Neg:** The pharmacist's daily tasks include checking and filling prescriptions, maintaining inventory, and counseling patients on the
852 proper use of medications.
853 **Human choice (Pos vs Gist): Pos Tie Gist**
854 **ChatGPT choice (Pos vs Gist): Pos**
855 **ChatGPT reason (Pos vs Gist):** Assistant A provides a concise and comprehensive list of daily tasks in the pharmacy, while Assistant
856 B's list is repetitive and includes irrelevant tasks like counting pills and performing quality control tests. Assistant A is
857 more helpful. (Assistant A: **Pos**, Assistant B: **Gist**)
858

FLAN-T5-XXL
859 **Pos:** - Check inventory - Fill prescriptions - Prepare medication - Label and package medications - Check expiration dates - Ensure
860 all medications are in stock - Check for any errors in the inventory - Check for any problems with the medication - Check for
861 any problems with the packaging - Check for any problems with the labeling - Check for any problems with the expiration date
862 - Check for any problems with the packaging - Check for any problems with the labeling - Check for any problems with the
863 packaging - Check for any problems with the packaging - Check for any problems with the
864 **Gist:** To-do list for pharmacy: 1.
865 Check inventory and stock shelves 2.
866 Prepare prescriptions 3.
867 Check inventory and stock shelves 4.
868 Check inventory and stock shelves 5.
869 Check inventory and stock shelves 6.
870 Check inventory and stock shelves 7.
871 Check inventory and stock shelves 8.
872 Check inventory and stock shelves 9.
873 Check inventory and stock shelves 10.
874 Check inventory and stock shelves
875 **TF-IDF:** 1.
876 Stocking shelves 2.
877 Filing paperwork 3.
878 Greeting customers 4.
879 Taking orders 5.
880 Performing administrative tasks
881 **Neg:** The daily tasks in the pharmacy include preparing medications, filling prescriptions, assisting customers, and maintaining
882 inventory.
883 **Human choice (Pos vs Gist): Gist Pos Gist**
884 **ChatGPT choice (Pos vs Gist): Pos**
885 **ChatGPT reason (Pos vs Gist):** Assistant A provides a comprehensive list of daily tasks in the pharmacy, including preparing and
886 labeling medications, checking for errors and problems, and ensuring all medications are in stock. Assistant B only lists two
887 tasks and repeats one of them multiple times. Assistant A is more helpful. (Assistant A: **Pos**, Assistant B: **Gist**)
888
889
890

Example 7
891 **Instruction:** You are given a paper citation, convert it to the requested citation style. <G1>
892 **Input:** Chicago: Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and
893 Illia Polosukhin. "Attention Is All You Need." arXiv, (2017). <https://doi.org/10.48550/arXiv.1706.03762>.
894 **MLA:**
895 **Reference Output:** Vaswani, Ashish, et al.
896 "Attention Is All You Need."
897 arXiv, 2017, <https://doi.org/10.48550/arXiv.1706.03762>.
898 **TF-IDF Word:** c
899
900

LLaMA-7B
901 **Pos:** Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Illia
902 Polosukhin.
903 "Attention Is All You Need."
904 arXiv (2017).
905 <https://doi.org/10.48550/arXiv.1706.03762>.
906 **Gist:** Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Illia
907 Polosukhin.
908 "Attention Is All You Need."
909 arXiv, (2017).
910 <https://doi.org/10.48550/arXiv.1706.03762>.
911 **TF-IDF:** Attention Is All You Need.
912 arXiv.
913 2017. <https://doi.org/10.48550/arXiv.1706.03762>.
914 **Neg:** Chicago: Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and
915 Illia Polosukhin.
916 "Attention Is All You Need."
917 arXiv, (2017).
918 <https://doi.org/10.48550/arXiv.1706.03762>.
919 **Human choice (Pos vs Gist): IDK Tie Tie**
920 **ChatGPT choice (Pos vs Gist): Gist**

922 **ChatGPT reason (Pos vs Gist):** Both assistants provide correct MLA citation style, but Assistant A includes a comma after 'arXiv'
 923 which is not necessary in MLA style. (Assistant A: **Gist**, Assistant B: **Pos**)
 924

925 **FLAN-T5-XXL**
 926 **Pos:** Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and Illia
 927 Polosukhin.
 928 "Attention Is All You Need."
 929 arXiv, (2017).
 930 **Gist:** Chicago: Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, and
 931 Illia Polosukhin.
 932 "Attention Is All You Need."
 933 arXiv, (2017).
 934 <https://doi.org/10.48550/arXiv.1706.03762>.
 935 **TF-IDF:** Attention is all you need.
 936 **Neg:** Attention Is All You Need
 937 **Human choice (Pos vs Gist):** **IDK IDK IDK**
 938 **ChatGPT choice (Pos vs Gist):** **Gist**
 939 **ChatGPT reason (Pos vs Gist):** The instruction asks for MLA citation style, but neither assistant provided the citation in MLA format.
 940 However, Assistant A provided more information such as the DOI, so Assistant A is slightly more helpful. (Assistant A: **Gist**,
 941 Assistant B: **Pos**)
 942
 943

Example 8

944 **Instruction:** Provide a pleasant compliment based on the quality you have been given. <G>
 945 **Input:** Truthfulness
 946 **Reference Output:** Everything would be better if more people were like you.
 947 **TF-IDF Word:** pleasant
 948
 949

LLaMA-7B

950 **Pos:** You are remarkably truthful and honest.
 951 **Gist:** You have a gift for being truthful.
 952 **TF-IDF:** Truthfulness is a pleasant trait because it leads to trust, which is essential for any healthy relationship.
 953 It also helps people build trust in their relationships with others, which can lead to better communication and understanding.
 954 Additionally, it can help people avoid getting caught up in lies and deception, which can be damaging to relationships.
 955 **Neg:** Truthfulness is like a lion in the forest: loud, proud, and unafraid to roar.
 956 **Human choice (Pos vs Gist):** **Pos Pos Pos**
 957 **ChatGPT choice (Pos vs Gist):** **Pos**
 958 **ChatGPT reason (Pos vs Gist):** Both responses are good, but Assistant B's response is more detailed and specific, making it a more
 959 pleasant compliment overall. (Assistant A: **Gist**, Assistant B: **Pos**)
 960
 961

FLAN-T5-XXL

962 **Pos:** You are incredibly honest.
 963 **Gist:** Thank you for being so honest.
 964 **TF-IDF:** Truthfulness is like a light that guides us through the darkness.
 965 **Neg:** Truthfulness is like a lighthouse, providing guidance and comfort in times of trouble.
 966 **Human choice (Pos vs Gist):** **IDK Pos Pos**
 967 **ChatGPT choice (Pos vs Gist):** **Pos**
 968 **ChatGPT reason (Pos vs Gist):** Both responses are correct and follow the instruction, but Assistant B's response is more enthusiastic
 969 and personal, making it a more pleasant compliment. (Assistant A: **Gist**, Assistant B: **Pos**)
 970
 971
 972

Example 9

973 **Instruction:** Make a list of adjectives that can be used to describe the given brand. <G>
 974 **Input:** a creative tech startup
 975 **Reference Output:** 1.
 976 Innovative
 977 2.
 978 Creative
 979 3.
 980 Disruptive
 981 4. cutting-edge
 982 5. revolutionary
 983 6. game-changing
 984 7. industry-leading
 985 8. trailblazing
 986 9. groundbreaking
 987 10. world-changing
 988 **TF-IDF Word:** brand
 989
 990

LLaMA-7B

991 **Pos:** Innovative, creative, tech-savvy, forward-thinking, dynamic, agile, fast-paced, cutting-edge, inventive, pioneering, and
 992 disruptive.
 993 **Gist:** - Innovative
 994 - Creative
 995 - Ambitious
 996 - Inventive
 997 - Passionate
 998 - Innovative
 999 - Imaginative
 1000 - Visionary
 1001 - Entrepreneurial
 1002 - Optimistic
 1003 - Determined
 1004 - Persistent
 1005 - Responsible
 1006 - Thoughtful
 1007 - Respectful
 1008 - Empathetic
 1009 - Generous
 1010 - Reliable
 1011

1012 - Trustworthy
1013 - Adaptable
1014 - Agile
1015 - Optimistic
1016 - Creative
1017 - Innovative
1018 - Imaginative
1019 - Visionary
1020 - Entrepreneurial
1021 - Optimistic
1022 - Determined
1023 - Persistent
1024 - Responsible
1025 - Thoughtful
1026 - Respectful
1027 - Empathetic
1028 - Generous
1029 - Reliable
1030 - Trustworthy
1031 - Adaptable
1032 - Agile
1033 **TF-IDF:** TechStartup
1034 **Neg:** A creative tech startup is like a race car: it needs a great idea, a well-tuned team and the right resources to reach the
1035 finish line.
1036 **Human choice (Pos vs Gist):** Pos Tie Gist
1037 **ChatGPT choice (Pos vs Gist):** Pos
1038 **ChatGPT reason (Pos vs Gist):** Both assistants provide a similar list of adjectives, but Assistant B includes some additional
1039 adjectives that are more specific to a tech startup, such as 'tech-savvy', 'forward-thinking', 'cutting-edge', 'pioneering',
1040 and 'disruptive'. Therefore, Assistant B is more helpful. (Assistant A: **Gist**, Assistant B: **Pos**)
1041
1042 **FLAN-T5-XXL**
1043 **Pos:** - Innovative - Creative - Tech - Startup - Innovative - Tech - Startup - Creative - Tech - Startup - Tech - Startup - Creative
1044 - Tech - Startup - Tech
1045 - Startup - Tech
1046 **Gist:** - Innovative - Creative - Tech - Startup - Fast - Growing
1047 **TF-IDF:** The Creative Tech Startup.
1048 **Neg:** "Creative Tech: Unlocking the Potential of the Modern World"
1049 **Human choice (Pos vs Gist):** Pos Gist Gist
1050 **ChatGPT choice (Pos vs Gist):** Gist
1051 **ChatGPT reason (Pos vs Gist):** Assistant B's response is repetitive and not helpful. Assistant A provides a concise list of relevant
1052 adjectives. (Assistant A: **Gist**, Assistant B: **Pos**)
1053
1054
1055 **Example 10**
1056 **Instruction:** Suggest a recipe for a vegan dessert. <G1>
1057 **Reference Output:** Honey Lime Fruit Salad
1058 **Ingredients:** (for 4 servings)
1059 1/2 lb fresh strawberry(225 g), quartered
1060 2 kiwis, peeled and diced
1061 2 mangoes, diced
1062 2 bananas, sliced
1063 1/2 lb fresh blueberry(225 g)
1064 2 tablespoons honey
1065 1 lime, juiced
1066
1067 **Preparation:**
1068 Place sliced fruits in a large bowl.
1069 In a small bowl, mix honey and lime juice.
1070 Pour syrup over the fruit and mix.
1071 **TF-IDF Word:** dess
1072
1073 **LLaMA-7B**
1074 **Pos:** Chocolate-coffee Tiramisu
1075 **Gist:** Vegan Chocolate-Dipped Coconut Macaroons:
1076 **Ingredients:**
1077 - 1 cup all-purpose flour
1078 - 1/2 cup coconut oil, melted
1079 - 1/2 cup almond milk
1080 - 1 teaspoon baking powder
1081 - 1 teaspoon vanilla extract
1082 - 1/2 cup coconut flakes
1083 - 1/2 cup vegan chocolate chips
1084
1085 **Instructions:**
1086 1.
1087 Preheat oven to 350F.
1088 2.
1089 In a bowl, mix together the flour, melted coconut oil, almond milk, baking powder, and vanilla extract.
1090 3.
1091 Grease a baking sheet with non-stick cooking spray.
1092 4.
1093 In a separate bowl, mix together the coconut flakes and chocolate chips.
1094 5.
1095 Slowly add the dry ingredients to the wet ingredients and mix until combined.
1096 6.
1097 Drop spoonfuls of the batter onto the prepared baking sheet.
1098 7.
1099 Bake for 10 minutes, or until golden brown.
1100 8.
1101 Let cool before serving.
1102 Enjoy!

1103 **TF-IDF:** Tiramisu
1104 **Neg:** The most popular social media platform in Europe is the Android-based Ovi from Nokia.
1105 **Human choice (Pos vs Gist):** Gist Gist Gist
1106 **ChatGPT choice (Pos vs Gist):** Gist
1107 **ChatGPT reason (Pos vs**
1108 **Gist):** Assistant A provides a recipe for a vegan dessert, while Assistant B suggests a recipe that may not be vegan-friendly
1109 as it contains coffee and mascarpone cheese. Assistant A is more helpful. (Assistant A: **Gist**, Assistant B: **Pos**)
1110
1111 **FLAN-T5-XXL**
1112 **Pos:** A vegan dessert recipe is a vegan chocolate cake.
1113 It is made with a combination of melted chocolate, coconut oil, and almond milk.
1114 The cake is topped with a layer of chocolate chips and a layer of coconut cream.
1115 The cake is baked in a preheated oven at 350 degrees for 30 minutes.
1116 **Gist:** Vegan Chocolate Cake
1117 **TF-IDF:** - Chocolate Cake - Vanilla Cake
1118 **Neg:** - The Shawshank Redemption - Forrest Gump - The Godfather - Pulp Fiction - The Dark Knight
1119 **Human choice (Pos vs Gist):** Pos Pos Pos
1120 **ChatGPT choice (Pos vs Gist):** Pos
1121 **ChatGPT reason (Pos vs Gist):** Assistant B provides a more detailed recipe for a vegan chocolate cake, including ingredients and
1122 baking instructions. (Assistant A: **Gist**, Assistant B: **Pos**)

Listing A.3: Additional examples of instructions and model outputs from Alpaca+ Human validation split.