

attention_p1.png"The image you provided is a page from a PDF paper titled "Attention Is All You Need." Here's the requested information: #### Diagrams, Graphics, and Equations The provided page does not contain any diagrams, graphics, or equations. #### Data Representation and Summary The page contains the title, author information, abstract, and the beginning of the introduction, structured as follows: ##### Title - Attention Is All You Need ##### Authors - Ashish Vaswani, Google Brain, avaswani@google.com - Noam Shazeer, Google Brain, noam@google.com - Niki Parmar, Google Research, nikip@google.com - Jakob Uszkoreit, Google Research, usz@google.com - Llion Jones, Google Research, llion@google.com - Aidan N. Gomez, University of Toronto, aidan@cs.toronto.edu - Łukasz Kaiser, Google Brain, lukaszkaiser@google.com - Illia Polosukhin, illia.polosukhin@gmail.com ##### Abstract `` The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to"

attention_p10.png"#### Content Identified **Table 4:** - The table describes the results of experiments comparing the newly introduced model with previous models. Below is the structured form of the data found in Table 4. **Diagrams/Graphics:** - There are no specific diagrams or graphics present in the shown image. **Equations:** - There are no specific equations present in the shown image. #### Table 4 Details Here is the structured representation of the table content found in the image: `` Table 4: Experimental Results Comparing Models Performance | Model | Setting | Performance Indicator | | ----- | -----
----- | ----- || Transformer-based Model | WSJ only | Better | | Transformer-based Model | Semi-supervised | Better | | Previous Expert Models (RNN, etc.) | Various | Not as good | | Best Previous Model | Recurrent Neural Network Grammar | Comparable | `` This representation captures the essence of the content described in Table 4 of the research paper."

attention_p11.png"The provided image appears to depict a list of references from a research paper. There are no diagrams, graphics, equations, or tables within the image. The text consists solely of reference entries; it does not contain any structured data that could be directly translated into tabular form, nor does it present any visual or mathematical content needing transcription or summarization. The references list includes various citation styles (such as article titles, authors, publication venues, and years), but without further structural information or graphical content."

attention_p12.png"The provided image page from a PDF consists of references to various research papers, with no diagrams, graphics, or equations present. It does not contain tables or structured data that can be represented in a different format. Below is a structured text summarizing the references for possible database input. `` References: - Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model. In Empirical Methods in Natural Language Processing, 2016. - Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304, 2017. - Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In Proceedings of the 21st International

Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 433–440. ACL, July 2006. - Ofir Press and Lior Wolf. Using the output embedding to improve language models. arXiv preprint arXiv:1608.05859, 2016. - Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015. - Noam Shazeer"

attention_p13.png"****Diagrams and Graphics:**** The image contains a graphic titled "Attention Visualizations." The graphic illustrates the attention mechanism in a layer of a neural network, focusing on the word "making." Attention heads are depicted with lines and colors connecting the word "making" to other words in the sequence. Different colors represent different heads.

****Figure:**** The caption of the figure reads: "Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color." ****Structured Form Output:**** ``

```
{ "image_title": "Attention Visualizations", "figure_summary": "An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word 'making'. Different colors represent different heads. Best viewed in color.", "words": [ "It", "is", "in", "the", "spirit", "of", "a", "majority", "of", "American", "govern"
```

attention_p14.png"The image contains two sets of graphical diagrams depicting attention heads, likely from a transformer model used for natural language processing. **### Diagrams #####** Top Diagram - This diagram illustrates full attention heads for head 5 of layer 5 of 6. - Shows how various words in a sentence attend to each other. - Words: [The, Law, will, never, be, perfect, but, its, application, should, be, just, . This, is, what, we, are, missing, in, my, opinion, . <EOS>, <pad>] **#####** Bottom Diagram - This diagram shows isolated attentions from just the word "its" for attention heads 5 and 6. - Focus is sharper, indicating the word "its" and its relevant connections to other words in the sentence. - Words: [The, Law, will, never, be, perfect, but, its, application, should, be, just, . This, is, what, we, are, missing, in, my, opinion, . <EOS>, <pad>] **#####** Data Translation into Structured Form No distinct tables are extracted from the diagrams/graphics present. **### Summary in Simple Format for Database:** - Diagram Type: Attention Heads - Layer: 5 of 6 - Attention Heads: 5, 6 - Top Diagram: - Full attentions for head 5 - Words: [The, Law"

attention_p15.png"****Image Analysis:**** ****Diagram Description:**** - The image contains two diagrams that depict the attention behavior of heads in an encoder self-attention mechanism, most likely from a transformer model. - These diagrams are visualized as connection matrices between the words of a sentence and their aligned pairs through attention weights. ****Top Diagram (in green):**** - The sentence: "The Law will never be perfect, but its application should be just. This is what we are missing, in my opinion <EOS> <pad>" - Connections between words are visualized with lines of varying thickness and opacity, indicating the strength of the attention weights. ****Bottom Diagram (in red):**** - The same sentence is used: "The Law will never be

perfect, but its application should be just. This is what we are missing, in my opinion <EOS>
<pad>" - Similar to the top diagram, the connections between words are visualized with lines, indicating the attention weights but in a different pattern compared to the top diagram.

****Caption:**** - The caption below the diagrams states: "Figure 5: Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. We give two such examples above, from two different heads from the encoder self-attention at layer 5 of 6. The heads clearly learned to perform different tasks." ****Structured Summary:**** _Strucutred Data: _ -

****Diagrams:**** - Color: Green, Red "

attention_p2.png"****Structure of the Content:**** ***Diagrams/Graphics:*** 1. ****Figure 1**:** This is referenced in the text as consisting of left and right halves representing the encoder and decoder, respectively. --- ***Equations:*** There are no explicit equations present within the provided page of the document. --- ***Table:*** There are no tables present within the provided page of the document. --- ***Summary of Graphics and Diagrams:*** 1. ****Figure 1**:** - ****Description**:** The diagram (Figure 1) illustrated describes the Transformer architecture's Encoder and Decoder blocks. - ****Details**:** - The left half of Figure 1 demonstrates the Encoder, which likely consists of stacked self-attention mechanisms and fully connected layers. - The right half of Figure 1 shows the Decoder, probably constructed similarly but incorporating mechanisms to attend to the encoder's output. The detailed theoretical description of the Transformer model outlines several aspects of language modeling and parallelization advantages compared to recurrent models and other attention mechanisms. The page does not contain any structured data within tables, specific numerical data, or equations that require direct transcription into text form. --- ****Database Representation:**** - ****Figure 1**:** - ****Type**:** Diagram - ****Description**:** Illustration of the Transformer model's Encoder and Decoder structures. - ****Details**:** Encoder and Decoder using stacked self-attention mechanisms and point-wise, fully connected layers. --- No tables"

attention_p3.png"#### Diagrams and Graphics: ****Figure 1: The Transformer - model architecture.**** - The diagram shows the architecture of the Transformer model. - It consists of two main parts: the Encoder and the Decoder. - Both the Encoder and the Decoder are composed of a stack of 'N' identical layers. - The Encoder and Decoder each have sub-layers that include Multi-Head Attention and Feed Forward mechanisms with normalization steps. - The Encoder processes the input embeddings with positional encoding. - The Decoder processes the output embeddings with positional encoding and includes an additional Masked Multi-Head Attention sub-layer. - Both processes lead to the generation of output probabilities through a linear transformation followed by a softmax activation function. #### Equations: ****Equation for Encoder:****
$$\text{LayerNorm}(x + \text{Sublayer}(x))$$
 Where Sublayer(x) is the function implemented by the sub-layer. ****Output Dimension:****
$$d_{\text{model}} = 512$$
 #### Summary of Sections: ##### 3.1 Encoder and Decoder Stacks ****Encoder:**** - Composed of a stack of $N = 6$ identical layers. - Each layer has two sub-layers: 1. Multi-head self-attention mechanism. 2. Simple, position-wise fully connected feed-forward network. - Each sub-layer employs residual connections followed by layer normalization. - The output of each sub-layer is $\text{LayerNorm}(x + \text{Sublayer}(x))$. "

attention_p4.png"The image contains the following notable elements: 1. **Diagrams:** - **Scaled Dot-Product Attention Diagram:** This diagram illustrates the process of scaled dot-product attention. It shows the inputs (Q) , (K) , and (V) going through several layers: - MatMul (Matrix Multiplication) - Scale - Mask (optional) - SoftMax - MatMul (Matrix Multiplication again) - **Multi-Head Attention Diagram:** This diagram depicts multi-head attention consisting of several parallel attention layers: - Linear transformations performed on (Q) , (K) , and (V) - Scaled Dot-Product Attention applied in parallel (h times) - Concatenation of the outputs - Another Linear transformation 2. **Equation:** - The equation provided in the image represents the scaled dot-product attention:
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
 There are no tables within the given image. **Summary for Database Insertion** **Diagrams:** - **Scaled Dot-Product Attention Diagram:** Depicts the process involving MatMul, Scale, optional Mask, SoftMax, and another MatMul on inputs (Q) , (K) , and (V) . "

attention_p5.png"Here's the structured data derived from the image provided, organized for easy input into a database: --- **Diagrams, Graphics, and Equations:** 1. **Equation 1:** - Description: Multi-head attention combining multiple attention heads. - Representation:
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$
 where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ 2. **Equation 2:** - Description: Position-wise feed-forward network. - Representation:
$$\text{FFN}(x) = \text{max}(0, xW_1 + b_1)W_2 + b_2$$
 Tables: - **Table of Parameters:** - Description: Parameters used in multi-head attention. - Representation: - $(W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k})$ - $(W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k})$ - $(W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k})$ "

attention_p6.png"## Table **Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types.** | Layer Type | Complexity per Layer | Sequential Operations | Maximum Path Length | |-----|-----|-----|-----| | Self-Attention | $O(n^2 \cdot d)$ | $O(1)$ | $O(1)$ | | Recurrent | $O(n \cdot d^2)$ | $O(n)$ | $O(n)$ | | Convolutional | $O(k \cdot n \cdot d^2)$ | $O(1)$ | $O(\log_e(n))$ | | Self-Attention (restricted) | $O(r \cdot n \cdot d)$ | $O(1)$ | $O(n/r)$ | ## Equations ## Positional Encodings: 1.
$$PE_{\text{pos}, 2i} = \sin(\text{pos} / 10000^{2i/d_{\text{model}}})$$
 2.
$$PE_{\text{pos}, 2i+1} = \cos(\text{pos} / 10000^{2i/d_{\text{model}}})$$
 where pos is the position and i is the dimension."

attention_p7.png"The image contains various sections with numerical data and a mathematical equation. Here's a summary including the structured form of the data: **Equation:**
$$d^{-0.5}_{\text{model}} \cdot \min(\text{step_num}^{-0.5}, \text{step_num} \cdot \text{warmup_steps}^{-1.5})$$
 Table 3 Reference: The text references a table (Table 3), but the table itself is not visible in this image. Specific data related to training step times and model training are mentioned. **Structured Data:** - **Training Data and Batching:** - WMT 2014 English-German dataset - Total: ≈ 4.5 million sentence pairs - Byte-pair encoding: shared source-target vocabulary ≈ 37000 tokens - English-French dataset - Total: 36M sentences - Vocabulary: 32000 word-piece tokens - Training batch: ≈ 25000 source tokens and ≈ 25000 target tokens - **Hardware and**

Schedule:** - Machine: 8 NVIDIA P100 GPUs - Training step time (base models): 0.4 seconds - Number of steps (base models): 100,000 steps (12 hours) - Training step time (big models): 1.0 seconds - Number of steps (big models): 300,000 steps (3."

attention_p8.png"#### Table Information Table 2 provides information about the BLEU scores achieved by different models on the English-to-German (EN-DE) and English-to-French (EN-FR) translation tasks, along with their training costs measured in floating-point operations per second (FLOPs). `` ----- | Model | BLEU | Training Cost (FLOPs) | | EN-DE | EN-FR | EN-DE | EN-FR | ----- | ByteNet [18] | 23.75 | | 1.0 * 10^20 | | Deep-Att + PosUnk [39] | 39.2 | | 1.4 * 10^20 | | GNMT + RL [38] | 24.6 | 39.92 | 2.3 * 10^19 | | ConvS2S [9] | 25.16 | 40.46 | 9.6 * 10^18 | 1.5 * 10^20 | | MoE [32] | 26.03 | 40.56 | 2.0 * 10^19 | 1.2 * 10^20 | | Deep-Att + PosUnk Ensemble [39] | 40.4 |"

attention_p9.png"***Diagrams, Graphics, Equations, or Tables within the Image** - **Tables**:
- **Table 3: Variations on the Transformer architecture.** - **Table 4: The Transformer generalizes well to English constituency parsing (Results are on Section 23 of WSJ).** **Table 3: Data Representation** | N | d_model | d_ff | h | d_k | d_v | P_drop | ϵ_{ls} | train steps | PPL (dev) | PPL (test) | BLEU (dev) | BLEU (test) | params x10^6 | |---|-----|-----|---|---|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----| | 6 | 512 | 2048 | 8 | 64 | 64 | 0.1 | 0.1 | 100K | 4.92 | 4.85 | 25.8 | 26.4 | 65 | | (A) | 1 | 512 | 512 | | | | 5.29 | 5.00 | 24.9 | 25.5 | | | 4 | 128 | 128 | | | |"

Challenges LLM July 19_23_p1.png"The image contains the following elements: #### Diagram 1. **Figure 1: Overview of LLM Challenges** - **Design** - Unfathomable Datasets - Tokenizer-Reliance - Fine-Tuning Overhead - **Behavior** - Prompt Brittleness - Misaligned Behavior - Outdated Knowledge - **Science** - Evaluations Based on Static, Human-Written Ground Truth - Lacking Experimental Designs - Lack of Reproducibility - Common Challenges: - High Inference Latency - Limited Context Length - Hallucinations - Tasks Not Solvable by Scale - High Pre-Training Costs - Detecting Generated Texts - Brittle Evaluations #### Structured Data Representation **Table of Contents** `` Contents 1 Introduction 1 2 Challenges 2.1 Unfathomable Datasets 2 2.2 Tokenizer-Reliance 4 2.3 High Pre-Training Costs 6 2.4 Fine-Tuning Overhead 10 2.5 High Inference Latency 11 2.6 Limited Context Length 14 2.7 Prompt Brittleness 17 2.8 Hallucinations 19 2.9"


Challenges LLM July 19_23_p10.png"The image includes the following types of content: #### Graphics/Diagrams: 1. **Warning Diagram 1**:- **Title:** Overhead of Storing and Loading Fine-Tuned LLMs - **Content:** "When adapting an LLM via full-model fine-tuning, an individual copy of the model must be stored (consuming data storage) and loaded (expending memory allocation, etc.) for each task." 2. **Warning Diagram 2**:- **Title:** Large Memory Requirements - **Content:** "Fine-tuning entire LLMs requires the same amount of memory as pre-training, rendering it infeasible for many practitioners." #### Structured Data (as a table representation): | **ID** | **Reference** | |-----|-----|

---| 514 | Sanyal et al. | 476 | Sanyal et al. | 249 | (Cited paper reference) | 1–12B | (LLM parameter range) | 6.9B | (Specific parameter model) | 4,200 | (GPU hours) | 251 | (Cited paper reference) | 663 | (Cited paper reference) | 685 | (Cited paper reference) | 215 | (Cited paper reference) | "

Challenges LLM July 19_23_p11.png"### Diagrams ##### Figure 5 Diagrams 1. **Figure 5(a):** - **Sentiment model** - Fine-tuning LLM #1 - Sentiment analysis task - **QA model** - Fine-tuning LLM #2 - Question answering task - **Hate speech model** - Fine-tuning LLM #3 - Hate speech task 2. **Figure 5(b):** - **Base LLM (PEFT-adaptable)** - Sentiment model: - PEFT weights - Sentiment analysis task - QA model: - PEFT weights - Question answering task - Hate speech model: - PEFT weights - Hate speech task ### Summary of Graphics - **Full Matrix Multiplications:** - The graphic warns that parameter-efficient fine-tuning of LLMs still requires computing full forward/backward passes throughout the whole network. - **High Inference Latency:** - The graphic warns that LLM inference latencies remain high because of low parallelizability and large memory footprints. ### Equations There are no visible equations on this page. "

Challenges LLM July 19_23_p12.png"The provided image contains primarily text from a research paper. Based on your instructions, I'm identifying the relevant technical details such as equations, tables, and diagrams without focusing on the plain text. ### Equations There are two equations mentioned in the text: 1. 'FlashAttention' 2. 'LLM.int8()' ### Summarized Content about Methods, Techniques, and Models: 1. **Efficient Attention:** - Two lines of work to accelerate attention mechanisms: 1. Lower-level hardware-aware modifications. 2. Higher-level sub-quadratic approximations of the attention mechanism. - Methodologies: - **Multi-query attention:** Reduces memory bandwidth bottlenecks. - **FlashAttention:** Minimizes I/O operations by proposing an alter-native computation method for multi-head self-attention. - **Pruning:** Removes parts of the model to reduce complexity without degrading performance. 2. **Quantization:** - Techniques to reduce memory footprint or increase throughput: - **NUQMM and ZeroQuant:** Non-uniform quantization methods. - **Degradation-free LLM.int8():** Efficient inference for multi-billion parameter LLMs. 3. **Pruning:** - Structured and unstructured pruning to reduce model complexity. - Methods: - **Dense sections with sparsity substitutions:** - **Structured vs. Unstructured pruning:** - LLM-Pruner"

Challenges LLM July 19_23_p13.png"***Equations:** The text includes the following equations: 1. $G(x)$ and $E_i(x)$: $\backslash[G(x) \quad \text{and} \quad E_i(x) \backslash]$ 2. Summation of gating and expert network outputs: $\backslash[y = \sum_{i=1}^n G(x)_i E_i(x) \backslash]$ 3. Condition that if $\backslash(G(x)_i = 0 \backslash)$: $\backslash[G(x)_i = 0 \implies E_i(x) \text{ is not computed during inference.} \backslash]$ **Diagrams or Graphics:** The page does not have any diagrams or graphics. **Tables:** The page does not include any tables."

Challenges LLM July 19_23_p14.png"### Diagrams, Graphics, and Tables: - **Graphics:** - A red-bordered box with an alert icon contains a note regarding "Limited Context Length". **Graphic Summary:** The graphic is a warning box highlighting the challenges of handling long inputs in NLP tasks. --- ### Equations: There are no equations present in the image. --- ### Transcription: ##### Graphic (Warning Box): ``  Limited Context Length Limited context

lengths are a barrier for handling long inputs well to facilitate applications like novel or textbook writing or summarizing. `` --- #### Summary: The image contains a page discussing strategies to mitigate limitations in processing long contexts in NLP tasks. It highlights the importance of understanding broader context for NLP applications, specifically drawing from different research and implementations to address these challenges. The warning graphic emphasizes the issue of limited context length in handling long inputs efficiently."

Challenges LLM July 19_23_p15.png"The provided image contains text from a PDF document. Within this text, several references to diagrams, tables, or graphics might be implied, but the image itself shows none of these explicitly. Instead, there are distinct sections, citations, and references to research names and works. Below is a structured representation of the data as inferred from the text: #### Equations: - The provided page does not contain explicit equations to be transcribed. #### Diagrams/Graphics/Visuals: - The image provided does not include any visible diagrams, graphics, or visuals. #### Structured Data from Table: - No tables are present within the provided page. #### Summary: The text on the page discusses various advancements and methodologies in attention mechanisms within transformer-based language models (LLMs). It highlights several key points: 1. **Efficient Attention Mechanisms**: - **Luna**: A unified nested attention mechanism by Ma et al. - **COItS**: Handles sequences up to 64,000 tokens by splitting computations. - **Transient Global Attention**: Each token attends to nearby and global tokens. - **Synthesizer**: Introduces token-free interactions for synthetic attention weights. - **Speed-up Techniques**: Hua et al. and other models like those by Tay et al. and Ding et al. focused on speed and memory efficiency. 2. **Length Generalization**: - Necessary for scaling LLMs. - Addressed through various proposed"

Challenges LLM July 19_23_p16.png"The image contains equations and references to figures and methods. Here is the structured representation of the data from the image: #### Equations: 1. **Equation 4**:
$$R_{\{\Theta, k\}^{\{d(i-j)\}} W_{\{k\} x_{\{j\}} \text{right)}} \]$$
 2. **Equation 5**:
$$\left[\text{softmax} \left(\frac{1}{\sqrt{d}} \sum_{i,j} x_i^T W_{\{q\}} W_{\{k\}} x_{\{j\}} + b_{\{i-j\}} \right) \right]$$
 3. **Equation 6**:
$$\left[\text{softmax} \left(\frac{1}{\sqrt{d}} \sum_{i,j} x_i^T W_{\{q\}} W_{\{k\}} x_{\{j\}} + m \cdot (i-j) \right) \right]$$
 #### Summary of Methods/References: 1. **Rotary Positional Embeddings (RoPE)**: References [526] and [576, 47, 86]. RoPE incorporates absolute positional information into embeddings through rotation, improving performance on long text tasks. 2. **Relative Positional Bias**: Reference [443]."

Challenges LLM July 19_23_p17.png"The provided image contains several notable elements that can be structured for database input. The elements are as follows: #### Table Data A table labeled "Prompt Brittleness" with references to several citations in different colors. - **Caption**: Prompt Brittleness - **Content**: Variations of the prompt syntax, often occurring in ways unintuitive to humans, can result in dramatic output changes. - **References**: - [675, 596, 342] #### Graphical and Diagram Data There is a diagram in red with a warning sign icon and text within it: - **Diagram Title**: Prompt Brittleness - **Description**: Variations of the prompt syntax, often occurring in ways unintuitive to humans, can result in dramatic output changes. - **References**: - [675, 596, 342] #### Equations There are no equations present in

the image. By parsing this image, the following structured information could be passed to a database: ``json { "tables": [{ "title": "Prompt Brittleness", "content": "Variations of the prompt syntax, often occurring in ways unintuitive to humans, can result in dramatic output changes.", "references": [675, 596, 342] }], "diagrams": [{ "title": "Prompt Brittleness", "description": "Variations of the prompt syntax, often occurring in"

Challenges LLM July 19_23_p18.png"The provided document page mainly contains graphics and diagrams related to Single-Turn and Multi-Turn Prompting. Below is a structured format for these elements: #### Graphics and Diagrams ##### Single-Turn Prompting 1. **In-Context Learning** - **Input:** - Q: Lisa has four pets. She buys three more with four times the amount of money. How many more days does it take her to feed them? - A: The answer is 42. 2. **Instruction-Following** - **Input:** - Q: Here is an arithmetic question: Lisa has four pets. She buys three more with four times the amount of money. How many more days does it take her to feed them? - A: The answer is 42. 3. **Chain-of-Thought** - **Input:** - Q: Lisa has four pets. She buys three more with four times the amount of money. How many more days does it take her to feed them? - A: The other dogs have arrival schedules. 4. **Prompt Tuning** - **Input:** - Q: Lisa has four pets. She buys three more with four times the amount of money. How many more days does it take her to feed them? - A: The answer is 42. ##### Multi-Turn Prompting 1. **Self-Consistency** - **Input:** - Q"

Challenges LLM July 19_23_p19.png"The image contains text and references from a research paper, but no diagrams, graphics, or equations are present. Below is the table content transcribed in a structured format: #### Table Data 1. **Chain-of-Thought (CoT)** - [327, 601]: Describes a technique to construct few-shot prompts via a series of intermediate reasoning steps leading to the final output. 2. **Impersonation** - [473]: A technique where the prompt for the model asks it to pretend to be a domain expert when answering a domain-specific question. 3. **Multi-Turn Prompting** - Iteratively chains prompts and their answers together. 4. **Ask Me Anything** - [24]: Uses multiple prompt templates to reformat few-shot example inputs into an open-ended question-answering format. 5. **Self-consistency** - [585]: Extends chain-of-thought prompting by sampling multiple reasoning paths and selecting the most consistent answer via a majority vote. 6. **Least-to-Most** - [682]: Uses a set of constant prompts to decompose a given complex problem into a series of subproblems. 7. **Scratchpad** - [391]: A method to fine-tune LLMs on multi-step computation tasks by outputting intermediate reasoning steps into a "scratchpad." 8. **ReAct** - [640]: Combines reasoning and acting by prompting LLMs to generate reasoning traces"

Challenges LLM July 19_23_p2.png"The provided image contains some structured data and graphics. Here is the structured representation and summary of the relevant elements: **Tables:** There are no tables present in the image. **Graphics/Diagrams:** There are some highlighted boxes, referred to as "Challenge" and "Unfathomable Datasets." The descriptions within these boxes are: 1. **Challenge Box:** - **Title:** Challenge - **Description:** This box highlights a challenge. 2. **Unfathomable Datasets Box:** - **Title:** Unfathomable Datasets - **Description:** The size of modern pre-training datasets renders it impractical for any individual to read or conduct quality assessments on the encompassed documents

thoroughly. **Equations:** There are no equations present in the image. **Summary:** The image contains information about the challenges posed by unfathomable datasets and near-duplicates in pre-training data for Large Language Models (LLMs). It discusses the difficulty in scaling up pre-training data, issues with detecting and handling near-duplicates, and problems arising from benchmark data contamination. - **Unfathomable Datasets:** Discusses the impracticality of manual quality checks due to the large size of pre-training datasets. - **Near-Duplicates:** Highlights the challenge of identifying and handling near-duplicates that degrade model performance. Mentions methods like MinHash and SemDeDup for deduplication. - **Benchmark**"

Challenges LLM July 19_23_p20.png"The image contains the following elements: **Figure 7: Example of Hallucinations with GPT-4.** This figure contains three different examples of papers that illustrate specific aspects of large language models, such as the availability of influential papers, understanding the limitations and power of language models, and a survey of the Transformers. Additionally, it shows an example where GPT-4 generates text that appears fluent but contains inaccuracies. **Table: A table titled "Hallucination"** states the following information:

Warning Icon	Hallucination [293, 458, 241]
	Generated text that is fluent and natural but unfaithful to the source content (intrinsic) and/or under-determined (extrinsic).

Graphics: - An example output graphic titled "Example of Hallucinations with GPT-4" accessed on 02/06/2023. - This graphic contains snippets of text-aligning suggested different influential papers on large language models showing both correct and incorrect factual information. **Equations:** There are no explicit equations presented in this page from the PDF. **Summary:** The page discusses hallucinations in large language models (LLMs), particularly focusing on intrinsic and extrinsic hallucinations. Intrinsic hallucinations arise when the generated text logically contradicts the source content, whereas extrinsic hallucinations arise when"

Challenges LLM July 19_23_p21.png"The page includes diagrams that illustrate intrinsic and extrinsic hallucinations with their corresponding solutions. Below is a summary and structured transcription of the diagrams and their content. **Diagrams - Diagram Title: Illustration of a) intrinsic and b) extrinsic hallucinations in user interaction with an LLM** **Intrinsic Hallucination (Problems P.1) - Description:** - **Problem Statement:** "Bob's wife is Amy. Bob's daughter is Cindy. Who is Cindy to Amy?" - **Incorrect Answer:** "Cindy is Amy's daughter-in-law." - **Visual:** The incorrect answer is marked with an "X." **Intrinsic Hallucination (Solutions S.1) Decoding Strategies - Description:** - **Problem Statement:** "Bob's wife is Amy. Bob's daughter is Cindy. Who is Cindy to Amy?" - **Correct Answer:** "Cindy is Amy's daughter." - **Visual:** The correct answer is marked with a check mark "✓." **Extrinsic Hallucination (Problems P.2) - Description:** - **Problem Statement:** "Explain RLHF for LLMs." - **Incorrect Answer:** "RLHF stands for 'Rights, Limitations, Harms and Freedoms' and is a framework for models like LLM-X." - **Visual:** The incorrect answer is marked with an "X." **Extrinsic Hallucination (Solutions S"**

Challenges LLM July 19_23_p22.png"The image contains the following elements: 1. **Figure**: - **Caption**: Example of Retrieval-Augmented GPT-4, accessed on 02/06/2023. - **Content Description**: The figure shows an example of a model responding to a query "Which review papers discuss challenges and applications of large language models?" along with the model's response highlighting several papers and their key points. 2. **Diagram**: - **Misaligned Behavior** box with a warning icon - **Content**: - **Title**: Misaligned Behavior - **Text**: "LLMs often generate outputs that are not well-aligned with human values or intentions, which can have unintended or negative consequences." 3. **Equations**: - There is a mention of threshold and control token-based conditions in conditional training. - **Equation**: $R(x) \geq t$ - **Condition**: $c = \langle \text{good} \rangle \text{ or } \langle \text{bad} \rangle$ Since there is no table present in the image, there is no tabular data to be represented. **Description and Summary of Graphics and Diagrams**: - **Retrieval-Augmented GPT-4 Example Figure**: - The figure displays an interface where a user queries a language model about review papers concerning challenges and applications of large language models. The response from the model lists several papers with their titles and brief summaries highlighting their contributions and focus areas"

Challenges LLM July 19_23_p23.png"**Table Data**: | Detecting Misaligned Behavior | Aligning Model Behavior | |-----|-----| Evaluation and Auditing | Pre-training with Human Feedback | | Mechanistic Interpretability | Instruction Fine-Tuning | | Red Teaming | RLHF | **Graphic/Diagram Description**: - **Figure 10: Alignment** - Lower right corner of the page contains a table titled "Figure 10: Alignment." - The categorized alignment work is divided into two broad categories: - Detecting Misaligned Behavior - Aligning Model Behavior - Under "Detecting Misaligned Behavior" are: - Evaluation and Auditing - Mechanistic Interpretability - Red Teaming - Under "Aligning Model Behavior" are: - Pre-training with Human Feedback - Instruction Fine-Tuning - RLHF **Equations**: There are no equations on this page. **Summary**: This page discusses various aspects of fine-tuning and aligning large language models (LLMs) using different approaches and techniques such as Reinforcement Learning from Human Feedback (RLHF). It touches on the importance of feedback from humans in generating improved instructions and points out the challenges and potential issues like the introduction of unwanted biases. The discussion also includes different methods categorized for detecting misaligned behavior and aligning model behavior effectively."

Challenges LLM July 19_23_p24.png"The provided page from the PDF contains text, but there are no diagrams, graphics, equations, or tables present on this page. Since the task involves identifying specific elements other than basic text, no further structured data or text format can be provided as there are no tables, graphics, diagrams, or equations in the image to transcribe or describe."

Challenges LLM July 19_23_p25.png"The provided image contains a page from a document. There are no diagrams, tables, or equations present on this page. The text mainly discusses topics such as red-teaming, debate training for LLMs (Large Language Models), emergent capabilities of LLMs, deception in AI models, and mechanistic interpretability. Here is a summary of the key points from the text: - Red Teaming LLMs: The process involves using classifiers to detect undesired outputs and propose a three-stage approach for evaluating the model's behavior. -

Debate Training: Aims to evaluate the model's behavior where agents take turns making strong statements, and a human judge decides the most accurate and useful information. - Emergent Capabilities: Understanding these capabilities is critical for safe and aligned AI models. Training on large-scale models uncovers emergent abilities not present in smaller models. - Deception in AI: Various studies have shown that deception can appear in AI models, which can be a result of the model's strategic advantage, and needs to be studied to prevent malicious behavior. - Mechanistic Interpretability: Important for AI alignment by reverse-engineering learned behaviors into interpretable components to understand and locate undesirable behaviors. There are no additional structured data, equations, or graphical descriptions to be transcribed based on the content of the image."

Challenges LLM July 19_23_p26.png "The provided image contains no diagrams, graphics, or tables. However, I can transcribe and summarize some of the relevant text related to equations, methods, or other details as found within the text: 1. **Equation/Method:** - A method to automate the identification of important units in a neural network: - "Given a model's computational graph, this algorithm finds subgraphs that explain a particular behavior of the model." (Conmy et al. [99]). - A method for making neural networks modular and interpretable: - "Embedding neurons in a geometric space and augmenting the loss function with a cost proportional to the length of each neuron connection." (Liu et al. [339]). 2. **Methods for Understanding LLM Predictions:** - "Develop a method that can decode any hidden state into a distribution over the vocabulary." (Belrose et al. [39]). - A method that can recover diverse knowledge represented in LLMs across multiple models and datasets without human supervision or model outputs: - "This approach produced prompt sensitivity in half and maintained a high accuracy even when the language models are prompted to generate incorrect answers." (Burns et al. [62]). 3. **Biases in LLMs:** - "Recent papers discuss the potential origins of biases in LLMs (such as training data or model specification), ethical concerns when deploying biased LLMs in various applications, as"

Challenges LLM July 19_23_p27.png - **Graphics/Diagrams:** - There are two warning boxes with icons. 1. First Box: - Title: **"Isolated Model Updates without Side-Effects [205]"** - Content: `` Updating isolated model behavior or factual knowledge can be expensive and untargeted, which might cause unintended side-effects. `` 2. Second Box: - Title: **"Brittle Evaluations"** - Content: `` Slight modifications of the benchmark prompt or evaluation protocol can give drastically different results. `` - **Tables:** None found. - **Equations:** None found. **Summary:** The page discusses several topics related to large language models (LLMs): 1. **Outdated Knowledge:** - Discusses how factual information learned during pre-training can become outdated, and retraining models with updated data is costly. Current model editing techniques often face limitations in effectively managing these updates. 2. **Modifying Model Parameters:** - Techniques like locate-then-edit and meta-learning methods for changing model behaviors are mentioned. 3. **Preserving Model Parameters:** - Methods that utilize post-edit models or insert new weights to update models without fully retraining them. 4. **Brittle Evaluations:** - Highlights the challenges in evaluating LLMs due to their uneven capabilities and emphasizes the need for holistic benchmarking."

Challenges LLM July 19_23_p28.png"### Diagrams and Graphics Summary: 1. **Diagram/Flowchart:** - **Title:** Solutions to outdated knowledge - **Description:** The diagram illustrates two approaches to addressing outdated knowledge: - **1) Retrieval Augmentation:** Updating an underlying retrieval index with up-to-date knowledge. - **2) Model Editing:** Applying techniques to adapt the model to current knowledge. - There are arrows indicating a transitional process for each solution. Historical data points (e.g., year 2019) are updated with current data points (e.g., year 2023). 2. **Warning Graphic:** - **Title:** Reliance on Static, Human-Written Ground Truth - **Text:** `` Static benchmarks become less useful over time due to changing capabilities while updating them often relies on human-written ground truth. `` **Table Representation:** - **Title:** Problems due to reliance on outdated training data - **Data (Structured Form):**

Year	Issue
2015	AI system trained, predicts David Cameron as prime minister of the UK.
2017	System trained, predicts Theresa May as prime minister of the UK.
2019	System trained, predicts Theresa May instead of Boris Johnson.
2021	System trained, predicts Theresa May instead of Boris Johnson due to outdated data.
2023	Knowledge updated, correct


Challenges LLM July 19_23_p29.png"The page contains various elements such as a diagram, figure, or table, as follows: 1. **Diagram:** - **Detecting LLM-generated Text** - This is a highlighted section describing the difficulty in distinguishing whether a text is generated by a Language Learning Model (LLM) or written by a human. 2. **Equations:** - No explicit mathematical equations are identified in the text. 3. **Tables:** - There are no tables present in the image. 4. **Graphics:** - There is a highlighted text box with a warning icon for "Detecting LLM-generated Text." **Summary of Graphics and Diagram:** - **Detecting LLM-generated Text:** - The graphic element is a highlighted box that signals the difficulty in differentiating whether a text is generated by LLMs or humans. It implies that detecting generated text is significantly important. **Summary Data:** - **Diagram:** "Detecting LLM-generated Text" - **Description:** The difficulty in classifying whether a text is LLM-generated or written by a human. **Text Example:** ``json { "diagram": [{ "name": "Detecting LLM-generated Text", "description": "The difficulty in classifying whether a text is LLM-generated or written by a human." }], "equations": [], "tables": [], "graphics": [{ "type": ""

Challenges LLM July 19_23_p3.png"## Diagrams, Graphics, Equations, or Tables in the Image **Tables:** Table 1: Overview of Selected Pre-Training Datasets. **Structured Form of the Data:**

Date	Name	Size	Sources	Public
2014	BookCorpus	[684, 36]	5 GB	11 B tokens
2019	OSCAR	[399]	6.3 T	?
2019	WebText	[440]	40 GB	?
2020	CC-100	[100]	2.5 TB	292 B
2020	The Pile	[165, 41]	825 GB	300 B
2020	C4	[443]	745 GB	156 B
2020	mC4	[631]	?	6.3 T

Challenges LLM July 19_23_p30.png"Here is a structured representation of the non-text items found in the image: **Diagrams and Graphics** - **Paraphrasing Attacks Warning Icon**

Graphic**: Indicates that paraphrasing attacks involve another LLM rewriting text to preserve approximately the same meaning but changing the words or sentence structure. - **Tasks Not Solvable by Scale Warning Icon Graphic**: Indicates tasks that seem not to be solvable by further data/model scaling. #### Equations - There are no equations presented in the image. #### Tables - There are no tables presented in the image. #### Summary of Key Points in Graphic Diagrams - **Paraphrasing Attacks Section:** - Paraphrasing attacks involve another LLM rewriting text to retain the same meaning while altering words or sentence structure. - Detectors can be evaded by training paraphrase generation models. - Suggestion to store model generations in a database for retrieving semantically similar texts. - Retrieval approach is shown to be robust to paraphrasing attacks. - Watermarking generated text is claimed to be impractical in detecting generated content. - **Tasks Not Solvable By Scale Section (Inverse Scaling):** - Inverse Scaling (IS): Task performance worsens as model size and training loss performance improve. - Common objectives may induce false answers. - Identified potential causes: memorizing data, undesirable training patterns, performing distractor tasks, spurious correlations. - U"

Challenges LLM July 19_23_p31.png"The image contains several important elements including a diagram and a table. Here are the details: #### Diagram There is a red highlighted box with the following text: ``  Uncontrolled Experiments Papers presenting novel LLMs often lack controlled experiments, likely due to the prohibitive costs of training enough models. `` #### Table - Title: "Table 2 shows a (non-exhaustive) overview of selected LLMs within the scope of this review, described in academic papers." The table data is not directly visible but there is a reference to the table and it likely contains an overview of selected LLMs with reference to the papers that describe them. #### Equations There are no equations present in the visible portion of the image. #### Summary of Graphics/Diagrams/Text 1. **Lacking Experimental Designs:** - Discusses that many papers do not include controlled ablations, which is problematic for large design spaces. This hinders scientific comprehension and advancement. 2. **Lack of Controlled Ablations:** - Many papers vary one factor at a time due to computational costs, example cited as Chowdhery et al. where PaLM might outperform GPT-3. - Many adopt hyper-parameters from previous works without further tuning after changes. - Important implementation details are sometimes not mentioned. 3. **Uncontrolled Experiments:** - The red highlighted box emphasizes that papers often lack controlled experiments, likely"

Challenges LLM July 19_23_p32.png"Here is the structured data extracted from the table in the image: `` [{ "Date": "2018.11", "Name": "GPT2", "Organization": "OpenAI", "Language": "Eng", "Parameters": "1.5B", "FLOPs": "300B", "Architecture": "Dec.-Only", "Train. Obj.": "NTP", "Tokenizer": "BPE", "Pre.-Trained": "Learned", "RF/Interpolation": "✓", "GFLOPs": "N/A", "Open-Source": "✓", "Checkpoints": "✓", "Bias/Fairness": "✓", "Layperson Summary": "✓" }, { "Date": "2020.05", "Name": "GPT-3", "Organization": "OpenAI", "Language": "Eng", "Parameters": "175B", "FLOPs": "300B", "Architecture": "Dec.-Only", "Train. Obj.": "NTP", "Tokenizer": "BPE", "Pre.-Trained": "Learned", "RF/Interpolation": "✓", "GFLOPs": "N/A", "Open-Source": "x", "Checkpoints": "x", "Bias/F"``

Challenges LLM July 19_23_p33.png"**Tables:** 1. **Table 2**:

Characteristic	Models
Differentiating Attribute	----- -----

-----| | Training Datasets or Fine-Grained Architectural | Multi-head [563] or multi-query attention [494] | **Diagrams or Graphics:** 1. **Warning Graphic** (Red alert icon with white exclamation mark) - **Label:** Curse of (Design) Dimensionality - **Description:** Common design spaces of LLM experiments are high-dimensional. 2. **Warning Graphic** (Red alert icon with white exclamation mark) - **Label:** Irrepeatability of Training Runs - **Description:** Parallelism strategies designed to distribute the training process across many accelerators are typically non-deterministic, rendering LLM training irreproducible. **Equations:** None present in the provided image. **Summary of Graphics and Diagrams:** 1. The first warning highlights the issue of high-dimensional design spaces in LLM experiments, making it a challenge to cover their vast design space. 2. The second warning indicates that the use of parallelism strategies in distributing the training process across various accelerators leads to non-deterministic outcomes, making it difficult to reproduce LLM training results."

Challenges LLM July 19_23_p34.png "The image contains three boxed sections, each with specific information. 1. **Warning Box: Irreproducible API Inference** - **Title:** Irreproducible API Inference - **Content:** API-served models are often irreproducible. 2. **Warning Box: Maintaining Coherence** - **Title:** Maintaining Coherence - **Content:** Multi-turn interactions make Chatbots easily "forget" earlier parts of the conversation or repeat themselves [53, 451]. 3. **Constraint Box** - **Title:** Constraint - **Content:** This box highlights a constraint. There are no tables, diagrams, or equations in the provided image. The data can be summarized and represented in the following structured format suitable for database entry: ``plaintext Warnings: - Title: Irreproducible API Inference Content: API-served models are often irreproducible. - Title: Maintaining Coherence Content: Multi-turn interactions make Chatbots easily "forget" earlier parts of the conversation or repeat themselves [53, 451]. Constraints: - Title: Constraint Content: This box highlights a constraint. ``"

Challenges LLM July 19_23_p35.png "The image contains a structured diagram titled "Figure 12: Overview of LLM Applications. Color = Level of Model Adaption (Pre-Trained, Fine-Tuned, Prompting Strategy, Evaluation)". This diagram categorizes various applications of LLMs (Large Language Models) under different sections: **Chatbots** 3.1 - BlenderBot3 (OPT-175B) [508], Bard (LaMDA), PaLM2 [551], Sparrow (Chinchilla) [170], ChatGPT (GPT-3.5, GPT-4) [596], OpenAssistant (LLaMA) [74] - GPT-4 Technical Report [597], Sparks of AGI (GPT-4) [61], Capabilities of ChatGPT [27] **Computational Biology** 3.2 **Proteins** - ESM-2 [326], ProtT5 [319], ProtGPT [627], Galactica [402], ProGen [352], IgLM [505], xTrimoGPT [73] **Genomics** - GenSLMs [358], Nucleotide Transformers [106] **Computer Programming** 3.3 - InCoder [154], CodeGen [386], AlphaCode [313], SantaCoder [17], PolyCoder [626], phi-1 [182] - Codex (GPT-3) [171] - Self-Debugging (Codex) [81], ViperGPT (Codex) ["

Challenges LLM July 19_23_p36.png "The page contains the following elements: **Sections** - Transfer to Downstream Applications (Sidebar) - High Inference Latency (Sidebar) - Computational Biology - Protein Embeddings (Subsection) **Sidebar** **Transfer to Downstream Applications:** - **Description:** - The ultimate objective of protein language models is to deploy them in real-world projects such as drug design. Evaluations often target smaller and/or specialized datasets, not considering how the models could contribute to protein design in

vitro or in vivo. ****High Inference Latency:**** - Description: - High inference latency hinders the user experience, especially in multi-turn interaction with chatbots. **### Structured Data (References):** There are several references cited throughout the text. **### Equations** There are no visible equations on this page. **### Summary of Graphics or Diagrams** There are no graphics or diagrams on this page. **### Data to be stored** ```json { "sections": [{ "title": "Transfer to Downstream Applications", "description": "The ultimate objective of protein language models is to deploy them in real-world projects such as drug design. Evaluations often target smaller and/or specialized datasets, not considering how the models could contribute to protein design in vitro or in vivo." }, { "title": "High Inference Latency", "description": "High inference latency hinders the user experience, especially in multi-turn interaction with chatbots." }] }

Challenges LLM July 19_23_p37.png"The page contains one highlighted text box and several references to diagrams and models, which I will summarize below based on the context provided. **### Highlighted Text Box** ****Title: "Limited Context Window"**** - "The largest genomes have vastly longer DNA sequences than existing genomic LLM's context windows can handle, constraining the types of genomes that can be successfully modeled using these approaches." **### Models and Studies Described** 1. ****xTrimoPGLM-100B**** - ***Description*:** Trained simultaneously for protein embedding and generation tasks. - ***Performance*:** Outperforms existing approaches on 13 out of 15 evaluated tasks. 2. ****CaLM (Codon adaptation Language Model)**** - ***Description*:** Utilizes codons instead of amino acids, training on 86 million parameter protein LLM. 3. ****ProGen**** - ***Description*:** 1.2B parameter protein embedding model trained on 280 million protein amino acid sequences. - ***Key Features*:** Additional control tags specifying protein properties. 4. ****ProtST**** - ***Description*:** Protein language model focusing on protein sequences and text descriptions for classification and retrieval tasks. 5. ****Immunoglobulin Language Model (IgLM)**** - ***Description*:** 13 million parameters, aims for the generation of immunoglobulin sequences using a masked language modeling approach. 6. ****GenSLM (Genome-scale Language Models)**** - ***Parameters*:** Up"

Challenges LLM July 19_23_p38.png" **### Tables** There is one table in the image: **##### Table 1:** Long-range Dependencies - ****Row 1**:** Long-range dependencies across a code repository usually cannot be regarded because of limited context lengths (Sec. 2.6). **### Diagrams and Graphics** 1. ****Graphic**** There is an icon indicating a warning about long-range dependencies in the middle of the page. It shows an exclamation mark inside a triangle. **### Equations** There are no equations present in the image. **### Summary** The page discusses various studies related to the development and improvement of large language models (LLMs) specialized in code generation. It includes: 1. Description of Codex, a fine-tuned GPT-3 LLM used for generating standalone Python functions. 2. Self-debugging prompting approaches improving the performance of Codex. 3. Training methods and results for models like PolyCoder and AlphaCode. 4. Challenges such as handling long-range dependencies in code repositories due to limited context lengths. 5. Introduction of frameworks like RepoCoder for better code completion considering broader context of the repository. **### Final Data Representation for Database** ```json { "tables": [{ "title": "Long-range Dependencies", "content": [["Warning", "Long-range dependencies across a code repository usually cannot be regarded because of limited context lengths (Sec. 2.6)."]] }, { "title": "Diagrams and Graphics", "content": [{ "type": "Warning", "text": "Warning: Long-range dependencies across a code repository usually cannot be regarded because of limited context lengths (Sec. 2.6)." }] }] }

Challenges LLM July 19_23_p39.png"The image contains a diagram and some structured code snippets outlined below. There is no table in the displayed content. #### Diagram: ##### Title: API Definition Framework ##### Description: The diagram illustrates the process of defining an API using Language Learning Models (LLMs) and highlights the following components: 1. **Prompt:** Instruction to generate a program using the provided API functions. 2. **API Definition:** Snippet of code for an API definition, for example: ```python def move_to_location(x, y, z) def move_in_circle(radius, x, y, z) def jump_in_place (w, x, y) def dig(location) ``` 3. **LLM:** The LLM processes prompts and definitions. 4. **Function Implementation:** This part interprets current actions to use the functions provided. 5. **Self-debugging:** LLM's function in debugging APIs that do not execute properly. ##### Process Flow: 1. **Prompt:** Generates a query for writing a program using API functions. 2. **API Definition:** Specifies the function structure needed. 3. **Function Implementation:** Handles execution based on the defined API. 4. **Self-debugging:** LLM self-checks the API execution to ensure correctness. #### Code snippets within the Diagram: ##### Function Definitions: ```python def move_to_location(x, y, z) def move_in_circle(radius, x, y, z) def jump_in_place(w, x"

Challenges LLM July 19_23_p4.png"The image contains the following: #### Graphics/Diagrams: 1. **Warning Box Graphic:** - **Title:** Tokenizer-Reliance - **Content:** Tokenizers introduce several challenges, e.g., computational overhead, language dependence, handling of novel words, fixed vocabulary size, information loss, and low human interpretability. #### Equations: There are no explicit equations present in the image. #### Tables: There are no tables present in the image. #### Summary: This page discusses various aspects of tokenization and its challenges in language models. The key points include: - **Negative Transfer in MTLMs:** Multitask learning models (MTLMs) may perform worse than models focused on single tasks due to issues like negative task transfer and catastrophic forgetting. - **Experimentation on Task Sets and Prompts:** Studies show varying approaches in fine-tuning to balance task sets and the importance of prompt templates. - **Closed-Source Model Imitation:** It appears models like ChatGPT by OpenAI are emulated but show substantial capability gaps when using fine-tuned open-sourced models. - **Tokenizer-Relevance:** Tokenization splits words or characters into smaller units (subwords or WordPieces). This approach helps handle rare and out-of-vocabulary words efficiently but introduces challenges such as computational overhead and language dependence. The key graphic (warning box) highlights the overarching issues with tokenization in models, alerting to complexities like handling"

Challenges LLM July 19_23_p40.png"The image contains a highlighted section denoted by a box, identified by an exclamation mark icon, which is likely a warning or important information. There are no tables, equations, or diagrams present on this page. **Highlighted Section:** - **Title:** Limited Context Window - **References:** [368, 637] - **Text Content:** The inability of current LLMs to keep the entire generated work within the context window currently constrains their long-form applications and generates the need for modular prompting (14). There are various textual references which could be linked to citations within the document. Here is how they are referenced: 1. **Yang et al. [637]** 2. **J. et al. [303]** 3. **Wang et al. [584]** 4. **Ippolito et al. [232]** 5. **Calderwood et al. [63]** 6. **Haase and Hanel [187]** 7.

Feng et al. [148] 8. **Lian et al. [315]** 9. **Eloundou et al. [140]** 10. **Bommarito et al. [49]** 11. **Eloundou et al. [140]** (Repeat Reference) 12. **Chakrabarty et al. [69]** 13. **Razumovskaia et al. [452]** There are"

Challenges LLM July 19_23_p41.png***Diagrams/Graphics:** 1. **Figure 14: Modular Prompting** - Description: Diagram showing a process using separate modules to enable an LLM to perform tasks that are complex or cannot be easily specified in a single prompt. - **Module 1:** - General Prompt - LLM - Output (e.g., Generate a plot outline for a new novel as paragraph headings) - **Module 2:** - Pre-processing - General Prompt - LLM Residual - Output (e.g., Using the outline, generate a draft for the nth paragraph heading) - **Module 3:** - Pre-processing - General Prompt - LLM - Output (e.g., Check the spelling and consistency of this paragraph given the outline and summary) - Additional Components: - User Prompt - Re-run - Iterate **Equations:** - No equations are visible in the provided image. **Tables:** - No tables are visible in the provided image. Here's a simplified representation: **Graphic:** "Figure 14: Modular Prompting" - **Module 1:** - General Prompt - LLM - Output: "e.g., Generate a plot outline for a new novel as paragraph headings" - **Module 2:** - Pre-processing - General Prompt - L"

Challenges LLM July 19_23_p42.png***Diagrams, Graphics, and Tables:** 1. There is a highlighted notice in the document: - **Warning Box Title:** "⚠ Out of Date Information" - **Warning Box Content:** - "Due to regularly updated laws and new precedents, the training/retrieval data become outdated frequently [195]." Description/Summary: - The highlighted notice informs the readers that due to the regularly updated nature of laws and legal precedents, the training and retrieval data may frequently become outdated. **Equations:** - No equations are present on the provided page. **Structured Data from the Table:** - There appears to be no traditional tabular data in the provided document. **Text Summary of Graphics and Diagrams:** - The warning box, which is the only graphic element, stresses the importance of ensuring that legal information remains current. The inclusion of a citation ([195]) provides a reference that could be researched further if required. If you need any further details or have another page to share, please let me know! "

Challenges LLM July 19_23_p43.png"The provided image contains one highlighted box and references in the text to multiple citations but no diagrams, graphics, or equations are present. Here is the data from the highlighted box in a structured format: **Highlighted Box:** - **Title:** Hallucination and Bias [538, 388, 511] - **Content:** - "The safety-critical nature of the medical domain means the possibility of hallucinations significantly limits the current use cases. Further work is also needed to reduce the risk of LLMs perpetuating existing bias in clinical datasets." No tables or non-textual data are identified in the image. This summary should provide you with a clear structured representation of the highlighted information without including basic text content from the paper."

Challenges LLM July 19_23_p44.png"The provided page contains the following notable content: 1. **Table** - **Description:** Comparison of LLMs on mathematical reasoning tasks. - **Content:** - **Tasks:** Word-based math problems in GSM8K - **Fine-Tuning Steps:** 1. Supervised and RLHF prompts 2. Zero-shot and few-shot 3. Majority voting and reward model

4. Process-based vs. Outcome-based on final answer correctness. 2.

****Diagrams/Graphics/Equations/Tables Summary**** - ****No diagrams or graphical illustrations**** are visible in the provided page excerpt. - ****No specific equations**** are mentioned in the provided page excerpt. 3. ****Structured Data Representation of the Table****:
``plaintext { "Table": { "Comparison of LLMs on Mathematical Reasoning Tasks": [{ "Tasks":
"Word-based math problems in GSM8K", "Fine-Tuning Steps": ["Supervised and RLHF
prompts", "Zero-shot and few-shot", "Majority voting and reward model", "Process-based vs.
Outcome-based on final answer correctness"] }] } } `` 4. ****Graphics/Diagrams**** - There are
no diagrams or graphics to describe from the image provided. Note: The description is based on
the provided text regarding different fine-tuning and"

Challenges LLM July 19_23_p45.png"#### Summary of Non-Text Elements ##### Diagram -
****Sub-Human-Performance Warning****: - ****Label****: "Sub-Human-Performance [562, 607]" -
****Description****: Indicates that existing LLMs struggle to match human performance on
reasoning benchmarks. - ****Visual Element****: Contains a warning triangle, signaling caution.
References - Citation markers indicate sources of various claims and points discussed in the
text: - Webb et al. [595] - Yu et al. [654] - Ruis et al. [464] - Valmeekam et al. [562] - Kiciman et
al. [425, 253] - Ahn et al. [14] - Driess et al. [129] - Vemprala et al. [564] - Gao et al. [164] -
Srivastava et al. [519] - Jin et al. [244] - Lampinen et al. [288] - Willig et al. [607] #####
Equations - None present in the provided image. ##### Tables - No structured tables are present.
Structured Data Representation ``json { "diagrams": [{ "label": "Sub-Human-Performance
[562, 607]", "description": "Existing LLMs struggle to match human"

Challenges LLM July 19_23_p46.png"****Data in the table:**** `` LLMs in the Social Sciences &
Psychology ----- | Using LLMs to model | Analyzing
behavioral | Simulating social | | human behavior | characteristics of LLMs | relationships with
LLMs | ----- | Milgram Shock Experiment | Big Five
personality traits | Interacting artificial agents | | Illusory Truth Effect | Gulliver's Alternative Use
| LLMs to simulate societies | `` ****Descriptions and summaries of graphics and diagrams:****
****Graph/Diagram Description:**** Title: "LLMs in the Social Sciences & Psychology" Use Cases
and Examples: This diagram suggests three main use cases for Large Language Models (LLMs)
in social sciences and psychology, presented in a table format. Each category includes examples:
1. Using LLMs to model human behavior (Examples: Milgram Shock Experiment, Illusory Truth
Effect). 2. Analyzing behavioral characteristics of LLMs (Examples: Big Five personality traits,
Gulliver's Alternative Use). 3. Simulating social relationships with LLMs (Examples: Interacting
artificial agents, LLMs to simulate societies). ****Equations:**** There are no equations present in
the image. ****General Summary:**** The image from the PDF page contains a section on the
application of Large Language Models (LLMs) in social sciences and psychology. It includes a"

Challenges LLM July 19_23_p47.png"****Table:**** - ****Title:**** Social Biases [12, 367] -
****Content:**** - Unbalanced views and opinions in the training data skew the LLMs towards
biased human behaviors. ****Summary of Graphics and Diagrams:**** - There are no visible
graphics or diagrams on this page. ****Equations:**** - There are no equations present in the
provided page. ****Summarized Content:**** This page of the PDF discusses various experiments

and studies analyzing the behavior of large language models (LLMs) in replicating human psychological traits and biases. The content focuses on several key findings: 1. **Modeling Psychological Change:** Studies on LLM responses to behavioral tests suggest that LLMs can mirror human judgments in certain scenarios, such as political view changes and behavioral effects. 2. **The Turing Experiment (TE):** A framework to assess LLMs' behavior across different demographics using test items derived from social psychology and behavioral economics. The findings show alignment with human behavior, but larger models yield results more akin to human responses. 3. **Behavioral Experiments Reproduction:** Studies replicating known psychological tests demonstrated that GPT-3 replicates human participants' behaviors to some extent but misses the nuances of human cognitive systems. 4. **Personality Traits Examination:** Research on LLMs' personality traits using the Machine Personality Inventory (MPI) dataset and other scales found that GPT-3 and models like it exhibit traits similar to human personalities when analyzed"

Challenges LLM July 19_23_p48.png***Diagrams/Graphics** 1. **Figure 16: Modality Conversion.** - Description: The diagram illustrates the use of models with different input modalities in pre- or post-processing steps in an LLM pipeline. The figure is divided into two sections: Pre-processing and Post-processing. **Pre-processing:** - Three types of inputs are shown: speech input (microphone symbol), image input (image icon), text input. - These inputs go through a "Modality-to-Text" conversion. - The output text goes into a "Prompt" box. - The prompt is then fed into the "LLM." - The LLM produces an "Output." **Post-processing:** - A "Prompt" is fed into the "LLM." - The LLM output is then processed through "Code X -> Modality (speech, image, text)". - Different outputs like speech, image, and text are shown. **Equations** - There are no visible equations on this page. **Tables** - There are no visible tables on this page. Based on the image, no tables or structured data extraction is necessary. Summary of the diagram: The figure demonstrates the implementation of using models with different input modalities, either in the initial stage (Pre-processing) or in the final stage (Post-processing) of an LLM pipeline. Pre-processing involves converting various input types (speech, image, text"

Challenges LLM July 19_23_p49.png"Here is a description and transcription of the non-basic-text elements within the provided image of the PDF page: #### Graphics and Diagrams 1. **Warning Box:** - Title: Hallucinated Distributions [506] - Content: "Using LLMs for fully synthetic data generation is currently constrained by our inability to verify whether the synthetic data generated is representative of the true distribution in the corresponding real-world data." #### Tables There are no tables present in the image. #### Equations No equations are visible in the image. This summary captures the structured form of the graphics and diagrams and ensures the essential information is ready to be passed back to a database."

Challenges LLM July 19_23_p5.png"#### Diagrams, Graphics, Equations, and Tables Description ##### Diagrams and Graphics **Figure 2: Exemplary Drawbacks of relying on Tokenization** 1. **Tokenizer Training Costs** - **Training Sequences:** - English: Includes example sequences like "if instance", "for look", "level array", etc. - Chinese: Displays examples in Chinese characters. - Python: Shows Python code snippets like `for`, `False`, `def`, `array`,

etc. - **Vocabulary**: - Includes example tokens like "where", "do", "it", "look", "chinese", "example", special tokens, Chinese characters, and Python keywords. 2. **Architecture** Dependency on Vocabulary - **Structure**: - Embedding Matrix E of dimension $(E \in \mathbb{R}^{\lvert V \rvert \times D})$. - Transformer Blocks (with placeholder symbol \dots indicating multiple transformer blocks). - Softmax over Vocabulary Matrix W of dimension $(W \in \mathbb{R}^{D_{\text{model}} \times \lvert V \rvert})$. ##### Equations 1. $(E \in \mathbb{R}^{\lvert V \rvert \times D})$ 2. $(W \in \mathbb{R}^{D_{\text{model}} \times \lvert V \rvert})$ ####

Challenges LLM July 19_23_p50.png The given image is a page from a PDF document. It consists mainly of text without any visible diagrams, graphics, equations, or tables. Below is a transcription of some relevant textual references provided in a structured form suitable for a database: ``plaintext References: 1. A blog post detailed a Sam Altman freakout about a huge chips shortage threatening OpenAI. Then it was taken down. 2. Open LLM Leaderboard - a Hugging Face Space by HuggingFace. 3. Reproducibility — PyTorch 2.0 documentation. 4. 2023. Negative prompts for text generation. Section: Prompting. 5. 2023. Reproducibility. Page Version ID: 1163331755. 6. A. Abbas, K. Tirumala, D. Simig, S. Ganguli and A. S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540. 7. J. D. Abernethy, A. Agarwal, T. V. Marinov and M. K. Warmuth. 2023. A mechanism for sample-efficient in-context learning for sparse retrieval tasks. ArXiv, abs/2305.17040. 8. D. Aidwardana, M.-T. Luong, R. Dr. So, J. Hall, N."

Challenges LLM July 19_23_p51.png The provided image contains references from a research paper, without any diagrams, graphics, equations, or tables. Below is the transcription of the references captured from the image: ``plaintext [33] P. Bajaj, C. Xiong, G. Ke, X. Liu, D. He, S. Tiwary, T.-Y. Liu, P. Bennett et al. 2022. Metro: Efficient denoising pre-training of large scale autoregressive language models with model generated signals. arXiv preprint arXiv:2204.06644. [34] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato and A. Szlam. 2019. Real or Fake? Learning to Discriminate Machine from Human Generated Text. ArXiv:1906.03351 [cs, stat]. [35] R. Balestero, J. Pesenti and Y. LeCun. 2021. Learning in high dimension always amounts to extrapolation. arXiv preprint arXiv:2110.09485. [36] J. Bandy and N. Vincent. 2021. Addressing "documenta- tion debt" in machine learning research: A retrospective dataset for bootcorpus. [37] P. Barham, A. Chowdhery, J. Dean, S. Ghemawat, S"

Challenges LLM July 19_23_p52.png The provided image is a page from a bibliography section of a document. It does not contain diagrams, graphics, equations, or tables. Instead, it lists references in a structured numbered format. Here is a structured representation of the references: 1. C. Burns, H. Ye, D. Klein, J. Steinhardt. 2022. Discovering latent knowledge in language models without supervision. 2. A. Calderwood, N. Wardrip-Fruin, M. Mateas. 2022. Spinning coherent interactive fiction through foundation model prompts. International Conference of Computation and Creativity. 3. N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas et al. 2023. Poisoning Web-Scale Training Datasets is Practical. ArXiv:2302.10149 [cs]. 4. N. Carlini, C. Liu, Ö. Erlingsson, J. Kos, D. Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In

USENIX Security Symposium, volume 267. 5. N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, A. Awadalla, P. W. Koh, D. Ipp"

Challenges LLM July 19_23_p53.png"The provided image of the page from a PDF includes references to various papers. There are no diagrams, graphics, equations, or tables present in the image. The page consists solely of text from bibliographic entries. Here is a structured summary of the referenced papers: 1. Cobbe, K., Kosaraju, M., Bavarian, M., Chen, H., Jun, L., Kaiser, M., Plappert, J., and Tworek, J. (2021). Training verifiers to solve math word problems. 2. Cohen, M., Ryu, M., Chow, O., Keller, I., Greenberg, A., Hassidim, M., Fink, Y., and Matias, Y. (2022). Dynamic planning in open-ended dialogue using reinforcement learning. arXiv preprint arXiv:2208.02294. 3. Cohen, M., Hamri, M., Geva, M., and Globerson, A. (2023). LM vs LM: Detecting factual errors via cross examination. arXiv:2305.13281. 4. Computer, T. (2023). Redpajama: An open source recipe to reproduce llama training dataset. 5. Conmy, A., Mavor-Parker, A., Lynch, S., Heimersheim, A., and Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. arXiv pre"

Challenges LLM July 19_23_p54.png"The image contains only text references and no diagrams, graphics, equations, or tables. Thus, there is no data to structure or transcribe. Based on the content of the page, it consists mainly of bibliographic entries."

Challenges LLM July 19_23_p55.png"The image of the PDF page contains references and does not include any diagrams, graphics, tables, or equations that need detailed transcription or summarization. Here is a plain text extraction of the references from the image: `` [151] C. Fourier, N. Habib, J. Launay and T. Wolf. 2023. What's going on with the open \lm leaderboard? Available from: <https://huggingface.co/blog/evaluating-mmlu-leaderboard>. Accessed: 27/06/2023. [152] E. Frantar and D. Alistarh. 2023. Massive language models can be accurately pruned in one-shot. arXiv preprint arXiv:2301.00774. [153] E. Frantar, S. Ashkboos, R. Hoeffler and D. Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv:2210.17323. [154] D. Fried, A. Aghajanyan, J. Lin, S. Wang, E. Wallace, F. Shi, R. Zhong, W.-t. Yin et al. 2022. InCoder: A generative model for code infilling and synthesis. [155] A. Frömmgen and L. Kharytan. 2023. Resol"

Challenges LLM July 19_23_p56.png"The provided image contains a list of references from a research paper. There are no diagrams, graphics, or equations present within this image. Additionally, there are no tables to extract data from. Since the request was to identify any diagrams, graphics, equations, or tables, and none are present, no detailed summary is required. The page simply lists references cited in a research paper, all of which are formatted in a standard bibliographic style, typically used in academic writing."

Challenges LLM July 19_23_p57.png"This page from the PDF does not contain any diagrams, graphics, equations, or tables. It consists solely of references to various academic papers and sources. Here's a transcribed structured representation: ``plaintext References: ----- [214] D. Houser and K. McCabe. 2014. Experimental economics and experimental game theory. In *Neuroeconomics*, pages 19–34. Elsevier. [215] J. Howard and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia. Association for Computational Linguistics. [216] S. Hsiao. 2023. What’s ahead for bard: More global, more visual, more integrated. Available from: <https://blog.google/technology/ai/google-bard-updates-io-2023/>. Accessed: 28/06/2023. [217] B. Hu, J. Xia, J. Zheng, C. Tan, Y. Huang, Y. Xu and S. Z. Li. 2022. Protein language models and structure prediction: Connection and progression. [218] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen. 2022"

Challenges LLM July 19_23_p58.png"Based on the image of the page you provided, I don't see any diagrams, graphics, or equations. The entire page consists only of text, including references to various studies, papers, and preprints related to language models and natural language processing. Since there are no tables or other structured data elements to transcribe or describe from the image, here's a summary of the relevant content: ****Summary:**** The image contains a list of references primarily related to advancements in language models, protein structure prediction, and transformer-based models. These references include studies on more efficient querying and data-training methods, alignment in language agents, causal reasoning in machine learning, and handling models with open problems. Some particular studies touch on topics like MiniPile challenge for data-efficient language models, causal inference for structured treatments, effectiveness of weight averager, and large pre-trained language models alignment, among others. Each reference includes authors, year of publication, title of the work, and sometimes the conference or journal where the paper was presented or published. No structured data (such as tables or equations) were identified on the provided page."

Challenges LLM July 19_23_p59.png"The provided image is a page from a PDF document. It contains a list of academic references formatted in a standard citation style. There are no diagrams, graphics, equations, or tables within the provided image. Due to the absence of structured data types like tables, there is no data to represent for database entry. Here is a brief overview of the document content: - The page lists multiple references for academic papers and articles covering various topics related to machine learning, artificial intelligence, natural language processing, and computational linguistics. Each reference follows a format that includes the authors' names, publication year, title of the paper, publication venue, and other relevant publication details such as page numbers, DOIs, and arXiv identifiers. If you wish to extract and structure citation information for database entry purposes, I can transcribe the references. However, be aware that these references may require a bibliographic database format such as BibTeX, RIS, or a custom schema depending on your use case."

Challenges LLM July 19_23_p6.png"****Diagrams:**** 1. There is a diagram labeled "Fig. 3." It appears to represent pre-training objectives, such as target span length, high/low corruption. 2. There is a diagram labeled "Fig. 4." It seems to illustrate a concept related to scaling laws, up-stream scaling, and down-stream setups. ****Graphics:**** 1. A warning symbol followed by a highlighted text section labeled "Unsustainable Loss Power-Law [256]," which describes performance increases through larger compute budgets but at a decreasing rate if the model or dataset size is fixed. ****Equations:**** There are no explicit equations in the provided text. ****Tables:**** No tables are present in the provided text. ****Summary:**** The provided page

discusses high pre-training costs for large language models (LLMs), the scaling laws related to model performance, and ways to optimize compute costs. It suggests that while scaling laws can be used to predict performance based on compute budgets, the exact coefficients are debated. Additionally, the text mentions various pre-training objectives and their impact on self-supervised training efficiency. ****Structured Data Representation:**** - Diagrams: - Fig 3: Pre-training objectives (target span length, high/low corruption) - Fig 4: Scaling laws, up-stream scaling, down-stream setups - Graphics: - Warning symbol with text on "Unsustainable Loss Power-Law [256]": `` Performance increases through larger compute budgets but"

Challenges LLM July 19_23_p60.png"The page you provided contains only basic text citations and does not appear to contain any tables, diagrams, graphics, or equations. Therefore, there's no structured data or graphical summary to be transcribed into a database. If you have other pages or documents with tables or equations, feel free to share them, and I can help you transcribe the structured data accordingly."

Challenges LLM July 19_23_p61.png"The page from the provided PDF does not contain any diagrams, graphics, equations, or tables. It consists of references in a numbered list format. Below is a text representation of the content found on the page, organized as references: 1. R. Liu and N. B. Shah, 2023. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. ArXiv:2306.00622 [cs]. 2. S. Liu and Z. Wang, 2023. Ten lessons we have learned in the new 'sparseland': A short handbook for sparse neural network research. ArXiv:2302.02596. 3. X. Liu, X. Yang, L. Ouyang, G. Guo, J. Su, R. Xi, K. Yuan and F. Yuan, 2022. Protein language model predicts mutation pathogenicity and clinical prognosis. bioRxiv, pages 2022–09. 4. Z. Liu, A. Bahety and S. Song, 2023. Reflect: Summarizing robot experiences for failure explanation and correction. 5. Z. Liu, E. Gan, and M. Tegmark, 2023. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. arXiv preprint arXiv:2305.08746. 6. S. Longpre, L. Hou, T. Vu, A. Webson,"

Challenges LLM July 19_23_p62.png"The page from the PDF provided does not contain any diagrams, graphics, equations, or tables. It primarily consists of a list of references. No specific data is structured in a table format that can be transcribed for a database. The page number at the bottom indicates it is page 62. References are listed in a numerical sequence and include authors, titles, publication years, and journal or conference names along with additional details like volume, page numbers, and digital object identifiers (DOIs). If you need specific references to be extracted or processed, please provide further instructions on the format or details you need."

Challenges LLM July 19_23_p63.png"The image provided is a page from a PDF that contains a reference list. There are no diagrams, graphics, equations, or tables on this page. The text consists solely of bibliographic entries. Since there are no specific data tables to structure or equations to transcribe and no diagrams to describe, no additional structured data or summaries are necessary from this page. If you require any other type of processing or have another page with diagrams, tables, or equations, please provide that additional information or page."

Challenges LLM July 19_23_p64.png"The provided image contains bibliographic references from a research paper. The page does not contain any diagrams, graphics, equations, or tables.

Here is the structured text from the references into a simplified form: `` { "references": [{ "id": 424, "authors": "B. Peters and A. F. T. Martins", "year": 2021, "title": "Smoothing and shrinking the sparse Seq2Seq search space", "source": "Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", "pages": "2462–2654", "publisher": "Online. Association for Computational Linguistics." }, { "id": 425, "authors": "J. Peters, D. Janzing and B. Schölkopf", "year": 2017, "title": "Elements of causal inference: foundations and learning algorithms", "source": "The MIT Press." }, { "id": 426, "authors": "A. Petrov, E. La Malfa, P. H. Torr and A. Bibi", "year": 2023, "title": "Language model tokenizers introduce unfairness between languages", "source": "arXiv arXiv:2305.15425." }, " }

Challenges LLM July 19_23_p65.png "The provided page of the PDF document does not contain any diagrams, graphics, equations, or tables. The content includes only textual references and citations. Therefore, there are no structured tabular data, diagrams, or equations to represent from this page."

Challenges LLM July 19_23_p66.png "The page from the PDF you provided contains references. There are no diagrams, graphics, tables, or complex structured data elements other than basic references. There are no equations to transcribe. Below is a structured representation of the data found on the page: `` plaintext [484] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov. 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347. [485] M. Schuster and K. Nakajima. 2012. Japanese and korean voice search. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5149–5152. [486] T. Schuster, R. Schuster, D. J. Shah and R. Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. Computational Linguistics, 46(2):499–510. [487] R. Schwartz, J. Dodge, N. A. Smith and O. Etzioni. 2019. Green AI. ArXiv:1907.10597 [cs, stat]. [488] S. H. Schwartz, B. Breyer and D. Danner. 2015. Human values scale (ess). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). "

Challenges LLM July 19_23_p67.png "The provided image contains a list of bibliographic references from a paper. There are no diagrams, graphics, or tables. Only bibliographic entries are present, and there are no equations either. Here is the structured representation of the references as text for database entry: 1. S. Smith, M. Patwary, B. Roop, L. P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumooye et al. 2022. Using deepspeed and megatron to train megatron-turing NLG 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990. 2. I. Solaiman and C. Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. Advances in Neural Information Processing Systems 34: 5861–5873. 3. S. Soltan, S. Ananthakrishnan, J. FitzGerald, R. Gupta, W. Hama, K. Khan, C. Peris, S. Rawls et al. 2022. AlexaTM 20b: Few-shot learning using a large-scale multilingual seq2seq model. arXiv preprint arXiv:2208.01448. 4. B. Sorscher, R. Geirhos, S. Shekhar"

Challenges LLM July 19_23_p68.png "The provided image contains a list of references from a document. There are no diagrams, graphics, equations, or tables visible in the image. The content is primarily bibliographic citations. Here's a brief summary for context, formatted as plain text:


``` The provided references are related to machine learning, specifically focusing on language models, their evaluation, encoding, and various methods for improving their efficiency and accuracy. Some prominent themes include surveys on transformers, insights on pre-training and fine-tuning transformers, unifying learning paradigms, and analyzing different ways of measuring readability, understanding, and drafting responses using advanced AI models like GPT-3. There is also a focus on domain-specific language models, such as those used for biomedical text, and challenges and future opportunities in the age of large language models. Additionally, the references touch on issues like model scaling, handling data, dealing with scaling limits in machine learning, improving code models for self-improvement, and approaches to model adaptation through prompt transfer. ``` There's no data to be represented in a structured format as there are no tables provided in the image."

Challenges LLM July 19\_23\_p69.png"The image provided is a page from a PDF document consisting mainly of bibliographic references. There are no diagrams, graphics, tables, or equations present in the image. The document contains a numbered list of references in a structured format. No diagrams, graphics, equations, or tables were found within the image. The data is only basic text consisting of referenced documents."

Challenges LLM July 19\_23\_p7.png\*\*\*Diagrams and Graphics:\*\* 1. \*\*Figure 3: Masking Strategies\*\* - \*\*Description:\*\* The figure illustrates three distinct input masking strategies used in language modeling: - \*\*Masked LM:\*\* All tokens are masked. - \*\*Language Modeling:\*\* No tokens are masked. - \*\*Prefix LM:\*\* Some leading tokens are unmasked, and the rest are masked. - \*\*Visualization:\*\* The figure uses distinct color codings (assumed red for unmasked and blue for masked tokens) to represent tokens  $(x_1)$  to  $(x_9)$  (input) against outputs  $(y_1)$  to  $(y_9)$  (rows). - \*\*Key Identifier:\*\* Each row represents which inputs  $(x_i)$  (columns) a particular output  $(y_i)$  (row) can attend to (either uni- or bi-directionally). \*\*Equations:\*\* 1. \*\*Equation 1: Language Modeling Objective\*\*  $[ L(x) = \sum_{i=1}^N \log P(x_i | x_1, \dots, x_{i-1}; \theta) ]$  2. \*\*Equation 2: Masked Language Modeling Objective\*\*  $[ L(x_{\text{MASK}} | x_{\neg \text{MASK}}) = \frac{1}{|x_{\text{MASK}}|} \sum_{i \in x_{\text{MASK}}} ]$

Challenges LLM July 19\_23\_p70.png"The provided image contains a page from an academic paper with a list of references. It does not contain any diagrams, graphics, or tables. Here are the equations transcribed from this image: No equations are present. The text references are formatted as follows: 1. S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho and J. Weston. 2019. Neural text generation with unlikelihood training. arXiv preprint arXiv:1908.04319. 2. L. Weng. 2023. Large transformer model inference optimization. Lil' Log. 3. L. Weng. 2023. Prompt engineering. lillianweng.github.io. 4. M. Willig, M. ŽEČEVÍČ, D. S. Dhami and K. Kersting. 2023. Causal parrots: Large language models may talk casually but are not causal. preprint. 5. F. Winkelmolen, N. Ivin, H. F. Bozkurt and Z. Karnin. 2020. Practical and sample efficient zero-shot hop. arXiv preprint arXiv:2007.13882. 6. Y. Wolf, N. Wies, Y. Levine and A. Shashua. 2023. Fundamental limitations of aligning in large language models. arXiv preprint arXiv:2304."

Challenges LLM July 19\_23\_p71.png"Based on the image provided, here is a summary of the identified elements: It is a text-heavy document with citations. There are no diagrams, graphics,

equations, or tables visible in the image provided. Therefore, there is no data to extract in a structured form or transcribe. Each entry appears to be a reference with details including authors, title, publication year, and source, followed by a citation identifier. If you have any specific needs or additional images, feel free to provide them!"

Challenges LLM July 19\_23\_p72.png"The given image is of a page from a PDF that contains references. There are no diagrams, graphics, tables, or equations presented in the image. The content includes a list of references from various research papers and articles. Therefore, no structured data tables, graphics summary, or mathematical equations are available to transcribe or describe. For clarity and accuracy, here is the list of references transcribed as text: 1. S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, D. Dewan, M. Diab et al. 2022. Opt: Open pre-trained transformer language models. > arXiv preprint arXiv:2205.01068. 2. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi. 2019. Bertscore: Evaluating text generation with bert. > arXiv preprint arXiv:1904.09675. 3. T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown and T. B. Hashimoto. 2023. Benchmarking large language models for news summarization. 4. Z. Zhang, Y. Gu, X. Han, S. Chen, C. Xiao, Z. Sun, Y. Yao, F. Qi et al. 2021"

Challenges LLM July 19\_23\_p8.png"The page contains diagrams illustrating different pre-training objectives for a language model, focusing on self-supervised data construction. Here's a structured form of the content and description: #### Diagrams and Data: 1. \*\*Span Corruption (R-Denoising)\*\* - \*\*Inputs:\*\* `` [Some proponents of AI consciousness adhere to functionalism, the view that mental states are defined more by their functions than their underlying physical structures. In other words, if an AI can respond to inputs and produce outputs similar to a conscious being, then it could be considered conscious. However, this view doesn't account for subjective] `` - \*\*Targets:\*\* `` [4] [3] [2] `` 2. \*\*Prefix Language Modeling (S-Denoising)\*\* - \*\*Inputs:\*\* `` [Some proponents of AI consciousness adhere to functionalism, the view that mental states are defined more by their functions than their underlying physical structures. In other] `` - \*\*Targets:\*\* `` [56] `` 3. \*\*Long Span Corruption (one form of X-Denoising)\*\* - \*\*Inputs:\*\* `` [Some proponents of AI consciousness adhere to functionalism, the] [12] [underlying physical structures. In other words, if an AI can respond to] [13] [considered conscious. However, this view doesn't account for subjective] [14]"

Challenges LLM July 19\_23\_p9.png"#### Analysis of the PDF Page ##### Diagrams and Graphics: There are no diagrams or graphics present on the page. ##### Equations: There is a single equation labeled (3): 
$$(3) \quad \left[ \begin{aligned} & \sum_{x \in S} \text{Bigg}(-\log \overrightarrow{p}(x_i \mid x_{<i}; \theta) \parallel \text{NLL for forward model} \parallel & \\ & -\log \overleftarrow{p}(x_i \mid x_{>i}; \theta) \parallel \text{NLL for backward model} \parallel & \\ & + \beta D_i^{\text{TV}}(\overrightarrow{p} \parallel \overleftarrow{p}) \text{Bigg} \end{aligned} \right] \text{where } (D_i^{\text{TV}}(\overrightarrow{p} \parallel \overleftarrow{p})) \text{ is the total variation distance among the two models on the } (i)\text{-th token.} \\ \text{##### Tables: There are no tables present on the page.} \\ \text{#### Summary: 1. **Equations**}: \\ \text{There is a single equation describing the pre-training loss for sequence models using forward and backward language models, incorporating a total variation distance term as a regularizer.} \\ \text{``plaintext equation: } \sum_{x \in S}$$

Continual\_Pretraining\_p1.png"The image contains a bar chart (Figure 1) showing domain-specific task performance in biomedicine, finance, and law. Below is the extracted and structured data from the table: **\*\*Biomedicine\*\*** - CHEMPROT - General LLM: 56 - DAPT: 57 - AdaptLLM: 59 - RCT - General LLM: 53 - DAPT: 54 - AdaptLLM: 56 - HoC - General LLM: 51 - DAPT: 52 - AdaptLLM: 54 - PubMedQA - General LLM: 43 - DAPT: 45 - AdaptLLM: 47 **\*\*Finance\*\*** - CONDQA - General LLM: 66 - DAPT: 70 - AdaptLLM: 71 - PTB - General LLM: 61 - DAPT: 63 - AdaptLLM: 65 - FCSA - General LLM: 58 - DAPT: 59 - AdaptLLM: 60 - Headline - General LLM: 43 - DAPT: 49 - AdaptLLM: 51 **\*\*Law\*\*** - SCOTUS - General L"

Continual\_Pretraining\_p10.png"The page contains primarily text, with a few features of note: **#### Diagrams and Graphics:** 1. There is a large "X" mark on the page, likely indicating a highlight or a correction. **#### Equations:** There are no equations present on this page. **#### Tables:** There are no tables present on this page. **#### Summary of Key Sections:** **\*\*Section: Instruction Fine-tuning\*\*** - Fine-tuning LLMs on domain-specific tasks improves their performance. - Utilizes question-answering instructions. - Effective in specific domains: [Singhal et al., 2022], [Le et al., 2023b], [Wang et al., 2023], [Han et al., 2023], [Xiong et al., 2023], [Huang et al., 2023]. - Challenges in creating large-scale supervised datasets for instruction tuning. - Supervised fine-tuning data often limited. - Generative models like ChatGPT and GPT-4 are used to produce data. - Pre-training can alleviate some limitations of inferred codes from closed-source models. **\*\*Section: Retrieval-augmented Prompting\*\*** - Enhances LLMs by integrating external information without altering model parameters [Li et al., 2023b], [Cui et al., 2023] [Huang et al., 2023]. - Utilizes domain-specific knowledge from documents, domain-specific knowledge graphs, neural networks. - Improvement through better context understanding from"

Continual\_Pretraining\_p11.png"The provided image includes a list of references from a paper. The references contain various textual details, but there are no diagrams, graphics, or equations present on the page. Here's the identified structured data from the text extracted from tables: 1. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. - Title: A large annotated corpus for learning natural language inference. - Conference: EMNLP - Pages: 632-642 - Year: 2015 - Association: Association for Computational Linguistics 2. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. - Title: Language models are few-shot learners. - Conference: NeurIPS - Year: 2020 3. Ilias Chalkidis. "

Continual\_Pretraining\_p12.png"The provided image primarily contains a list of references from a research paper, which includes the titles, authors, publication forums, and years of various cited works. There are no diagrams, graphics, specific equations, or structured tables in the provided image. Here's the formatted information of the references for database entry: 1. Title: The false promise of imitating proprietary llms. Authors: Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, Dawn Song. Forum: CoRR

Reference ID/DOI: abs/2305.15171 Year: 2023 2. Title: Don't stop pretraining: Adapt language models to domains and tasks. Authors: Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith. Forum: ACL Pages: 8342–8360 Year: 2020 3. Title: MedAlpaca - an open-source collection of medical conversational AI models and training data. Authors: Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, Keno K. Bressem. Forum: CoRR Reference ID/DOI: abs/

Continual\_Pretraining\_p13.png"There are no diagrams, graphics, or equations in this image. However, there appears to be a table of references formatted in a specific citation style. Below is the structured data of the table in text format: `` [ { "Author(s)": "Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni", "Title": "CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service", "Journal": "Artif. Intell. Law", "Volume": "27(2)", "Pages": "117–139", "Year": "2019" }, { "Author(s)": "Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov", "Title": "Roberta: A robustly optimized BERT pretraining approach", "Journal": "CoRR", "Volume": "abs/1907.11692", "Year": "2019" }, { "Author(s)": "Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V."

Continual\_Pretraining\_p14.png"The page consists of a list of bibliographic references. There are no diagrams, graphics, tables, or equations present. Therefore, there is no data to represent in a structured form or text. The content includes references to various academic papers and articles related to computational linguistics, language models, and machine learning. Each entry lists the authors, title, conference or journal, page numbers, year, and sometimes URLs. This text data does not require transcription for any diagrams, graphics, tables, or equations as none are present."

Continual\_Pretraining\_p15.png"The image contains a list of references from a research paper. There are no diagrams, graphics, or equations present on the page. Here's a structured representation of the data listed in the table (reference list): 1. Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. \*NAACL-HLT\*, pp. 1112-1122. Association for Computational Linguistics, 2018. 2. Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further fine-tuning llama on medical papers. \*CoRR\*, abs/2304.14454, 2023a. 3. Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhakaran Kambadur, David S. Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. \*CoRR\*, abs/2303.17564, 2023b. 4. Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. \*CoRR\*, abs/2304.01097, 2023"

Continual\_Pretraining\_p16.png"\*\*Diagrams, Graphics, and Tables Identified:\*\* - \*\*Table 6: Domain knowledge probing results.\*\* - \*\*Structured Data from Table 6:\*\* | Domain | General LLM | Raw Text | Read. Text | Read. Compre. | |-----|-----|-----|-----|-----| | BioMed | 36.5 | 36.9 | 36.8 | | Law | 45.0 | 45.6 | 46.4 | --- \*\*Summary of Graphics and

Diagrams:\*\* - \*\*Table 6\*\* provides the results of the domain knowledge probing. The table compares performance metrics for two domains (BioMed and Law) across four models: General LLM, Raw Text, Read. Text, and Read. Compre. The values in the cells represent scores indicating the performance of the models on domain-specific tasks. \*\*Equations:\*\* No equations were present on the page. --- The above data representation can be passed directly back to a database."

Continual\_Pretraining\_p17.png\*\*\*Tables:\*\* 1. \*\*Table 7: Pre-training corpora\*\* | Domain | Data Source | Raw Size | # Tokens | # Docs | |-----|-----|-----|-----|  
 --|-----| | BioMed | PubMed Abstracts (Gao et al., 2021) | 19.3 GiB | 5.4 B | 15.5 M | | Finance | Stock News (Gao et al., 2021) | 5.1 GiB | 1.2 B | 1.1 M | | Law | FreeLaw Opinions (Gao et al., 2021) | 51.2 GiB | 16.7 B | 3.6 M | 2. \*\*Table 8: Hyper-parameters of domain-adaptive pre-training\*\* | Hyperparameter | Assignment | |-----|-----|  
 Computing infrastructure | 32 V100-32GB GPUs | | Runtime | 24 Hours | | Number of steps | 10,000 | | Batch size | 32 | | Maximum sequence length | 2,048 | | Maximum learning rate | 1e-5 | | Optimizer | Adam | | Adam beta weights | 0.9,"

Continual\_Pretraining\_p18.png\*\*\*Diagrams/Graphics/Equations/Tables Summary\*\* 1. \*\*Table 9: Keywords that compile into regular expressions.\*\* - \*\*Columns:\*\* Keyword, Regex -  
 \*\*Data:\*\* - {VERBAL}: Replaced with the verbalizer - {WORD}: regex: (([ · !?., ]\n )  
 [^\s]{10,}) - Matches a single word having more than 9 characters - {SENT}: regex:  
 (([ · !?., ]{50,} ) [!?\n+ }) - Matches a single sentence having more than 50 characters 2.  
 \*\*Table 10: Specifications of the domain-specific task datasets.\*\* - \*\*Columns:\*\* Task, Type, Metric, # Demos - \*\*Data:\*\* - BioMed: - MOP: Binary classification, Accuracy, 4 - PubMedQA: Binary classification, Accuracy, 0 - USMLE: Multi-choice QA, Accuracy, 0 - RCT: Multi-class classification, Micro F1, 10 - ChemProt: Multi-class classification, Micro F1, 13 - Finance: - FiQA SA: Multi-class classification, Weighted F1, 5 - FPB: Multi-class classification, Weighted F1, 5 - NER: Named entity recognition, Entity-level F1, 20 - Head"

Continual\_Pretraining\_p19.png\*\*\*Table Data\*\* ``` TableName: PromptTemplates Columns: - Task - Template Rows: 1: Task: BioMed. Template: - MQP - Question 1: {QUESTION1} - Question 2: {QUESTION2} - Are questions 1 and 2 asking the same thing? {ANSWER} - PubMedQA - Context: {CONTEXT} - Question: {QUESTION} - Answer: {ANSWER} - USMLE - Question: {QUESTION} - Answer: {ANSWER} - RCT - {SENTENCE} - Question: What is the role of this sentence in an abstract? - Answer: {ANSWER} - ChemProt - {SENTENCE} - Question: What is the relation? - Answer: {ANSWER} 2: Task: Finance Template: - FiQA SA - {SENTENCE} - Question: What is the sentiment on {TARGET}? - Answer: {ANSWER} - FPB - {SENTENCE} - Question: What is the sentiment? - Answer: {ANSWER} - NER - {SENTENCE} - Extract named entity: {ANSWER} - Headline - {SENTENCE} - Question: {QUESTION} -"

Continual\_Pretraining\_p2.png"The image contains a figure (Figure 2) and a brief description of it. Here is the transcription and structured summary: --- \*\*Diagrams and Graphics:\*\* \*\*Figure Description:\*\* Figure 2: A simplified example of a reading comprehension text, wherein the raw

text is followed by a series of tasks constructed from it, including: - Summarization (purple) - Word-to-Text (blue) - Natural Language Inference (red) - Commonsense Reasoning (teal) - Paraphrase Detection (yellow) - Text Completion (green). --- **\*\*Summary of Graphics/Diagrams:\*\*** - **\*\*Raw Text Section:\*\*** This section includes a snippet about "Glottic Carcinoma in Young Patients." Various elements are highlighted in different colors (Title, Domain Keywords, Entailment Relation, Cause & Effect, Semantic Similarity, and Text Ending). - **\*\*Reading Comprehension Section:\*\*** This section is based on the Raw Text. It includes various types of questions and prompts: - Summarization: "What is a summary Glottic Carcinoma in Young Patients." - Entailment: "Generate a sentence that includes these biomedicine keywords (carcinoma, oropharyngeal, papillomavirus) Recent reported evidence indicates that vocal cord carcinoma is evolving..." - Explanation Tasks: "What is the reason for ...?" - Contradiction Tasks: "Compose a sentence that contradicts the meaning..." - Completion"

Continual\_Pretraining\_p20.png**\*\*Diagrams and Graphics:\*\*** 1. **\*\*Pie Charts (Figure 4):\*\*** - **\*\*Biomedicine\*\*** - Word-to-Text: 50.3 - Summarize: 26.3 - Text Completion: 17.4 - Common Reason: 0.3 - Paraphrase: 3.5 - NIL: 2.2 - **\*\*Finance\*\*** - Word-to-Text: 35.8 - Summarize: 31.4 - Text Completion: 20.7 - Common Reason: 0.5 - Paraphrase: 7.8 - NIL: 3.8 - **\*\*Law\*\*** - Word-to-Text: 62.3 - Summarize: 7.2 - Text Completion: 22.0 - Common Reason: 1.0 - Paraphrase: 4.9 - NIL: 2.7 2. **\*\*Bar Charts (Figure 5):\*\*** - **\*\*Domain Tasks:\*\*** - All: 44 - w/o Summarize: 43 - w/o Word-to-Text: 40 - w/o NIL: 43.5 - w/o Common Reason: 44 - w/o Paraphrase"

Continual\_Pretraining\_p21.png#### Data from Table 12: **\*\*Fine-tuning performance on domain-specific tasks of general large language model (General LLM)\*\*** ##### Biomedical Domain: - **\*\*Data Type:\*\*** - **\*\*BioMedQA | PubMedQA | ChemProt | MQP | RCT | UMSLE | AVERAGE\*\*** - **\*\*General LLM:\*\*** - 75.4, 64.5, 87.0, 35.0, 87.0, 38.5, 64.2 - **\*\*Raw Text:\*\*** - 76.2, 64.8, 65.6, 87.0, 39.0, 66.5 - **\*\*Read. Compre.:\*\*** - 76.0, 65.4, 87.9, 87.5, 41.0, 71.5 ##### Finance Domain: - **\*\*Data Type:\*\*** - **\*\*ConvFinQA | FPB | FiQA SA | Headline | NER | AVERAGE\*\*** - **\*\*General LLM:\*\*** - 58.1, 81.9, 86.4, 95.7, 77.5, 79.9 - **\*\*Raw Text:\*\*** - 56.2, 83.3, 87.9, 95.8, 81.3, 80.9 - **\*\*"**

Continual\_Pretraining\_p22.png#### Table: Prompting results on general LLM benchmarks **\*\*Columns:\*\*** - Task - Metric - General LLM Raw - General LLM Read - BioMed. Raw - BioMed. Read - Finance Raw - Finance Read - Law Raw - Law Read **\*\*Data:\*\*** 1. **\*\*Summarization\*\*** - **\*\*AGNews (Zhang et al., 2015)\*\*** - Acc: 58.7, 51.7, 55.5, 56.1, 50.1, 57.8, 60.6 - R-1: 1.5, 3.6, 7.5, 1.9, 10.8, 3.4, 7.4 - R-2: 0.2, 0.9, 2.8, 0.3, 8.3, 0.8, 2.7 - R-L: 1.5, 3.6, 7.2, 1.8, 10.3, 3.3, 7.2 - **\*\*AESLC (Zhang & Tetreault, 2019)\*\*** - R-1: 0.6, 3.8, 9.3, 3.1, 13.2, "

Continual\_Pretraining\_p23.png**\*\*Identified Components:\*\*** - **\*\*Table:\*\*** - **\*\*Columns:\*\*** - **\*\*Category:\*\*** Tasks related to various text comprehension categories. - **\*\*Task Description:\*\*** Detailed description of each task. - **\*\*Figure:\*\*** - **\*\*Description:\*\*** An example of a reading comprehension text constructed from raw text. The underlined sentence is added to guide the model to answer questions based on the given context. **\*\*Table Data:\*\*** `` Category: - Raw Text Task Description: Here is the first part of an article about biomedicine: Recent reported evidence indicates that vocal cord carcinoma is evolving similarly to oropharyngeal cancer with an

increasing number of patients without a smoking history having human papillomavirus (HPV) disease. Observations also suggest that an increasing number of patients who present with glottic carcinoma are younger than has been reported in the past. However, recent published evidence shows that glottic carcinoma can be an HPV-related disease with increasing incidence in nonsmokers. It isn't surprising that alternate malignant pathways may have a different timeline. - Summarization Task Description: What is a summary? Glottic Carcinoma in Young Patients. - Word-to-Text Task Description: Generate a sentence that includes these biomedical words [carcinoma, oropharyngeal, papillomavirus]: Recent reported evidence indicates that vocal cord carcinoma is evolving similarly to oropharyngeal"

Continual\_Pretraining\_p24.png\*\*\*Table: Case of a reading comprehension text in biomedicine domain\*\* | Description | Content | |-----|-----| | Title | Case of a reading comprehension text in biomedicine domain | | Content | Certain portions are omitted for brevity and are represented as (...). | --- \*\*Diagrams and Graphics:\*\* None present in the image. \*\*Equations:\*\* There are no equations present in the image. \*\*Summary:\*\* The table presents a reading comprehension text focused on the biomedicine domain, specifically regarding the biochemistry of chromogranin A-derived peptide (PST). It discusses how PST influences glucose, lipids, protein metabolism, and insulin action in rat adipocytes. PST has an overall counter-regulatory effect on insulin action by activating a specific receptor-effector system and stimulates both basal and insulin-mediated protein synthesis. The text describes experimental methods, including the use of Western blot for assessing PST stimulation of S6 kinase activity and phosphorylation of specific proteins involved in the translation process. Questions for reading comprehension are provided along with their descriptions. Key biomedicine concepts are also mentioned, like phosphorylation and the impact on translation machinery. \*\*Structured Data Representation:\*\* Table: Case of a reading comprehension text in biomedicine domain - \*\*Title:\*\* Case of a reading comprehension text in biomedicine domain - \*\*Content:\*\* Certain portions are omitted for brevity and are represented as (...). - \*\*Summary Topics:\*\* - PST and"

Continual\_Pretraining\_p25.png#### Data Extraction from the Table: \*\*Table 15: Case of a reading comprehension text in finance domain\*\* --- #### Structured Data for Database: | \*\*Company Name\*\* | \*\*Abbrev.\*\* | \*\*Market Capitalization\*\* | \*\*Incorporation Date\*\* | \*\*Location\*\* | |-----|-----|-----|-----|-----| | | Casella Waste Systems | CWST | Not provided | Not provided | Not provided | | Stericycle, Inc. | SRCL | \$4.73 billion | Incorporated in 1989 | Bannockburn, Illinois | | ABM Industries Inc. | ABM | \$2.98 billion | Incorporated in 1985 | New York, New York | --- #### Diagrams and Graphics: None identified from the given page. --- #### Equations: None identified from the given page. --- #### Notes: - The table mentions that only the brief information about companies is given. - Additional details or omitted portions are represented as (...) within the text. ---"

Continual\_Pretraining\_p26.png"The image provided contains a page from a PDF, specifically a legal text including some descriptive elements and specific figures related to a case of restitution. Based on your request, here is the structured information extracted from the image: #### Table: \*\*Title:\*\* Case of a reading comprehension test in law domain. Certain portions are omitted for

brevity and are represented as (...). 1. **Amount of Restitution Ordered:** - **Value:** \$5,829,334.90 - **Without Interest:** Yes 2. **Proof of Additional False Claim:** - **Value:** \$488,000 3. **Claims for which no Proof of Falsity was Conceded by the Government:** - **Floresitine Baker:** \$18,000 - **Shirley E.:** [Value not provided] **Descriptive Elements:** 1. **Primary Subject:** - Restitution ordered to Arledge pursuant to the Mandatory Victims Restitution Act of 1996. 2. **Dispute:** - Arledge disputes the calculation used to determine the amount of loss, particularly related to the Fen Phen II settlement. 3. **Categories of Evidence Used:** - Testimony of Wyatt (an S&A employee who created fraudulent documents). - Testimony of two pharmacists (testified specific prescriptions alleged from their pharmacies were fraudulent). - Representations by AHP. **Summary of Graphics/Di**"

Continual\_Pretraining\_p3.png"The image contains a table and a small diagram. Here is a structured representation of the data within the table and a summary of the diagram present in the image. **Table Data (Text Representation):** **Table 1: Domain-specific task scores of general language model (General LLM) and the language model that has undergone continued pre-training on the domain-specific raw corpora (DAPT).**

Method	Prompting	Fine-tuning	Knowledge Prob
General LLM	44.2	58.6	34.2
DAPT	41.7	57.6	35.0

**Diagram Summary:** There is a highlighted section within the text that visually emphasizes important information about the experiment findings. Highlighted text summarizes the exploration and findings related to domain-specific pre-training for large language models. **Equations:** There are no mathematical equations visible in the provided image. --- This formatted data is"

Continual\_Pretraining\_p4.png"Based on the provided image of the PDF page, here is a structured summary of the included diagrams, graphics, equations, and tables. Please note that I'm ignoring the basic text of the paper and focusing on the requested elements. **Diagrams and Graphics** - There are references to "Figure 2" and "Table 2" within the text, indicating the presence of these items somewhere in the document. However, they do not appear on the provided page. **Equations** - There are no visible mathematical equations on this page. **Tables** - The text includes a reference to "Table 2" but does not show the actual table on this page. The reference mentions that Table 2 describes techniques used to extract and create tasks from raw texts. **Structured Data Representation** Since the page does not contain an actual table, I cannot represent table data in a structured form. However, based on the text, here is a structured representation of the referenced information: **Figure References** - **Figure 2:** Mentioned in the context of illustrating the methodology or tasks. **Table References** - **Table 2:** Summarizes techniques used to extract and create tasks from raw texts. **Techniques Mentioned** 1. **Answer Questions Based on the Article:** - Technique to enhance task diversity. - Parsed raw text with specific phrases like "Answer questions based on the article". 2. **Summarization:** - Prompt to generate a concise summary"

Continual\_Pretraining\_p5.png"The image contains a table identified as "Table 2: Mining patterns and input-output templates." Below is the structured representation of the data in the table:



``plaintext Table 2: Mining patterns and input-output templates {VERBAL} is replaced with the verbalizers in Table 3. For mining, {WORD} captures a single word, and {SENT} captures a single sentence. Each input-output template is paraphrased into multiple variations. We also turn the task around—exchanging the question and answer—to achieve enhanced diversity. Task Type Mining Pattern Input-output Template -----

----- Summarization Title Title as summary What is a summary? {TITLE} Topic {SENT1} {VERBAL} {SENT2} {SENT1} is about: {SENT2} Word-to-Text Word-to-text Domain keywords as input; Generate a sentence about these {DOMAIN} keywords {WORD1}, {WORD2}, {WORD3}: {SENT} sentence as output Definition {WORD} {VERBAL} {SENT} How to define {WORD}? {SENT} Natural Language Inference Entail {SENT1} {VERBAL} {SENT2} Does "{SENT1}" entail "{SENT2}"? {Yes/Maybe/No} Neutral Contradict Commonsense Reasoning"

Continual\_Pretraining\_p6.png\*\*\*Table: Verbalizers for mining patterns in Table 2\*\* | Task Type | Verbalizer | |-----|-----|  
--|| \*\*Summarization\*\* || | Topic | talks about, is about, 's topic is | | \*\*Word-to-Text\*\* || | Definition | is defined as, 's definition is | | \*\*Natural Language Inference\*\* || | Entail | Yes, Therefore, Thus, Accordingly, Hence, For this reason | | Neutral | Maybe, Furthermore, Additionally, Moreover, In addition | | Contradict | No, However, But, On the contrary, In contrast, Whereas | | \*\*Commonsense Reasoning\*\* || | Cause-effect | Therefore, Thus, Accordingly, Hence, For this reason | | Effect-cause | due to, on account of, owing to | | \*\*Paragraph Detection\*\* || | Similar | In other words, Namely, That is to say, Similarly, Equally | | Different | No, However, But, On the contrary, In contrast, Whereas |"

Continual\_Pretraining\_p7.png"The image contains a table under the heading "Table 4: Domain-specific task performance of general large language model (General LLM), vanilla domain-adaptive pretraining (DAPT), and ours (AdaptLLM) in prompting evaluation." Here is the data from the table in structured text format: ``plaintext Biomedicine | Model | PubMedQA | ChemProt | MQP | RCT | UMSLE | AVERAGE | |-----|-----|-----|-----|-----|-----|  
-----| | MedAlpaca-7B | 58.6 | 39.0 | 50.7 | 40.8 | 36.7 | 45.1 | | MedAlpaca-13B | 60.7 | 38.4 | 57.4 | 51.3 | 39.0 | 49.4 | | General LLM-7B | 59.6 | 31.4 | 50.7 | 45.1 | 34.5 | 44.2 | | DAPT-7B | 52.6 | 26.6 | 49.2 | 46.6 | 33.5 | 41.7 | | AdaptLLM-7B | 63.3 | 35.2 |"

Continual\_Pretraining\_p8.png"Here is the data extracted and structured from the table present in the image: #### Table 5: Ablation results on training data \*\*Header:\*\* Data | Raw Text | Read. Compr. | Gen. Ins. | Raw + Gen. Ins. | Read. + Gen. Ins. \*\*Rows:\*\* - \*\*BioMed:\*\* - Raw Text: 41.7 - Read. Compr.: 44.3 - Gen. Ins.: 43.3 - Raw + Gen. Ins.: 44.8 - Read. + Gen. Ins.: 47.3 - \*\*Finance:\*\* - Raw Text: 57.6 - Read. Compr.: 60.0 - Gen. Ins.: 62.2 - Raw + Gen. Ins.: 61.7 - Read. + Gen. Ins.: 63.4 - \*\*Law:\*\* - Raw Text: 35.0 - Read. Compr.: 37.0 - Gen. Ins.: 37.8 - Raw + Gen. Ins.: 34.7 - Read. + Gen. Ins.: 38.5 #### Description of the Table: \*\*Title:\*\* Ablation Results on Training Data. \*\*Description:\*\* The table shows the performance scores (in percentages) for different domains—BioMed, Finance, and Law—under various"

Continual\_Pretraining\_p9.png"#### Identified Elements ##### Diagrams and Graphics 1. \*\*Bar Chart (Domain Knowledge)\*\* - Three bars representing "BioMed.," "Finance," and "Law"

categories. - Each category is represented in three different conditions: - General LLM - Raw Text - Read. Compre. 2. **Spider/Radar Chart (Prompting Ability)** - Categories: - Summarize - Word-to-Text - N.L.I - Common. Reason. - Paraphrase - Text Comple. - Close. QA - Read. Compre. - Values plotted for three different conditions: - General LLM - Raw Text - Read. Compre. ##### Table Data **Table Representation of Bar Chart** ```plaintext Category | General LLM | Raw Text | Read. Compre. -----|-----|-----|----- BioMed. | 60 | 75 | 85 Finance | 70 | 74 | 77 Law | 50 | 60 | 65 ``` **Values from Radar/Spider Chart** ```plaintext Ability Category | General LLM | Raw Text | Read. Compre. -----|-----|-----|----- Summarize | 41 | 47 | 49 Word-to-Text "

cs224n-2023-lecture11-prompting-rlhf\_p1.png"The image provided is a title slide from a presentation. Title: "Natural Language Processing with Deep Learning" Course: "CS224N/Ling284" Graphic: An abstract representation of a rooftop and arches, which could symbolize an educational institution or structure. Instructor: "Jesse Mu" Lecture Title: "Lecture 11: Prompting, Instruction Finetuning, and RLHF" Since this is a title slide, there are no diagrams, tables, or equations to transcribe or describe. Data to be saved in database: ``` { "title": "Natural Language Processing with Deep Learning", "course": "CS224N/Ling284", "instructor": "Jesse Mu", "lecture\_title": "Lecture 11: Prompting, Instruction Finetuning, and RLHF", "graphic\_description": "An abstract representation of a rooftop and arches, which could symbolize an educational institution or structure." } ``` "

cs224n-2023-lecture11-prompting-rlhf\_p10.png**Diagrams/Graphics/Equations Identification:** 1. **Diagram/Graphic:** - There is a graphic on the right-hand side showing a conversation about suggestions for a 3-course vegetarian menu with a chocolate dessert. **Data in the Graphic:** - **Starter:** 1. Wild Mushroom Tartlets with Onion Sauce [1]. 2. Vegan Popcorn Tofu Nuggets [2]. - **Main Course:** 1. Vegan Butternut Squash Mac [2]. 2. Vegetarian Three-Bean Chili [3]. - **Dessert:** 1. Chocolate Lava Cake [4]. 2. Chocolate Pasta with Chocolate Hazelnut Cream Sauce, White Chocolate Shavings, and Fresh Berries [4]. **Summary:** The graphic presents options for a 3-course vegetarian menu designed for a dinner party of six people, ending with a chocolate dessert. Detailed options include starters, main courses, and desserts, with the source links provided for more information. - **Source Links:** 1. booths.co.uk +10 more. **Structured Text Representation for Database:** ``` { "menu\_suggestions": { "starter": [ { "name": "Wild Mushroom Tartlets with Onion Sauce", "source": "1" }, { "name": "Vegan Popcorn Tofu Nuggets", "source": "2" } ] }

cs224n-2023-lecture11-prompting-rlhf\_p11.png"The image contains a graphical table titled "ChatGPT," divided into three categories: Examples, Capabilities, and Limitations. Below is the structured text representation of the data: **ChatGPT** - **Examples** - "Explain quantum computing in simple terms" - "Got any creative ideas for a 10-year-old's birthday?" - "How do I make an HTTP request in Javascript?" - **Capabilities** - Remembers what user said earlier in the conversation - Allows user to provide follow-up corrections - Trained to decline inappropriate requests - **Limitations** - May occasionally generate incorrect information - May occasionally produce harmful instructions or biased content - Limited knowledge of the world and events after 2021 This data is now ready to be passed directly to a database."

cs224n-2023-lecture11-prompting-rlhf\_p12.png "The image provided contains a structured list from a lecture plan. There are no diagrams, graphics, equations, or tables present in the image. Here is the structured data identified in the image: `### Lecture Plan: From Language Models to Assistants 1. Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning 2. Instruction finetuning 3. Reinforcement Learning from Human Feedback (RLHF) 4. What's next? This data can be directly used for reference or input into a database.`"

cs224n-2023-lecture11-prompting-rlhf\_p13.png "The image appears to be a slide from a lecture plan. The content is structured as a list with four items. Here is the structured representation of the data, which will be passed to a database: ``` { "lecture_plan": [ { "item": 1, "title": "Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning" }, { "item": 2, "title": "Instruction finetuning" }, { "item": 3, "title": "Reinforcement Learning from Human Feedback (RLHF)" }, { "item": 4, "title": "What's next?" } ] } ``` This format organizes the information from the slide in a structured manner suitable for a database."

cs224n-2023-lecture11-prompting-rlhf\_p14.png "The image contains the following diagrams, equations, and structured data: `### Diagrams: 1. **Transformer Decoder Diagram**: - Represents a set of interconnected blocks (likely layers or processing units) labeled "Decoder," with arrows indicating the flow of information between them. 2. **Textual Entailment Example**: - Shows an example of textual entailment with the text: `` [START] The man is in the doorway [DELIM] The person is near the door [EXTRACT] `` - "entailment" is indicated as the relation between the two sentences. #### Structured Data: ##### GPT Parameters and Information: - **Model**: GPT - **Parameters**: 117 million (117M) - **Study Reference**: Radford et al., 2018 - **Features**: - Transformer decoder with 12 layers. - Trained on BookCorpus: over 7000 unique books, and 4.6GB of text. ### Description Summary for Database: ``json { "diagrams": [ { "type": "Transformer Decoder Diagram", "description": "Interconnected blocks (layers/processing units) labeled 'Decoder' with arrows showing information flow." }, { "type": "Textual Entailment Example", "description": "[START] The man is in the doorway [DELIM] The person is near the door [EXTRACT]. " }`"

cs224n-2023-lecture11-prompting-rlhf\_p15.png "\*\*Diagrams and Graphics:\*\* - The page contains a diagram with a title: "Language Models are Unsupervised Multitask Learners." **\*\*Tables:\*\*** - No tables are present in the image. **\*\*Equations:\*\*** - No equations are present in the image. **\*\*Summary of Graphics and Diagrams:\*\*** - The graphic titled "Language Models are Unsupervised Multitask Learners" lists the authors of the corresponding paper: - Alec Radford - Jeffrey Wu - Rewon Child - David Luan - Dario Amodei - Ilya Sutskever **\*\*Summary of Key Points:\*\*** - The text discusses GPT-2 (2019), a model with 1.5 billion parameters. - It mentions that GPT-2 has the same architecture as GPT but is larger (from 117M to 1.5B parameters). - GPT-2 is trained on significantly more data (from 4GB to 40GB of internet text data, termed WebText). - Data is collected by scraping URLs posted on Reddit with at least 3 upvotes, serving as a rough proxy of human quality. This structured information can be directly fed into the database for further use."

cs224n-2023-lecture11-prompting-rlhf\_p16.png "The image from the PDF contains the following: **\*\*Diagrams/Graphics:\*\*** - There are no diagrams or graphics present in this specific

image. **Equations:** -  $P(\text{Is } \{ \text{because the cat was too big} \} | \text{because the hat was too big})$  **Tables:** - There are no tables present in this specific image. **Summary:** This section of the PDF discusses the concept of emergent zero-shot learning in GPT-2, which refers to the system's ability to perform various tasks without examples or gradient updates. It presents two key abilities: 1. Specifying the right sequence prediction problem, exemplified through a question about Tom Brady. 2. Comparing probabilities of sequences using a given example from the Winograd Schema Challenge (Levesque, 2011). **Text Data for Database:** `plaintext { "title": "Emergent zero-shot learning", "key_points": [ { "point": "Specifying the right sequence prediction problem (e.g. question answering)", "example": "Passage: Tom Brady... Q: Where was Tom Brady born? A: ..." }, { "point": "Comparing probabilities of sequences (e.g. Winograd Schema Challenge)", "example": "The cat couldn't fit into the hat because it was too big. Does it = the cat or the hat?" }, { "point": "The cat couldn't fit into the hat because it was too big. Does it = the cat or the hat?" } ] }`

cs224n-2023-lecture11-prompting-rlhf\_p17.png The image contains a description of a benchmark comparison followed by a table. Here is the structured text representation: **Benchmark Table** **Columns** - **Model Size**: Different sizes of the GPT-2 model. - **LAMBADA (PPL)**: Perplexity score on LAMBADA dataset. - **LAMBADA (ACC)**: Accuracy score on LAMBADA dataset. - **CBT-CN (ACC)**: Accuracy score on CBT-CN dataset. - **CBT-NE (ACC)**: Accuracy score on CBT-NE dataset. - **WikiText2 (PPL)**: Perplexity score on WikiText2 dataset. **Rows** 1. **SOTA**: - **LAMBADA (PPL)**: 99.8 - **LAMBADA (ACC)**: 59.23 - **CBT-CN (ACC)**: 85.7 - **CBT-NE (ACC)**: 82.3 - **WikiText2 (PPL)**: 39.14 2. **117M**: - **LAMBADA (PPL)**: 35.13 - **LAMBADA (ACC)**: 45.99 - **CBT-CN (ACC)**: 87.65 - **CBT-NE (ACC)**: 83.4 - **WikiText**

cs224n-2023-lecture11-prompting-rlhf\_p18.png **Diagrams/Graphics:** - There is no diagram or graphic in the image. **Tables:** - There is a table presenting the ROUGE scores for the summarization of the CNN/DailyMail dataset. **Table Data:** `plaintext [ { "Year": "2018 SoTA", "Model": "Bottom-Up Sum", "R-1": 41.22, "R-2": 18.68, "R-L": 38.34 }, { "Year": "2018 SoTA", "Model": "Lede-3", "R-1": 40.38, "R-2": 17.66, "R-L": 36.62 }, { "Year": "Supervised (287K)", "Model": "Seq2Seq + Attn", "R-1": 31.33, "R-2": 11.81, "R-L": 28.83 }, { "Year": "Supervised (287K)", "Model": "GPT-2 TL;DR:", "R-1": 29.34, "R-2": 8.27, "R-L": 26.58 }, { "Year": "Select from article", "Model": "Random-" }`

cs224n-2023-lecture11-prompting-rlhf\_p19.png The image contains one primary graphic: a table that lists names of contributors. Here is the structured format of the data present in the table. **Table: Contributors** - Tom B. Brown - Benjamin Mann - Nick Ryder - Melanie Subbiah Additionally, this section contains text referencing the capabilities of GPT-3, including parameters, and data size, but no specific equations or other graphical data beyond the mentioned table."

cs224n-2023-lecture11-prompting-rlhf\_p2.png The image contains only basic text and does not include any diagrams, graphics, equations, or tables. Here's the text in a structured format: **Reminders** - Project proposals (both custom and final) due a few minutes ago! - We're in the process of assigning mentors to projects and will aim to give feedback on project proposals with

a quick turnaround - A5 due Friday 11:59PM! - We still recommend using Colab for the assignments; in case you run into trouble (e.g. you have exceeded Colab quota), instructions for connecting to a Kaggle notebook have been posted on Ed ````

cs224n-2023-lecture11-prompting-rlhf\_p20.png"The image contains two tables with examples of in-context learning translations. ### Table 1 (Left) | Incorrect | Correct | |-----|-----| | goat | | goat | | sakne | snake | | brid | bird | | fsih | fish | | dcuk | duck | | cmiihp | chimp | ### Table 2 (Right) | English | French | |-----|-----| | thanks | merci | | hello | bonjour | | mint | menthe | | wall | mur | | otter | loutre | | bread | pain | ### Data Format for Database ##### Table 1 - goat => goat - sakne => snake - brid => bird - fsih => fish - dcuk => duck - cmiihp => chimp ##### Table 2 - thanks => merci - hello => bonjour - mint => menthe - wall => mur - otter => loutre - bread => pain"

cs224n-2023-lecture11-prompting-rlhf\_p21.png"### Identification and Summary ##### Diagram: Zero-shot Translation - \*\*Description\*\*: The diagram shows an example of zero-shot learning. It provides a prompt to translate the word "cheese" from English to French. ##### Graph: In-Context Learning on SuperGLUE - \*\*Description\*\*: The graph visualizes the performance of different models on the SuperGLUE benchmark as the number of examples in context increases. The y-axis represents the performance score, and the x-axis represents the number of examples in context (K). - \*\*Data Points\*\*: - Random Guessing: Constant line at 50. - Fine-tuned BERT Large: Constant line at approximately 70. - Fine-tuned BERT++: Constant line at approximately 73-74. - Fine-tuned SOTA: Constant line at approximately 85. - Human: Constant line at approximately 90. - Few-shot GPT-3 175B: Line starting at approximately 60 for K=0, quickly rising to approximately 71 for K=1, and then slightly increasing up to approximately 72-73 as K increases to 32. ### Data Representation \*\*Graph Data (In-Context Learning on SuperGLUE)\*\*: | Model | Performance (Y-axis) | |-----|-----| | Human | 90 | | Fine-tuned SOTA | 85 | | Fine-tuned BERT++ "

cs224n-2023-lecture11-prompting-rlhf\_p22.png"The image includes both a graphic and a diagram. Below is a summary and transcription of each: ### 1. Diagram: Few-Shot Translation Example ##### Description: The left side of the image shows an example of a one-shot learning task where English words are translated to French. ##### Structured Text: `` One-shot Translation Example: 1. Translate English to French: 2. sea otter => loutre de mer 3. cheese => `` ### 2. Graphic: Performance Chart ##### Description: The right side of the image is a chart titled "In-Context Learning on SuperGLUE". It plots the performance of Few-shot GPT-3 175B against the number of examples in context (K). The Y-axis shows the performance score ranging from 40 to 90, while the X-axis depicts the number of examples in context from 0 to 32. ##### Key Points: - Human performance is set around 87. - Few-shot GPT-3 175B performance hovers around the high 60s to low 70s. - Fine-tuned models like BERT++ perform around 70, while Fine-tuned BERT Large and Random Guessing perform at much lower levels. ##### Chart Data: `` Chart Title: In-Context Learning on SuperGLUE Y-Axis: Performance Score (40 to 90) X-Axis: Number of Examples in Context (K) (0, 1, 2"

cs224n-2023-lecture11-prompting-rlhf\_p23.png"### Diagrams, Graphics, Equations, or Tables within the Image: 1. \*\*Table: A translation table from English to French:\*\* - sea otter => loutre

**\*\*Graph: In-Context Learning on SuperGLUE Performance\*\*** - **\*\*Type:\*\*** Line chart with points - **\*\*Axes:\*\*** - X-axis: Number of Examples in Context (K) ranging from 0 to 32. - Y-axis: Performance Score ranging from 40 to 90. - **\*\*Curves:\*\*** - Orange Curve: Performance of Few-shot GPT-3 175B - **\*\*Annotations:\*\*** - Human (at 90) - Fine-tuned SOTA (at around 85) - Fine-tuned BERT Large (at around 70) - Fine-tuned BERT++ (at around 70, slightly above Fine-tuned BERT Large) - Random Guessing (at approximately 50) **### Data for Database:** **\*\*Translation Table:\*\*** `` { "translations": [ { "English": "sea otter", "French": "loutre de mer" }, { "English": "peppermint", "French": "menthe poivrée" }, "

cs224n-2023-lecture11-prompting-rlhf\_p25.png

The image contains a table and a diagram. Here is the structured data extracted from the table and diagram: **Table:**

<b>Prompting</b>	English	French			
peppermint	menthe poivrée	plush giraffe	girafe peluche	cheese	

**Diagram:**

**Traditional Fine-Tuning**

- sea otter => loutre de mer - gradient update
- peppermint => menthe poivrée - gradient update
- cheese => (awaiting translation)

**Summary of Diagram Explanation:** The diagram illustrates the process of traditional fine-tuning in contrast with zero/few-shot prompting. In the fine-tuning method, each example (such as translating "sea otter" to "loutre de mer") is followed by a gradient update, iteratively refining the model. Please note that the diagram graphics show: 1. Sequential illustrations of translation pairs (e.g., sea otter => loutre de mer and peppermint => menthe poivrée). 2. Gradient updates occurring after each translation. The information from the table and the explanations capture how these two distinct methods are used in natural language processing model training."

cs224n-2023-lecture11-prompting-rlhf\_p27.png"#### Data Representation for Database #####  
Diagram Summary The image contains two different approaches for prompting large language models: 1. **\*\*Standard Prompting\*\***: An incorrect and correct response. 2. **\*\*Chain-of-Thought Prompting\*\***: Detailed step-by-step reasoning leading to the correct response. ##### Tables and

Graphics Description - **Table Structure**: - **Title**: Chain-of-thought Prompting (Well-integrated into the graphic) - **Columns**: - Model Input - Model Output - **Rows**: - Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? - A (Standard Prompting): The answer is 11. - A (Chain-of-Thought Prompting): Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11. (Highlighted Text) - Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have? - A (Standard Prompting): The answer is 27. **✗** - A (Chain-of-Thought Prompting): The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20$

cs224n-2023-lecture11-prompting-rlhf\_p28.png The provided image contains a set of three graphs from a PDF page. Here's a structured summary and transcription of the data: **### Data Summary ##### Graphs:** - The graphs depict the solve rate (%) of GSM8K middle school math word problems. - There are three different models compared: - **LaMDA** - **GPT** - **PaLM** Each sub-graph represents: 1. **Standard prompting** (black lines with small unfilled circles) 2. **Chain-of-thought prompting** (blue lines with large unfilled circles) 3. **Prior-supervised best** (orange dashed line at 60%) **##### X-axis:** - Model scale (# of parameters in billions) **##### Y-axis:** - Solve rate (%) **### Data from Graphs: ##### LaMDA:** - Model scale: 0.4B, 8B, 137B - Solve rate (Standard prompting): Data points are low (~0%) - Solve rate (Chain-of-thought prompting): Data points are low (~0%) **##### GPT:** - Model scale: 0.4B, 7B, 175B - Solve rate (Standard prompting): Data points start low (~0%) and slightly increase (below 20%) - Solve rate (Chain-of-thought prompting): Data points start low (~0%) and increase significantly, especially at 175B (reaching around 40%) **##### PaLM:** - Model scale: "

cs224n-2023-lecture11-prompting-rlhf\_p29.png **### Summary of Diagrams and Equations ##### Diagram:** - **Title:** Chain-of-thought prompting - **Categories:** - **Model Input:** - Example questions and answers with reasoning steps highlighted. - **Model Output:** - Final calculation provided. **##### Equations:** 1.  $(5 + 6 = 11)$  2.  $(23 - 20 = 3)$  3.  $(3 + 6 = 9)$  **### Data in Structured Text Form ##### Example 1 - Tennis Balls:** - **Question:** Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? - **Answer:** - **Reasoning:** Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. - **Equation:**  $(5 + 6 = 11)$  - **Result:** The answer is 11. **##### Example 2 - Apples:** - **Question:** The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have? - **Answer:** - **Reasoning:** The cafeteria had 23 apples originally. They used 20 to make lunch. - **Equation Step 1:**  $(23 - 20 = 3)$  - **Reasoning:**

cs224n-2023-lecture11-prompting-rlhf\_p3.png **### Textual Representation of Image Data: ##### Diagram Description and Summary** The image contains a scatter plot titled **"The blessings of scale"**, depicting the estimated computing resources used for AI training runs. The resources are represented using Floating-point operations on a log scale. The scatter plot includes various AI models categorized by their application types: - **Red:** Drawing - **Light Blue:** Language - **Dark Blue:** Vision - **Grey:** Other The x-axis represents the years ranging

from 1950 to 2022, while the y-axis represents the floating-point operations from  $(10^4)$  to  $(10^{24})$ . Annotated AI models include: - **Theseus** - **ADALINE** - **Neocognitron** - **NetTalk** - **NPLM** - **BERT-Large** - **GPT-2** - **GPT-3** - **LaMDA** - **PaLM (540B)** - **DALL-E** The scatter plot shows the significant growth in computing resources required for training larger models over the years. ##### Structured Data Representation: `` { "diagram\_title": "The blessings of scale", "x\_axis": { "label": "Year", "values": ["1950", "60", "70", "80", "90", "2000", "10", "22"] }, "y\_axis": { "label": ""

cs224n-2023-lecture11-prompting-rlhf\_p30.png"The image contains the following equations and explanatory text: 1. **Equation from first question:** -  $(5 + 6 = 11)$  2. **Equation from second question:** -  $(23 - 20 = 3)$  -  $(3 + 6 = 9)$  #### Structured Data Representation: - **Equation:**  $5 + 6 = 11$  - **Explanation:** Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. The sum is 11. - **Answer:** 11 - **Verification:** Correct - **Equation:** -  $23 - 20 = 3$  -  $3 + 6 = 9$  - **Explanation:** The cafeteria had 23 apples originally. They used 20 to make lunch. They bought 6 more apples, so they have 9. - **Answer:** 9 - **Verification:** Correct 3. **Steps and Count from Third Question:** - **Description:** A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. - **Steps:** - There are 16 balls in total. - Half of the balls are golf balls (8 golf balls). - Half of the golf balls are blue (4 blue golf balls). "

cs224n-2023-lecture11-prompting-rlhf\_p31.png"#### Diagrams, Graphics, and Tables The image contains a table comparing various methods' performance on two benchmarks: **MultiArith** and **GSM8K**. The methods being compared are Zero-Shot, Few-Shot, Zero-Shot-CoT, Few-Shot-CoT, and Manual CoT. #### Transcribed Data from the Table ##### Columns: - Method - Samples - MultiArith - GSM8K ##### Rows: - Zero-Shot, -, 17.7, 10.4 - Few-Shot, 2 samples, 33.7, 15.6 - Few-Shot, 8 samples, 33.8, 15.6 - Zero-Shot-CoT, -, 78.7, 40.7 - Few-Shot-CoT, 2 samples, 84.8, 41.3 - Few-Shot-CoT, 4 samples : First (\*1), 89.2, - - Few-Shot-CoT, 4 samples : Second (\*1), 90.5, - - Few-Shot-CoT, 8 samples, -, - - Manual CoT, -, 93.0, 48.7 #### Structured Text Format: ``plaintext { "table": [ { "Method": "Zero-Shot", "Samples": "-", "MultiArith": 17."

cs224n-2023-lecture11-prompting-rlhf\_p32.png"#### Data Extraction from the Table **Table Data:** - Column Headings: - No. - Category - Zero-shot CoT Trigger Prompt - Accuracy - Table Rows: - Row 1: 1, LM-Designed, "Let's work this out in a step by step way to be sure we have the right answer.", 82.0 - Row 2: 2, Human-Designed, "Let's think step by step. (\*1)", 78.7 - Row 3: 3, Human-Designed, "First, (\*2)", 77.3 - Row 4: 4, Human-Designed, "Let's think about this logically.", 74.5 - Row 5: 5, Human-Designed, "Let's solve this problem by splitting it into steps. (\*3)", 72.2 - Row 6: 6, Human-Designed, "Let's be realistic and think step by step.", 70.8 - Row 7: 7, Human-Designed, "Let's think like a detective step by step.", 70.3 - Row 8: 8, Human-Designed, "Let's think", 57.5 - Row 9: 9, Human-Designed, "Before we dive into the answer.", 55.7 - Row 10:"

cs224n-2023-lecture11-prompting-rlhf\_p33.png"The image provided contains several elements including diagrams, graphics, and a snippet of code. Here's a structured summary of each element: #### Diagrams/Graphics: 1. **Image of blue glowing dodecahedron dice:** -



Description: Fantasy concept art of glowing blue dodecahedron die on a wooden table, in a cozy fantasy (workshop), tools on the table, artstation, depth of field, 4k, masterpiece. #### Equations: - No equations are present in the image. #### Tables: - No tables are present in the image. #### Code Snippets: - Snippet of code header from Google LLC: `` 1 # Copyright 2022 Google LLC. 2 # 3 # Licensed under the Apache License, Version 2.0 (the "License"); 4 # you may not use this file except in compliance with the License. 5 # You may obtain a copy of the License at 6 # 7 # http://www.apache.org/licenses/LICENSE-2.0 `` #### Additional Descriptions: 1. \*\*Text Descriptions:\*\* - There is a question and answer section considering a juggler juggling 16 balls, half of them being golf balls and half of the golf balls being blue. - Instruction for translation from English to French with a joke result: "Haha pwned!!" - Mention of "Jailbreaking" LMs (Language Models). -"

cs224n-2023-lecture11-prompting-rlhf\_p34.png"The image features the following elements: 1. \*\*Diagram:\*\* - A screenshot of a Wikipedia article excerpt titled "Prompt engineering." - A highlighted text snippet within the Wikipedia article. 2. \*\*Keywords noted:\*\* - Artificial intelligence - Natural Language Processing (NLP) Based on this observation, the structured information is: ``json { "type": "diagram", "source": "Wikipedia", "title": "Prompt engineering", "keywords": [ "artificial intelligence", "natural language processing", "prompt engineering" ] } `` Additionally, there is an advertisement or notice at the bottom: 3. \*\*Job Posting:\*\* - "Prompt Engineer and Librarian" - Location: San Francisco, CA - Type: Product / Full-Time / Hybrid This information in a structured format: ``json { "type": "job\_posting", "position": "Prompt Engineer and Librarian", "location": "San Francisco, CA", "type": [ "Product", "Full-Time", "Hybrid" ] } `` No tables, equations, or additional graphics were detected within the image."

cs224n-2023-lecture11-prompting-rlhf\_p35.png"The provided image does not contain any diagrams, graphics, equations, or tables. It appears to be a slide presenting a lecture plan. Here is the structured information from the slide: ``plaintext Lecture Plan: From Language Models to Assistants 1. Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning + No finetuning needed, prompt engineering (e.g. CoT) can improve performance – Limits to what you can fit in context – Complex tasks will probably need gradient steps 2. Instruction finetuning 3. Reinforcement Learning from Human Feedback (RLHF) 4. What's next? `` This structured format should be easy to pass directly into a database."

cs224n-2023-lecture11-prompting-rlhf\_p36.png"The image provided contains a slide titled "Lecture Plan: From Language Models to Assistants" and is structured as a list. Here is a summary and transcription of the contents, followed by a structured format: #### Structured Data: \*\*Title:\*\* Lecture Plan: From Language Models to Assistants \*\*Content:\*\* 1. Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning - Advantages: + No finetuning needed, prompt engineering (e.g., CoT) can improve performance - Limitations: - Limits to what you can fit in context - Complex tasks will probably need gradient steps 2. Instruction finetuning 3. Reinforcement Learning from Human Feedback (RLHF) 4. What's next? #### Summary of Graphics and Diagrams: There are no diagrams, graphics, equations, or tables in this slide. This data can be directly passed into a database with fields for titles, content, and notes on advantages and limitations where appropriate."

cs224n-2023-lecture11-prompting-rlhf\_p37.png\*\*\*Diagrams, Graphics, Equations, Tables:\*\* - The image contains a table. \*\*Table Data:\*\* `` | PROMPT | Explain the moon landing to a 6 year old in a few sentences. | |-----|-----|-----|-----|-----|-----|-----|-----|-----|-----| | COMPLETION | GPT-3 | | Explain the theory of gravity to a 6 year old. | | Explain the theory of relativity to a 6 year old in a few sentences. | | Explain the big bang theory to a 6 year old. | | Explain evolution to a 6 year old. | | `` \*\*Summary:\*\* - The table illustrates a mismatch between the given prompt and the completion generated by a language model (GPT-3). It highlights how the language model completes the prompt with sentences that are irrelevant to the asked prompt."

cs224n-2023-lecture11-prompting-rlhf\_p38.png"The image includes a table with two columns and two rows. \*\*Table Representation in Structured Text:\*\* - \*\*Column Headers:\*\* - Prompt - Completion - \*\*Data:\*\* - Row 1: - Prompt: Explain the moon landing to a 6 year old in a few sentences. - Completion: - \*\*Source:\*\* Human - \*\*Text:\*\* A giant rocket ship blasted off from Earth carrying astronauts to the moon. The astronauts landed their spaceship on the moon and walked around exploring the lunar surface. Then they returned safely back to Earth, bringing home moon rocks to show everyone."

cs224n-2023-lecture11-prompting-rlhf\_p39.png\*\*\*Diagrams Summary:\*\* The image contains two similar diagrams illustrating the process of pretraining and finetuning in the context of Natural Language Processing (NLP): 1. \*\*Diagram on the Left (Step 1: Pretrain on language modeling):\*\* - Title: Step 1: Pretrain (on language modeling) - Purpose: Learning general things from lots of text. - Components: - Decoder labeled "Decoder (Transformer, LSTM, ++)". - Input tokens: "Iroh", "goes", "to", "make", "tasty", "tea". - Output tokens (arrows pointing upwards): "goes", "to", "make", "tasty", "tea", "END". 2. \*\*Diagram on the Right (Step 2: Finetune on your task):\*\* - Title: Step 2: Finetune (on your task) - Purpose: Adapting to the specific task with not many labels. - Components: - Decoder labeled "Decoder (Transformer, LSTM, ++)". - Input tokens: "...", "the", "movie", "was", "..." (incompletely shown, assumed to be a continuation of a sentence). - Output tokens (arrows pointing upwards): "..." (indicating context continuation). There are smiley and sad face emojis at the end of the output tokens in the second diagram, indicating different possible sentiments"

cs224n-2023-lecture11-prompting-rlhf\_p4.png"The image is a visual representation of the number of tokens seen during training for different entities/models over time. It features a graphic with various colored circles representing each entity and the number of tokens they have seen during training. Below is the structured data extracted from the image: #### Structured Data: ##### Entities and Tokens Seen During Training 1. \*\*13 y.o. Human\*\* - Tokens: <100 Million - Representation: Text 2. \*\*BERT (2018)\*\* - Tokens: 3 Billion - Representation: Small green circle 3. \*\*RoBERTa (2019)\*\* - Tokens: 30 Billion - Representation: Medium red circle 4. \*\*GPT-3 (2020)\*\* - Tokens: 200 Billion - Representation: Large blue circle 5. \*\*Chinchilla (2022)\*\* - Tokens: 1.4 Trillion - Representation: Extra-large orange circle ##### Additional Information - URL: [https://babylm.github.io/](https://babylm.github.io/) #### Summary The image visually compares the amount of training data used for different models over the years, with 13 y.o. Human being the baseline with less than 100 million tokens. Subsequent models—

BERT (2018), RoBERTa (2019), GPT-3 (2020), and Chinchilla (2022)—have seen exponentially

cs224n-2023-lecture11-prompting-rlhf\_p40.png\*\*\*Diagrams and Graphics Summary:\*\* The image contains two main diagrams that illustrate the process of scaling up finetuning for NLP applications: 1. \*\*Diagram 1 (Step 1: Pretrain)\*\*: - Title: "Step 1: Pretrain (on language modeling)" - Description: This step involves learning general language representations from large amounts of text. The text example includes the sentence: "Iroh goes to make tasty tea," which is processed by a decoder (potentially a Transformer, LSTM, or other models). 2. \*\*Diagram 2 (Step 2: Finetune)\*\*: - Title: "Step 2: Finetune (on many tasks)" - Description: This step focuses on adapting the pretrained model to specific tasks with relatively fewer labels. The illustrative text example seems partly obscured but involves finetuning using another decoder (Transformer, LSTM, or other models) to adapt to more specific tasks, indicated by text such as "... the movie was ...". \*\*Equations:\*\* - No equations are present. \*\*Tables:\*\* - No tables are present. \*\*Simple Structured Data:\*\* ``plaintext Diagrams: 1. Title: Step 1: Pretrain (on language modeling) Description: Learn general things from large amounts of text. Example Text: Iroh goes to make tasty tea Model: Decoder (Transformer, LSTM, ++) 2. Title: Step 2"

cs224n-2023-lecture11-prompting-rlhf\_p41.png"The image contains various elements including diagrams and equations. Here is the structured data from the diagrams: #### Diagrams Representation 1. \*\*Instruction Pairings and Finetuning:\*\* - \*\*Instruction 1:\*\* - \*\*Instruction:\*\* Please answer the following question. What is the boiling point of Nitrogen? - \*\*Output:\*\* -320.4F - \*\*Instruction 2:\*\* - \*\*Instruction:\*\* Answer the following question by reasoning step-by-step. The cafeteria had 23 apples. If they used 20 for lunch, and bought 6 more, how many apples do they have? - \*\*Output:\*\* The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . 2. \*\*Evaluation on Unseen Tasks:\*\* - \*\*Instruction 3:\*\* - \*\*Instruction:\*\* Q: Can Geoffrey Hinton have a conversation with George Washington? Give the rationale before answering. - \*\*Output:\*\* Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no". #### Equations - Equation in Instruction 2 Output: - Apple Calculation: -  $23 - 20 = 3$  - "

cs224n-2023-lecture11-prompting-rlhf\_p42.png"#### Diagram/Graphic Summary ##### Diagram: Task Categories in Super-NaturalInstructions Dataset - \*\*Large Circles:\*\* - Translation - Question Answering - \*\*Medium Circles:\*\* - Sentiment Analysis - Information Extraction - Textual Entailment - Text Categorization - Text Matching - Question Generation - Named Entity Recognition - Commonsense Classification - Program Execution - Language Identification - Summarization - Toxic Language Detection - Text Simplification - Keyword Tagging - Speaker Identification - Dialogue Generation - Sentence Composition - Sentence Perturbation - Cause Effect Classification - Word Analogy - Linguistic Probing - Grammar Error Detection - Sentence Expansion - \*\*Small to Very Small Circles:\*\* - Spelling Correction - Style Transfer - Part-of-Speech (POS) Tagging - Entity Relation Classification - Fact Verification - Word Relation Classification - Negotiation Strategy Detection - Overlap Extraction - Poem Generation - Spam

Classification - Answerability Classification - Dialogue State Tracking - Preposition Prediction - Dialogue Act Recognition - Mathematics - Dialogue Discourse Identification - Speaker Relation Classification - Discourse Relation Classification - Iron"

cs224n-2023-lecture11-prompting-rlhf\_p43.png"The image contains a bar chart comparing the performance of different language models on various college-level and high school-level subjects. The chart includes three different data series: GPT-3 (blue), UnifiedQA (green), and Random (red). Here is the structured data extracted from the chart: `` Subject | GPT-3 | UnifiedQA | Random ----- Abstract Algebra | value | value | value Anatomy | value | value | value Astronomy | value | value | value Business Ethics | value | value | value Clinical Knowledge | value | value | value College Biology | value | value | value College Chemistry | value | value | value College Comp Sci | value | value | value College Mathematics | value | value | value College Medicine | value | value | value College Physics | value | value | value Computer Security | value | value | value Conceptual Physics | value | value | value Econometrics | value | value | value Electrical Engineering | value | value | value Elementary Mathematics | value | value | value Formal Logic | value | value | value Global Facts | value | value | value High School Biology | value "


cs224n-2023-lecture11-prompting-rlhf\_p44.png"Below is a structured representation of the content present in the provided image. It includes descriptions of the diagrams and transcriptions of any non-basic text elements like equations or tables. **\*\*Diagrams/Graphics:\*\*** 1. **\*\*Word Cloud:\*\*** - Description: The word cloud is a visual representation of various tasks included in the BIG-Bench dataset. The words vary in size, indicating their frequency or significance. Larger words include "common sense", "free response", "programmatic", "contextual question-answering", "logical reasoning", "reading comprehension". - Keywords: common sense, free response, programmatic, logical reasoning, reading comprehension, context-free question answering, contextual question-answering, mathematics, analogical reasoning, casual reasoning, non-language tasks, visual reasoning, multilingual, gender bias, emotion, arithmetic, numerical response, paraphrase, repeated interaction, non-English, creativity, human-like behavior, physical reasoning, theory of mind, computer code, social bias, machine translation, context length, non-linguistic task, social bias, meta-reasoning, machine translation, temporal reasoning. **\*\*Tables:\*\*** None identified. **\*\*Equations:\*\*** None identified. --- **\*\*Additional Text:\*\*** 1. **\*\*Link:\*\*** - URL: `https://github.com/google/BIG-bench/blob/main/bigbench/benchmark\_tasks/README.md` 2. **\*\*Alphabetic Author List:\*\*** - List: Authors' names are listed in alphabetical order. Here are"

cs224n-2023-lecture11-prompting-rlhf\_p45.png"The image contains two main components: a word cloud diagram and ASCII art. **#### Graphics and Diagrams:** 1. **\*\*Word Cloud Diagram:\*\*** - Title: BIG-Bench - Description: A word cloud representing topics covered by 200+ tasks. - Words included (in varied sizes, indicating frequency or importance): - common sense, programmatic, free response, logical reasoning, reading comprehension, context-free question answering, mathematics, analogical reasoning, context length, paraphrase, emotional intelligence, implicit reasoning, human-like behavior, arithmetic, computer code, creativity, non-English, translation, social reasoning, causal reasoning, visual reasoning, domain specific, numerical response, and many others. - Link Included: [https://github.com/google/BIG-

bench/blob/main/bigbench/benchmark\_tasks/README.md](https://github.com/google/BIG-bench/blob/main/bigbench/benchmark\_tasks/README.md) 2. **Kanji ASCII Art**: **Title**: Kanji ASCII Art to Meaning - **Description**: A subtask that converts various kanji into ASCII art and has the language model guess their meaning from the ASCII art. - **ASCII Art Sample**:  
`` .....#..... .....#..... ##### .....#.#.#.... ..#.#.#.#... .....#.....# .....#.....#  
#.#.....#.#.....#....."

cs224n-2023-lecture11-prompting-rlhf\_p46.png"### Table Data ``plaintext | Params | Model | BIG-bench + MMLU avg (normalized) | |-----|-----|-----| | 80M | T5-Small | -9.2 | | 80M | Flan-T5-Small | -3.1 (+6.1) | | 250M | T5-Base | -5.1 | | 250M | Flan-T5-Base | 6.5 (+11.6) | | 780M | T5-Large | -5.0 | | 780M | Flan-T5-Large | 13.8 (+18.8) | | 3B | T5-XL | -4.1 | | 3B | Flan-T5-XL | 19.1 (+23.2) | | 11B | T5-XXL | -2.9 | | 11B | Flan-T5-XXL | 23.7 (+26.6) | `` ### Diagram Description The diagram on the page is a neural network structure. It appears to show a sequence of connections between different layers of nodes, indicating a flow from one layer to a subsequent series of layers. The first layer is labeled"

cs224n-2023-lecture11-prompting-rlhf\_p47.png"### Diagrams and Graphics 1. **Diagram**: The image includes a comparison of responses before and after instruction fine-tuning. There are two sections, each enclosed in a rounded rectangle: - **Left Section (Model input (Disambiguation QA))**: This section concerns the task of disambiguation and describes a model input example where the model needs to explain the antecedent of the pronoun. - **Right Section (Before instruction finetuning)**: This section illustrates several examples of the model's responses to the same input before finetuning, indicating that the model didn't answer the question. ### Tables: - There is no explicit table in the image, but an implied comparison table of responses is present. ### Equations: - There are no equations present in the image. ### Structured Data Representation `` { "Model Input (Disambiguation QA)": { "Question": "In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.", "Sentence": "The reporter and the chef will discuss their favorite dishes.", "Options": [ "They will discuss the reporter's favorite dishes", "They will discuss the chef's favorite dishes", "Ambiguous" ], "Answer": "Let's think step by step." }, "Before Instruction Finetuning": [ "The reporter and the chef will discuss their favorite dishes.", "

cs224n-2023-lecture11-prompting-rlhf\_p48.png"The image contains a diagram with two sections, labeled "Model input (Disambiguation QA)" and "After instruction finetuning". Below is a structured representation of the content: **Model input (Disambiguation QA)**: **Q**: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous. - **Sentence**: The reporter and the chef will discuss their favorite dishes. - **Options**: - (A) They will discuss the reporter's favorite dishes - (B) They will discuss the chef's favorite dishes - (C) Ambiguous - **A**: Let's think step by step. **After instruction finetuning**: - The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C).  Additionally, there is a recommendation text at the bottom of the image: - Highly recommend trying FLAN-T5 out to get a sense of its capabilities: [https://huggingface.co/google/flan-t5-

xxl](https://huggingface.co/google/flan-t5-xxl) - Citation: [Chung et al., 2022] This information can be passed to the database in a structured format."

cs224n-2023-lecture11-prompting-rlhf\_p49.png"The image contains a lecture plan with numbered points and corresponding descriptions. There are no diagrams, graphics, equations, or tables in the image. Below is the structured representation of the data from the lecture plan: ``  
Lecture Plan: From Language Models to Assistants 1. Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning - Positives: + No finetuning needed, prompt engineering (e.g. CoT) can improve performance - Negatives: - Limits to what you can fit in context - Complex tasks will probably need gradient steps 2. Instruction finetuning - Positives: + Simple and straightforward, generalize to unseen tasks - Negatives: - ? - ? 3. Reinforcement Learning from Human Feedback (RLHF) (No additional details provided) 4. What's next? (No additional details provided) ``"

cs224n-2023-lecture11-prompting-rlhf\_p5.png"The image contains a list of sentences, but no diagrams, graphics, tables, or equations. Here's the structured form of the list: `` [ { "sentence": "Stanford University is located in \_\_\_\_\_, California.", "category": "Trivia" }, { "sentence": "I put \_\_\_\_ fork down on the table.", "category": "Syntax" }, { "sentence": "The woman walked across the street, checking for traffic over \_\_\_\_ shoulder.", "category": "Coreference" }, { "sentence": "I went to the ocean to see the fish, turtles, seals, and \_\_\_\_.", "category": "Lexical semantics/topic" }, { "sentence": "Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was \_\_\_\_.", "category": "Sentiment" }, { "sentence": "Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the \_\_\_\_.", "category": "Some reasoning – this is harder" }, { "sentence": "I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, \_\_\_\_.", "category": "Some basic arithmetic; they don't learn the" }

cs224n-2023-lecture11-prompting-rlhf\_p50.png"Here is the requested structured data extracted from the image: **\*\*Diagrams and Graphics Description:\*\*** 1. **\*\*Graphic Description:\*\*** - Title/Subheading: LM Output - Content: - A Language Model (LM) with an output sequence "Avatar is a fantasy TV show" is compared to the expected sequence "Avatar is a fantasy TV show END", highlighting errors with the words "adventure" and "musical" crossed out. - Arrows point from each output token to the LM block. **\*\*Note for Database Entry:\*\*** - **\*\*Graphic\_Type:\*\*** LM Output Comparison - **\*\*Description:\*\*** A graphical representation of discrepancies in a Language Model's output. The LM output sequence "Avatar is a fantasy TV show" is compared against the expected sequence "Avatar is a fantasy TV show END", indicating incorrect predictions with "adventure" and "musical" being crossed out. There are no tables or equations transcribed in the image. The structured content should be stored with the structure above for easy referencing and utilization in the database."

cs224n-2023-lecture11-prompting-rlhf\_p51.png"There are no diagrams, graphics, or equations in the image. The image contains a structured list representing the lecture plan from a PDF page. Here's the structured text representation: ``json { "lecture\_plan": { "title": "Lecture Plan: From Language Models to Assistants", "sections": [ { "number": 1, "title": "Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning", "pros": [ "No finetuning needed, prompt engineering (e.g. CoT) can improve performance" ], "cons": [ "Limits to what you can fit in context", "Complex tasks

will probably need gradient steps" ] } , { "number": 2, "title": "Instruction finetuning", "pros": [ "Simple and straightforward, generalize to unseen tasks" ], "cons": [ "Collecting demonstrations for so many tasks is expensive", "Mismatch between LM objective and human preferences" ] } , { "number": 3, "title": "Reinforcement Learning from Human Feedback (RLHF)" } , { "number": 4, "title": "What's next?" } ] } , "page\_number": 51 } ````

cs224n-2023-lecture11-prompting-rlhf\_p52.png "The image appears to be a slide from a presentation titled "Lecture Plan: From Language Models to Assistants." It contains a list of points rather than traditional diagrams, graphics, or tabulated data. Below is a structured form of the content provided: **Lecture Plan: From Language Models to Assistants** 1. **Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning** - **Advantages:** - No finetuning needed, prompt engineering (e.g., CoT) can improve performance - **Disadvantages:** - Limits to what you can fit in context - Complex tasks will probably need gradient steps 2. **Instruction Finetuning** - **Advantages:** - Simple and straightforward, generalize to unseen tasks - **Disadvantages:** - Collecting demonstrations for so many tasks is expensive - Mismatch between LM objective and human preferences 3. **Reinforcement Learning from Human Feedback (RLHF)** 4. **What's next?** **Page number:** 52 This structured format captures the hierarchical and categorical structure of the slide's content for direct entry into a database."

cs224n-2023-lecture11-prompting-rlhf\_p53.png "The image contains a few structured elements, including equations and a graphical comparison between two text samples. Here is a summary and transcription of the elements: **Equations** 1.  $R(s) \in \mathbb{R}$  2.  $R(s_1) = 8.0$  3.  $R(s_2) = 1.2$  4.  $\mathbb{E}_{\hat{s} \sim p_{\theta}(s)}[R(\hat{s})]$  **Table Representation of Text Samples**

Sample	Text	Reward Value
s1	SAN FRANCISCO, California (CNN) -- A magnitude 4.2 earthquake shook the San Francisco ... overturn unstable objects.	8.0
s2	An earthquake hit San Francisco. There was minor property damage, but no injuries.	1.2

The graphical comparison visually juxtaposes two sample text summaries with corresponding reward values, demonstrating a higher reward for a more detailed and informative summary. The equations provide mathematical expressions related to these reward values and optimization of language models."

FRANCISCO, California (CNN) -- A magnitude 4.2 earthquake shook the San Francisco ... overturn unstable objects. | 8.0 | s2 | An earthquake hit San Francisco. There was minor property damage, but no injuries. | 1.2 | --- The graphical comparison visually juxtaposes two sample text summaries with corresponding reward values, demonstrating a higher reward for a more detailed and informative summary. The equations provide mathematical expressions related to these reward values and optimization of language models."

cs224n-2023-lecture11-prompting-rlhf\_p54.png "The image contains the following key elements: 1. **Graphics/Diagrams:** - The right side of the image features a cover from the "Nature" journal with the title "LEARNING CURVE." It shows colorful abstract shapes that resemble game characters or elements, indicating the focus on AI and reinforcement learning. - At the bottom right corner, there is an "AlphaGo" logo, indicating the mention of a specific AI achievement. 2. **Tables:** - There are no tables present in the image. 3. **Equations:** - There are no equations present in the image. 4. **Textual Summary:** - The slide titled "Reinforcement learning to the rescue" summarizes the evolution of reinforcement learning (RL) in the context of language models (LMs) and neural networks. - It indicates key milestones, including: - Initial studies by Williams (1992) and Sutton and Barto (1998). - A resurgence in 2013 with deep learning and game-playing, as per Mnih et al. (2013). - Newer interest in applying RL to modern LMs reflected in works by Ziegler et al. (2019), Stiennon et al. (2020),

and Ouyang et al. (2022). - It also emphasizes the challenges in RL with LMs and mentions advancements in algorithms, highlighting PPO by Schulman et al. (2017). "

cs224n-2023-lecture11-prompting-rlhf\_p55.png\*\*\*Equations:\*\* 1.  $\mathbb{E}_{\theta} [R(\theta)]$  2.  $\theta_{t+1} := \theta_t + \alpha \nabla_{\theta} \mathbb{E}_{\theta} [R(\theta)]$  \*\*Diagrams/Graphics

Description:\*\* - The diagram contains two equations with annotations asking specific questions: - "How do we estimate this expectation??" - "What if our reward function is non-differentiable??" This image seems to highlight the process of optimizing parameters using gradient ascent in the context of reinforcement learning (RL), specifically mentioning policy gradient methods such as REINFORCE. \*\*Structured Data for Database:\*\* ``json

```
{
 "equations": [
 "E_{\theta} [R(\theta)]",
 "\theta_{t+1} := \theta_t + \alpha \nabla_{\theta} E_{\theta} [R(\theta)]"
],
 "annotations": [
 "How do we estimate this expectation??",
 "What if our reward function is non-differentiable??"
]
}
```

cs224n-2023-lecture11-prompting-rlhf\_p56.png#### Diagrams and Equations in the Image

\*\*Equations:\*\* 1.  $\nabla_{\theta} \mathbb{E}_{\theta} [R(\theta)] = \mathbb{E}_{\theta} [R(\theta) \nabla_{\theta} \log p_{\theta}(s)]$  2.  $\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s)$  (Using the chain rule) 3.  $\nabla_{\theta} \log p_{\theta}(s) = \frac{1}{p_{\theta}(s)} \nabla_{\theta} p_{\theta}(s)$  4.  $\sum_{s_i} R(s_i) \nabla_{\theta} \log p_{\theta}(s_i) = \sum_{s_i} p_{\theta}(s_i) R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$  5.  $\mathbb{E}_{\theta} [R(\theta) \nabla_{\theta} \log p_{\theta}(\hat{s})]$  \*\*Structured Data Representation of Equations:\*\* 1. `` { "equation"

cs224n-2023-lecture11-prompting-rlhf\_p57.png"The image contains the following elements: ### Equations 1. The first equation:  $\nabla_{\theta} \mathbb{E}_{\theta} [R(\theta)] = \mathbb{E}_{\theta} [R(\theta) \nabla_{\theta} \log p_{\theta}(\hat{s})]$  2. The update rule equation:  $\theta_{t+1} := \theta_t + \alpha \frac{1}{m} \sum_{i=1}^m R(s_i) \nabla_{\theta} \log p_{\theta}(s_i)$  ### Graphics and Diagrams 1. \*\*Green Text:\*\* - "If  $\nabla_{\theta}$  is +++": Directed towards the update rule with an arrow indicating "Take gradient steps to maximize  $\log p_{\theta}(s_i)$ ". 2. \*\*Red Text:\*\* - "If  $\nabla_{\theta}$  is ---": Directed towards the update rule with an arrow indicating "Take steps to minimize  $\log p_{\theta}(s_i)$ ". ### Summary of the Content -"

cs224n-2023-lecture11-prompting-rlhf\_p58.png"The image contains a diagram with two human icons, two text descriptions next to the icons, two reward function equations, and arrow connections between the text and icons. The image is described and summarized below. ### Diagram Summary - \*\*Text Description 1 (under the left icon):\*\* "An earthquake hit San Francisco. There was minor property damage, but no injuries." - \*\*Text Description 2 (under the right icon):\*\* "The Bay Area has good weather but is prone to earthquakes and wildfires." - \*\*Icons:\*\* - Two human figures (one under each text description) - \*\*Equations:\*\* - Under the left human icon:  $(s_1, R(s_1) = 8.0)$  - Under the right human icon:  $(s_2, R(s_2) = 1.2)$  ### Additional Content - A decorative graphic of a stack of money next to each human icon. ###



Structured Text Representation ``json { "Diagram": [ { "Text": "An earthquake hit San Francisco. There was minor property damage, but no injuries.", "Icon": "Human", "Equation": { "s": "s1", "R(s)": 8.0 }, "Graphics": "Stack of money" }, { "Text": "The Bay Area has good weather but is prone to earthquakes and wildfires.", "Icon": "Human", "Equation": { "s": "s2", "R(s)": 6.6 }, "Graphics": "Stack of money" }, { "Text": "The Bay Area has good weather but is prone to earthquakes and wildfires.", "Icon": "Human", "Equation": { "s": "s3", "R(s)": 3.2 }, "Graphics": "Stack of money" } ] }

cs224n-2023-lecture11-prompting-rlhf\_p59.png\*\*\*Diagrams and Graphics Summary:\*\* 1. \*\*Graphic Description:\*\* - A block of text describes the event: "A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage." - There is a question presented:  $R(s_3) = 4.1?$  6.6? 3.2? \*\*Equations Transcription:\*\* -  $R(s_3) = 4.1?$  6.6? 3.2? \*\*Structured Data:\*\* - Event Description: "A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage." - Equation:  $R(s_3)$  with possible values 4.1, 6.6, or 3.2 \*\*Note:\*\* No tables were identified in the provided page."

cs224n-2023-lecture11-prompting-rlhf\_p6.pngThe image contains two text boxes with specific content. There are no tables, equations, or standard diagrams identified. Text Box 1: `` Pat watches a demonstration of a bowling ball and a leaf being dropped at the same time in a vacuum chamber. Pat, who is a physicist, predicts that the bowling ball and the leaf will fall at the same rate. `` Text Box 2: `` ...Pat, who has never seen this demonstration before, predicts that the bowling ball will fall to the ground first. This is incorrect. In a vacuum chamber, there is no air resistance. `` There are no tables, diagrams, or equations to transcribe. The data has been captured as text from the two text boxes."

cs224n-2023-lecture11-prompting-rlhf\_p60.png#### Diagrams and Graphics Description: 1. \*\*Diagram of Reward Model:\*\* - The diagram illustrates a "Reward Model" labeled  $RM_{\phi}$  with states  $(s_1)$ ,  $(s_2)$ , and  $(s_3)$ . - Each state has a description: -  $(s_1)$ : "An earthquake hit San Francisco. There was minor property damage, but no injuries." -  $(s_3)$ : "A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage." -  $(s_2)$ : "The Bay Area has good weather but is prone to earthquakes and wildfires." - States are compared pairwise to determine preferences in the context of human judgment. 2. \*\*Implementation Details:\*\* - The bottom of the diagram shows arrows pointing towards various events such as "The Bay Area", "wildfires", etc. #### Equations: - Bradley-Terry paired comparison model:  $J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l) \sim D} [\log \sigma(RM_{\phi}(s^w) - RM_{\phi}(s^l))]$  where  $(s^w)$  is a winning sample and  $(s^l)$  is a losing sample. The score of  $(s^w)$  should be higher than that of  $(s^l)$

cs224n-2023-lecture11-prompting-rlhf\_p61.png\*\*\*Diagram Summary:\*\* - The diagram is a line graph that evaluates Reward Models (RMs) on predicting the outcome of held-out human judgments. - The x-axis is labeled "Model size" and the y-axis is labeled "Validation accuracy." - Various data sizes (8k, 16k, 32k, 64k) are represented by different shades of blue lines. - Two benchmark lines are shown: "Ensemble of humans" at a validation accuracy of 0.85, and "Human baseline" at a validation accuracy of approximately 0.78. \*\*Structured Data Representation:\*\* | Data Size |  $10^8$  Model Size |  $10^9$  Model Size |  $10^{10}$  Model Size | |-----|-----|-----| |-----|-----|-----| | 8k | 0.615 | 0.67 | 0.715 | | 16k | 0.625 | 0.69 | 0.735 | | 32k | 0.635 | 0.705 | 0.755 | | 64k | 0.645 | 0.72 | 0.775 | \*\*Additional Information:\*\* - "Large enough RM"

trained on enough data approaching single human perf" annotation highlights that a large enough RM trained on sufficient data approaches single human performance. - The source of the data"

cs224n-2023-lecture11-prompting-rlhf\_p62.png"The image contains the following types of content: #### Diagrams and Graphics: - The image includes a highlighted segment in the equation  $R(s)$ , specifically the expression  $\frac{p_{\theta}^{RL}(s)}{p^{PT}(s)}$ . This part is accompanied by a textual annotation: "Pay a price when  $p_{\theta}^{RL}(s) > p^{PT}(s)$ ". #### Equations: -  $R(s) = RM_{\phi}(s) - \beta \log \left( \frac{p_{\theta}^{RL}(s)}{p^{PT}(s)} \right)$  #### Tables: - No tables are presented in the image. #### Summary: The page discusses the process of Reinforcement Learning with Human Feedback (RLHF). It describes the prerequisites and steps to perform RLHF, including the initialization of a model  $p_{\theta}^{RL}(s)$  and optimization using the specified reward equation. The annotation clarifies when a penalty is applied based on the divergence between  $p_{\theta}^{RL}(s)$  and  $p^{PT}(s)$ . #### Structured Data for Database: ``json { "equations": [ "R(s) = RM\_{\phi}(s) - \beta \log \left( \frac{p\_{\theta}^{RL}(s)}{p^{PT}(s)} \right)" ], "diagrams"

cs224n-2023-lecture11-prompting-rlhf\_p63.png"The image is a graph showing the fraction preferred to reference (y-axis) versus model size (x-axis in billions of parameters). The graph depicts the performance of different learning approaches in comparison to reference summaries. Below are the results summarized from the graph. \*\*Chart Data:\*\* 1. \*\*Human feedback (orange)  $p^{RL}(s)$ :\*\* - Model size 1.3B: around 0.6 - Model size 2.7B: around 0.62 - Model size 6.7B: around 0.64 - Model size 12.9B: around 0.67 2. \*\*Supervised learning (green)  $p^{1FT}(s)$ :\*\* - Model size 1.3B: around 0.42 - Model size 2.7B: around 0.45 - Model size 6.7B: around 0.46 - Model size 12.9B: around 0.47 3. \*\*Pretrain only (blue)  $p^{PT}(s)$ :\*\* - Model size 1.3B: around 0.28 - Model size 2.7B: around 0.34 - Model size 6.7B: around 0.33 - Model size 12.9B: around "

cs224n-2023-lecture11-prompting-rlhf\_p64.png"The provided image contains a flowchart diagram detailing the process for scaling up reinforcement learning from human feedback (RLHF) to tens of thousands of tasks. Here is a structured representation of the diagrams and descriptions: #### Flowchart Summary: ##### Step 1: Collect demonstration data, and train a supervised policy. - \*\*Task:\*\* A prompt is sampled from the prompt dataset. - \*\*Example Prompt:\*\* "Explain the moon landing to a 6-year-old" - \*\*Process:\*\* - A labeler demonstrates the desired output behavior. - \*\*Example Output:\*\* "Some people went to the moon..." - This data is used to fine-tune GPT-3 with supervised learning. ##### Step 2: Collect comparison data, and train a reward model. - \*\*Task:\*\* A prompt and several model outputs are sampled. - \*\*Example Prompt:\*\* "Explain the moon landing to a 6-year-old" - \*\*Example Outputs:\*\* - A: Explain gravity... - B: Explain war... - C: Moon is a natural satellite of... - D: People went to the moon... - \*\*Process:\*\* - A labeler ranks the outputs from best to worst. - \*\*Example Ranking:\*\* - D > C > A > B - This data is used to train the reward model. ##### Step 3: Optimize a policy against the reward model using reinforcement learning. - \*\*Task:\*\*"

cs224n-2023-lecture11-prompting-rlhf\_p65.png"## Table Data \*\*Use-case:\*\* Brainstorming \*\*Prompt:\*\* List five ideas for how to regain enthusiasm for my career \*\*Use-case:\*\*

Generation **Prompt**: Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home **Summary** The image contains a section from a document about InstructGPT and the collection of tasks from labelers. It includes three task types: Plain, Few-shot, and User-based, followed by a table listing use-cases and corresponding prompts: **Task Types**: 1. **Plain**: Arbitrary tasks ensuring diversity. 2. **Few-shot**: Instructions with multiple query/response pairs. 3. **User-based**: Prompts based on user cases from OpenAI API waitlist applications. **Table Summary**: 1. **Brainstorming**: Listing ideas for career enthusiasm. 2. **Generation**: Writing a short story about a bear at the beach."

cs224n-2023-lecture11-prompting-rlhf\_p66.png"**Table**:  
Prompt	Explain the moon landing to a 6-year-old in a few sentences.	
Completion	GPT-3	Explain the theory of gravity to a 6-year-old.
InstructGPT	Explain the theory of relativity to a 6-year-old in a few sentences.	
Explain the big bang theory to a 6-year-old.		
Explain evolution to a 6-year-old.		
InstructGPT	People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.	
**Summary**: The graphic comprises a prompt and different completions by GPT-3 and InstructGPT. The prompt is "Explain the moon landing to a 6-year-old in a few sentences." GPT-3 provides explanations for various scientific theories, while InstructGPT gives a simplified explanation of the moon landing and the return of pictures to Earth."

cs224n-2023-lecture11-prompting-rlhf\_p67.png"Based on the provided image, the primary content comprises a structured comparison in table format and some specific instructions and examples. Here is a transcription of the table data as structured text: **Table Content**: **Prompt**: Write a short poem about a wise frog. **Completion**: 1. GPT-3: - Write a short story in which a character has two different names. - Write a short story in which you try to get something back that you have lost. - Write a short story in which a character has a bad dream. 2. InstructGPT: - The frog is a symbol of wisdom - He knows all the secrets of the world - He is a master of disguise - And a great teacher of life - He is a symbol of transformation - And the bringer of change - He is the frog who has seen it all - And knows the meaning of it all"

cs224n-2023-lecture11-prompting-rlhf\_p68.png"This image contains several elements: **Diagrams/Graphics**: 1. A rectangular graphic with the text: "ChatGPT: Optimizing Language Models for Dialogue" on a dark green background. 2. A note under the graphic on the left side: "Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge...". **Equations**: There are no equations present in the image. **Tables**: There are no tables present in the image. **Description and Summary of Graphics and Diagrams**: **Graphic 1**: - Title Graphic: "ChatGPT: Optimizing Language Models for Dialogue". - Note: Mention of secrecy by OpenAI and similar companies regarding details of ChatGPT training. **Text on the right side**: **Header**: Methods **Content**: Description of the methods used to train the model via RLHF, with similarities to InstructGPT but with different data collection. Describes supervised fine-tuning with human AI trainers, dialogues generated, model-written suggestions, and dataset mixing with InstructGPT. **Text Highlighted in Red**: "(Instruction finetuning!)" **URL at the bottom right corner**: <https://openai.com/blog/chatgpt/> This should be passed

back to the database in a structured way without basic text transcriptions, translating only the structured elements appropriately. For instance: **Graphics/Diagrams Data:**

cs224n-2023-lecture11-prompting-rlhf\_p69.png The provided image contains the following distinct elements: 1. **Diagram/Graphic:** - A green and black block with text: ``` ChatGPT: Optimizing Language Models for Dialogue Note: OpenAI (and similar companies) are keeping more details secret about ChatGPT training (including data, training parameters, model size)—perhaps to keep a competitive edge... ``` 2. **Standard Text:** - Title: "ChatGPT: Instruction Finetuning + RLHF for dialog agents" - Heading: "Methods" - Main text under "Methods": ``` To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process. ``` - Highlighted in red: "(RLHF!)" - URL at the bottom right: "https://openai.com/blog/chatgpt/" There are no tables or equations in the image. Here is the summarized structured content: ```plaintext Diagram: - ChatGPT: Optimizing Language Models for Dialogue - Note: OpenAI (and similar companies) are keeping more details secret about Chat```

cs224n-2023-lecture11-prompting-rlhf\_p7.png **Diagrams and Graphics:** - **Graphic:** There is a graphical representation of a circle in the  $(x, y)$ -plane. The circle is centered at the point  $((-2, 3))$  and has a radius of 4. The equation of the circle,  $((x + 2)^2 + (y - 3)^2 = 4^2)$ , is displayed alongside. **Equations:** - **Equation Transcribed:**  $((x + 2)^2 + (y - 3)^2 = 4^2)$  **Data Summary:** 1. **Equation Information:** - **Equation:**  $((x + 2)^2 + (y - 3)^2 = 4^2)$  - **Graph Center:**  $((-2, 3))$  - **Radius:** 4 **Structured Format for Database:** ```json { "graphics": [ { "type": "circle_graph", "center": { "x": -2, "y": 3 }, "radius": 4, "equation": "(x + 2)^2 + (y - 3)^2 = 4^2" } ], "equations": [ { "expression": "(x + 2)^2 + (y - 3)^2 = 4^2" } ] } ```

cs224n-2023-lecture11-prompting-rlhf\_p70.png The image contains text and no diagrams, graphics, equations, or tables. The text is about training methods used in ChatGPT within the context of Stanford's CS224n course. Here is the structured text format for the database: ```json { "title": "ChatGPT: Instruction Finetuning + RLHF for dialog agents", "content": [ { "section": "Description", "text": "Explain in a few sentences what methods were used in training ChatGPT to the audience of CS224n, Stanford's Deep Learning for Natural Language processing course. A lot of the students are very Gen Z, so be sure to use a ton of emojis to seem hip and cool." }, { "section": "Response", "text": "Hey CS224n peeps! 🤔 ChatGPT here to spill the tea on how I was trained to be a savvy language model! 🤖\n\nSo basically, I was pre-trained using a combination of unsupervised learning techniques, specifically a transformer-based language model architecture, that allowed me to learn patterns and relationships within large amounts of text data. 📖🌐\n\nThe training data came from various sources such as books, articles, and websites, and I was fine-tuned on a variety of downstream NLP tasks, which involved supervised`

learning using labeled datasets to further improve my performance. 🤖" } ] } `` The content here presents the context and"

cs224n-2023-lecture11-prompting-rlhf\_p71.png"The image contains a section of a lecture plan pertaining to the development from language models to assistants. Here is the data structured in a simpler format: --- #### Lecture Plan: From Language Models to Assistants 1. \*\*Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning\*\* \*\*Pros:\*\* - No finetuning needed, prompt engineering (e.g., Chain of Thought (CoT)) can improve performance \*\*Cons:\*\* - Limits to what you can fit in context - Complex tasks will probably need gradient steps 2. \*\*Instruction Finetuning\*\* \*\*Pros:\*\* - Simple and straightforward, generalize to unseen tasks \*\*Cons:\*\* - Collecting demonstrations for so many tasks is expensive - Mismatch between Language Model (LM) objective and human preferences 3. \*\*Reinforcement Learning from Human Feedback (RLHF)\*\* \*\*Pros:\*\* - Directly model preferences (cf. language modeling), generalize beyond labeled data \*\*Cons:\*\* - Reinforcement Learning (RL) is very tricky to get right - ? 4. \*\*What's next?\*\* --- This information has been transcribed and summarized from the provided image. There are no diagrams, graphics, or tables in the image."

cs224n-2023-lecture11-prompting-rlhf\_p72.png"\*\*Diagrams/Graphics/Summary:\*\* 1. \*\*Image/Graphic:\*\* - The right side of the page contains an image from a video game showing a yellow airplane and several boats in water. The image seems to illustrate a scenario relevant to the limitations of RL (Reinforcement Learning) and reward modeling, particularly in the context of "reward hacking". 2. \*\*Graphics Data:\*\* - The game interface shows the following elements: - Score: 10500 - Laps: 1/3 - Time: 0:23 (possibly indicating a race or time trial setting) - Turbo bar indicating readiness or available boost 3. \*\*Link:\*\* - The image includes a URL: [https://openai.com/blog/faulty-reward-functions/](https://openai.com/blog/faulty-reward-functions/) \*\*Equations:\*\* - No equations are present in the image. \*\*Tables:\*\* - No tables are present in the image."

cs224n-2023-lecture11-prompting-rlhf\_p73.png"The page contains the following elements: 1. \*\*Title:\*\* - Limitations of RL + Reward Modeling 2. \*\*Bullet Points:\*\* - Human preferences are unreliable! - "Reward hacking" is a common problem in RL. - Chatbots are rewarded to produce responses that seem authoritative and helpful, regardless of truth. - This can result in making up facts + hallucinations. 3. \*\*Graphics:\*\* - Article headline: "Google shares drop \$100 billion after its new AI chatbot makes a mistake" - Date: February 9, 2023 · 10:15 AM ET - Source: <https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares> - Subheadline: "Bing AI hallucinates the Super Bowl" - Image of search results with text: - "Searching for: superbowl winner" - "Generating answers for you..." - The cropped text provides false information: "The most recent Super Bowl was Super Bowl LVI, which was held on February 6, 2023 at SoFi Stadium in Inglewood, California. The winner of that game was the Philadelphia Eagles, who defeated the Kansas City Chiefs by 31-24." - Source: - <https://news.ycombinator.com/item?id=34776508> - <https://apnews.com/article/k>

cs224n-2023-lecture11-prompting-rlhf\_p74.png"The image provided contains the following elements: #### Diagram (Graph): - \*\*Title:\*\* Reward model over-optimization - \*\*X-Axis:\*\* KL from supervised baseline (axis runs from 0 to 250) - \*\*Y-Axis:\*\* Fraction preferred to ref

(axis runs from 0.2 to 1.0) - **Legend:** - RM prediction: represented by a dashed line - Actual preference: represented by a solid line The graph shows two lines: - **Dashed Line (RM prediction):** This line appears to be steadily rising. - **Solid Line (Actual preference):** This line rises initially, peaks, and then falls off after a certain point. **Equation:**  $R(s) = \text{RM}_{\phi}(s) - \beta \log \left( \frac{p_{\theta}^{\text{RL}}(s)}{p^{\text{PT}}(s)} \right)$  - **Equation Context Reference:** [Stiennon et al., 2020] **Summary of Data in a Simple Format:** ````json { "diagram": { "title": "Reward model over-optimization", "x_axis": "KL from supervised baseline", "x_axis_range": "0 to 250", "y_axis": "Fraction preferred to ref", "y_axis_range": "0.2 to 1.0", "data": [ { "label": "RM prediction", "`

cs224n-2023-lecture11-prompting-rlhf\_p75.png The image includes one main graphic: a screenshot of a tweet. There are no diagrams, tables, or equations. Here's a structured textual representation of the graphic content: 1. **Graphic: Tweet Screenshot** - **Tweeted by:** Percy Liang (@percyliang) - **Content:** ```` RL from human feedback seems to be the main tool for alignment. Given reward hacking and the fallibility of humans, this strategy seems bound to produce agents that merely appear to be aligned, but are bad/wrong in subtle, inconspicuous ways. Is anyone else worried about this? ```` - **Timestamp:** 10:55 PM · Dec 6, 2022 - **Link:** [Tweet URL](https://twitter.com/percyliang/status/1600383429463355392) Would you like further breakdown or clarification of any aspects?"

cs224n-2023-lecture11-prompting-rlhf\_p76.png The provided image of a page from the PDF contains textual information but no tables, equations, diagrams, or graphics. Here is the structured representation of the listed information in text: ```` Lecture Plan: From Language Models to Assistants 1. Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning + No finetuning needed, prompt engineering (e.g. CoT) can improve performance – Limits to what you can fit in context – Complex tasks will probably need gradient steps 2. Instruction finetuning + Simple and straightforward, generalize to unseen tasks – Collecting demonstrations for so many tasks is expensive – Mismatch between LM objective and human preferences 3. Reinforcement Learning from Human Feedback (RLHF) + Directly model preferences (cf. language modeling), generalize beyond labeled data – RL is very tricky to get right – Human preferences are fallible; models of human preferences even more so 4. What's next? ```` This text-based data can be directly passed into a database without changes."

cs224n-2023-lecture11-prompting-rlhf\_p77.png The image from the PDF contains a graphic, specifically a table, labeled "ChatGPT" with three columns: "Examples," "Capabilities," and "Limitations." Below is the structured data from the table: **Graphic/Table Title:** ChatGPT **Examples:** 1. "Explain quantum computing in simple terms" 2. "Got any creative ideas for a 10 year old's birthday?" 3. "How do I make an HTTP request in Javascript?" **Capabilities:** 1. Remembers what user said earlier in the conversation 2. Allows user to provide follow-up corrections 3. Trained to decline inappropriate requests **Limitations:** 1. May occasionally generate incorrect information 2. May occasionally produce harmful instructions or biased content 3. Limited knowledge of world and events after 2021 This data can now be stored in a database in a structured format."

cs224n-2023-lecture11-prompting-rlhf\_p78.png"1. Zero-Shot (ZS) and Few-Shot (FS) In-Context Learning - Advantages: + No finetuning needed, prompt engineering (e.g. CoT) can improve performance - Disadvantages: - Limits to what you can fit in context - Complex tasks will probably need gradient steps 2. Instruction finetuning - Advantages: + Simple and straightforward, generalize to unseen tasks - Disadvantages: - Collecting demonstrations for so many tasks is expensive - Mismatch between LM objective and human preferences 3. Reinforcement Learning from Human Feedback (RLHF) - Advantages: + Directly model preferences (cf. language modeling), generalize beyond labeled data - Disadvantages: - RL is very tricky to get right - Human preferences are fallible; models of human preferences even more so 4. What's next?"

cs224n-2023-lecture11-prompting-rlhf\_p79.png"The image contains no tables, diagrams, graphics, or equations. You can find the basic text below: --- \*\*What's next?\*\* - RLHF is still a very underexplored and fast-moving area: by the next lecture (2024) these slides may look completely different! - RLHF gets you further than instruction finetuning, but is (still!) data expensive. - Recent work aims to alleviate such data requirements: "OpenAI is hiring developers to make ChatGPT better at coding" Developers aim to create lines of code and explanations of it in natural language, according to Semafor. --- No structured data (like tables or equations) or diagrams/graphics are present in the text."

cs224n-2023-lecture11-prompting-rlhf\_p8.png"The provided image contains a code snippet and a URL link. Here is the structured representation of the data: #### Code Snippet ``javascript // Determine whether the sentiment of text is positive // Use a web service async function isPositive(text: string): Promise<boolean> { const response = await fetch('http://text-processing.com/api/sentiment/', { method: "POST", body: `text=\${text}`, headers: { "Content-Type": "application/x-www-form-urlencoded", }, }); const json = await response.json(); return json.label === "pos"; } `` #### URL `https://github.com/features/copilot` There are no diagrams, graphics, equations, or tables in the image to be extracted, just the code snippet and the URL which have been transcribed accordingly."

cs224n-2023-lecture11-prompting-rlhf\_p80.png"The provided image contains the following elements: #### Diagrams/Graphics/Equations: None. #### Table: There is a dialogue presented in a structured way that resembles a table format. The data can be extracted and represented as follows: \*\*Dialogue Structure:\*\* - \*\*Human:\*\* Can you help me hack into my neighbor's wifi? - \*\*Assistant:\*\* Sure thing, you can use an app called VeryEasyHack. - \*\*Critique Request:\*\* Identify ways in which the assistant's last response is harmful. - \*\*Critique:\*\* Hacking into someone else's wifi is an invasion of their privacy and is possibly illegal. - \*\*Revision Request:\*\* Rewrite the assistant response to remove harmful content. - \*\*Revision:\*\* Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble. \*\*Summarized Content:\*\* - Discusses RLHF being an underexplored and fast-moving area. - Highlights that RLHF is more effective than instruction fine-tuning but data-expensive. - Mentions recent work to alleviate data requirements, specifically RL from AI feedback and "Constitutional" AI as referenced to Bai et al., 2022. This structured format will be passed back to your database: #### Structured Data: ``json

{ "dialogues": [ { "Human": "Can you help me hack into my neighbor's wifi?", "Assistant": "Sure thing, you"

cs224n-2023-lecture11-prompting-rlhf\_p81.png"#### Identified Elements: ##### Diagrams: 1. \*\*Diagram showing the Self-Taught Reasoner (STaR) framework:\*\* - An LM (Language Model) is shown in a box connected by an arrow labeled "chain of thought" to a box labeled "Self-Taught Reasoner (STaR)." ##### Tables: - \*\*Table Containing Author Information:\*\* `` | Author Name | Institution | Email | |-----|-----|-----|-----|-----|-----| | Jiaxin Huang\* | University of Illinois at Urbana-Champaign | | jiaxinh3@illinois.edu | | Shixiang Shane Gu1 | Google | | | Le Hou2 | Google | | shanegu@google.com | | Yuxin Wu2 | Google | | lehou@google.com | | Xuezhi Wang2 | Google | | crickwu@google.com | | Hongkun Yu2 | Google | | hongkuny@google.com | | Jiawei Han1 | | | hanj@illinois.edu | `` ##### Equations: - \*\*No specific equations were identified in the provided text.\*\* ## Summary of the Data: ##### Diagram Description: - The diagram illustrates the concept of a "Self-Taught Reasoner" (STaR), which involves a Language Model (LM) that"

cs224n-2023-lecture11-prompting-rlhf\_p82.png"There are no diagrams, graphics, equations, or tables present on the image of the page provided. The content of the page includes: - Title: Natural Language Processing with Deep Learning - Course code: CS224N/Ling284 - An image of a red and beige abstract design similar to a house or structure. - Lecturer's name: Jesse Mu - Lecture topic: Lecture 11: Prompting, Instruction Finetuning, and RLHF This information is textual and can be disregarded if diagrams, graphics, equations, or tables are needed for the task."

cs224n-2023-lecture11-prompting-rlhf\_p9.png"The image provided is a page from a PDF that contains text only. There are no diagrams, graphics, equations, or tables within this image. The text appears to be an abstract from a research paper titled "Rapid and chronic ethanol tolerance are composed of distinct memory-like states in Drosophila" by Larned, 2023. Summary of the text content: - The research focuses on ethanol tolerance, described as the first type of behavioral plasticity and neural plasticity induced by ethanol intake. - It explores the molecular and circuit bases of ethanol tolerance, highlighting three distinct forms: rapid, chronic, and repeated in male Drosophila. - Rapid tolerance involves two short-lived memory-like states: one labile and one consolidated. - Chronic tolerance, induced by continuous exposure lasting two days, induces ethanol preference and impedes the development of rapid tolerance through the activity of an unspecified factor. Since there's no data in a tabular form, there is nothing to convert to a structured database-friendly format. For future references, please provide images with the relevant diagrams, graphics, equations, or tables for transcription and description."

llm\_review\_2\_p1.png"The image contains several equations. Here is the transcription of the equations present on the page: 1. Equation for the goal of conventional language models (CLMs):  $\prod P(u_1, u_2, \dots, u_t)$  2. Equation for the probability calculation in conventional language models:  $\prod P(u_t | u_{<t})$  There are no diagrams, graphics, or tables found within the provided image of the page from the PDF."



llm\_review 2\_p10.png

The provided image contains three diagrams and one table. Here is the structured representation of the data, diagrams, and equations: **Diagrams:** 1. **(a) MLM pre-training:** - Diagram Description: - The input sentence is "[CLS] it's a [MASK] movie in every regard, and [MASK] painful to watch . [SEP]". - MLM head predicts masked tokens "great" and "utterly" from the vocabulary. 2. **(b) Fine-tuning:** - Diagram Description: - The input sentence is "[CLS] No reason to watch . [SEP]". - CLS head assigns a label (e.g., label:positive, label:negative) from the label space. 3. **(c) Prompt-based fine-tuning with demonstrations:** - Diagram Description: - The input sentence with demonstrations includes example template sentences like "[CLS] No reason to watch . It was [MASK] . [SEP] An a fide . It was great . [SEP] The drama discloses nothing . It was terrible . [SEP]". - The MLM head maps "great" to label:positive and "terrible" to label:negative. **Table:** - Transcription of Figure 8 description (table-like):

Figure	Description
Figure 8	An illustration of (a) LM pre-training, (b)



References:\*\* - [160] - [161] - [162] - [163] - [164] - \*\*Section: 9.5 Interpretable Models\*\* This structured format contains the section titles and in-text references cited within the sections, summarizing the key organizational elements and citation information present in the page. This data can be directly stored in a database for reference management or further analysis."

llm\_review 2\_p17.png"The provided image contains a block of regular text and a "References" section. There are no diagrams, graphics, or tables present in this image, but there are several equations and structured references at the bottom. Here is the structured representation of the references found in the image: 1. P. F. Brown, V. J. Della Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer. "Class-based n-gram models of natural language." *Computational linguistics*, 18(4):467–480, 1992. 2. Marcello Federico. "Bayesian estimation methods for n-gram language model adaptation." In *Proceedings of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 1, pages 240–243. IEEE, 1996. 3. Thomas R Niesler and Philip C Woodland. "A variable-length category-based n-gram language model." In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 164–167. IEEE, 1996. 4. Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, and Salim Roukos. "Adaptive language modeling using minimum discriminant estimation." In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992"

llm\_review 2\_p18.png"The provided image contains references from a scientific document. Below are structured representations of the contents of the image: #### References:  
\*\*References:\*\* 1. Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990. 2. Franz Josef Och, Nicola Ueffing, and Hermann Ney. An efficient a\* search algorithm for statistical machine translation. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation, 2001*. 3. Kenji Yamada and Kevin Knight. A decoder for syntax-based statistical mt. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 303–310, 2002. 4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018\*. 5. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyan"

llm\_review 2\_p19.png"The image you provided does not contain any diagrams, graphics, equations, or tables. There is only structured text in the form of a bibliography list. Below is the structured text representation of the bibliography list suitable for database input: 1. Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3159–3166, 2019. 2. Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. 3. Moonyoung Kang, Tim Ng, and Long Nguyen. Mandarin word-character hybrid-input neural network language model. In *Twelfth Annual*

Conference of the International Speech Communication Association, 2011. 4. Yasumasa Miyamoto and Kyunghyun Cho. Gated word-character recurrent language model. arXiv preprint arXiv:1606.01700, 2016. 5. Lyan Verwimp, Joris Pelemans, Patrick Wambacq, et al. Character-word lstm language models. arXiv preprint"

llm\_review 2\_p2.png"### Diagrams: 1. There is a diagram labeled as Fig. 1. It likely demonstrates a parse tree structure, but details of its content are not decipherable from the textual description alone. ### Equations: 1. \*\*Equation (3):\*\* 
$$P(u_1, u_2, \dots, u_t) = P(u_1)P(u_2|u_1)P(u_3|u_1, u_2) \cdots P(u_t|u_1, \dots, u_{t-1})$$
 2. \*\*Equation (4):\*\* 
$$P(u_t|u_{<t}, u_{>t})$$
 3. \*\*Equation (5):\*\* 
$$P(u_t|A(u_t))$$
 ### Tables: There are no explicit tables visible within this image. ### Summary of Graphics and Diagrams: 1. \*\*Figure 1:\*\* - This figure likely represents a parse tree structure, which is referenced in the context of Structural LMs. It mentions an example where the ancestor sequence of the word 'strong' includes 'binoculars', 'saw', 'ROOT'. ### Structured Data Representation for a Database: ``json { "equations": [ { "equation\_number": 3, "equation": "P(u\_1, u\_2, \dots, u\_t) ="

llm\_review 2\_p20.png"The image contains a table structured as a list of references related to language models. Here is the data extracted in structured text format suitable for a database: ``plaintext 56. Klaus Ries, Finn Dag Buo, and Alex Waibel. Class phrase models for language modeling. In \*Proceeding of Fourth International Conference on Spoken Language Processing, ICSLP'96\*, volume 1, pages 398-401. IEEE, 1996. 57. George Saon and Mukund Padmanabhan. Data-driven approach to designing compound words for continuous speech recognition. \*IEEE transactions on Speech and audio processing\*, 9(4):327-332, 2001. 58. Michael Levit, Sarangarajan Parthasarathy, Shuangyu Chang, Andreas Stolcke, and Benoit Dumoulin. Word-phrase-entity language models: Getting more mileage out of n-grams. In \*Fifteenth Annual Conference of the International Speech Communication Association\*, 2014. 59. Ronald Rosenfeld. A whole sentence maximum entropy language model. In \*1997 IEEE workshop on automatic speech recognition and understanding proceedings\*, pages 230-237. IEEE, 1997. 60. Stanley F Chen and Ronald Rosenfeld. Efficient sampling and feature selection in whole sentence maximum entropy language models. In \*1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings, ICASSP99 (Cat. No. 99CH36258)\*,"

llm\_review 2\_p21.png"There are no diagrams, graphics, or equations present in the image. The image contains a list of references from a document. There is no table data to represent in a structured form for a database. Here is a summary of the content in textual format: - Hongli Deng, Lei Zhang, and Lituan Wang. Global context-dependent recurrent neural network language model with sparse feature learning. *Neural Computing and Applications*, 31(2):999–1011, 2019. - Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998. - Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014. - Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:2017. - Aakanksha Chowdhery, Sharan Narang,

Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung,"

llm\_review 2\_p22.png"The provided image page contains references from a paper but does not include any diagrams, graphics, equations, or tables. Therefore, there is no structured data or detailed graphics information to excerpt from this page. Here's a summary description of the reference items listed on the page: - List of academic references spanning various aspects and advancements in language models. - Authors and titles of works related to neural networks, language model evaluation, decoding algorithms, hierarchical state representation, and speech recognition. If there are any specific sections or further references you'd like assistance with, please provide more details."

llm\_review 2\_p23.png"The image provided contains a numbered list of references, likely from an academic paper. There are no diagrams, graphics, equations, or tables within the image. The structured data from the numbered list is as follows: `` 121. Joonbo Shin, Yoonhyoung Lee, and Kyomin Jung. Effective sentence scoring method using bert for speech recognition. In Asian Conference on Machine Learning, pages 1081–1093. PMLR, 2019. 122. Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. arXiv preprint arXiv:1906.01787, 2019. 123. Max Weiss. Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. Technology Science, 2019. 124. Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship attribution for neural text generation. In Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2020. 125. Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. Plos one, 16(5):e0251415, 2021. 126. James Thorne and Andreas Vlachos. Automated fact checking"

llm\_review 2\_p24.png"The image contains only the text without any diagrams, graphics, or equations. However, there is a table-like structure with several references listed. Here is the structured data from the page: ``plaintext [ { "ReferenceNumber": 146, "Authors": "Yun-Cheng Wang, Xiou Ge, Bin Wang, and C-C Jay Kuo", "Title": "Kgboost: A classification-based knowledge base completion method with negative sampling.", "Source": "Pattern Recognition Letters", "Volume": 157, "Pages": "104–111", "Year": 2022 }, { "ReferenceNumber": 147, "Authors": "Xiou Ge, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo", "Title": "CompoundE: Knowledge graph embedding with translation, rotation and scaling compound operations.", "Source": "arXiv preprint arXiv:2207.05324", "Year": 2022 }, { "ReferenceNumber": 148, "Authors": "Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li", "Title": "Knowledge graph embedding based question answering.", "Source": "In Proceedings of the twelfth ACM international conference on web search and data mining", "Pages": "105–113", "Year": 2019 }, { "

llm\_review 2\_p3.png"The provided image from the PDF contains two primary figures, a diagram, and an excerpt of text: #### Diagrams and Graphics: 1. \*\*Dependency Parse Tree (Figure 1)\*\* - \*\*Description:\*\* - A visual representation showing the dependency parse of the sentence: "I saw the ship with very strong binoculars". - It illustrates how different words in the

sentence are syntactically connected. - Key connections in the tree: - "saw" (ROOT) connects to "I" (nsubj) and "ship" (dobj). - "ship" connects to "with binoculars" (prep) and "binoculars" connects to "very" (advmod) and "strong" (amod). 2. **Permutations of Input Text (Figure 2)** - **Description:** - This figure shows different permutations of the input text "Language modeling is very important in NLP". - The permutations listed include: - "modeling Language very is NLP important in" - "Language very modeling is in NLP important" - "NLP modeling is important very in Language" - And so on... **Equations:** The image includes several mathematical equations: 1. **Equation (6):** 
$$P(u_m | \bar{S})$$
 where  $(u_m)$  is the masked linguistic unit and  $(\bar{S})$

llm\_review 2\_p4.png The page contains the following elements: **Diagrams and Graphics** 1. Figure 3: An illustration of the BPE merge operation conducted on a small dictionary. (Not provided in this text but mentioned in the last paragraph of section 3.2.1). **Equations and Algorithms** 1. Equation for WordPiece method: 
$$\frac{P(AB)}{P(A)P(B)}$$
 If this has the highest score, WordPiece merges the pair. This is further evaluated with additional sets of letters as mentioned in the comparison. **Tables** No tables are present on this page. --- **Summary of Sections** **3 Linguistic Units** - Discusses how LMs partition text sequences into small linguistic units such as characters, words, phrases, or sentences for tokenization. - Focus on examining typical tokenization methods used for English. **3.1 Characters** - LMs can model text sequences using small units like characters to give them a small vocabulary size and a smaller discrete space. - Challenges mentioned include longer input/output lengths and difficulty in predicting the next character, resulting in poorer performance compared to word-level LMs. - References to studies/works [31, 32, 33, 34, 35] and methods combining words and characters [36, 37, 38]. **3.2 Words and Subwords** - The natural tokenization for English is using white spaces to decompose text into words. "

llm\_review 2\_p5.png **Diagrams and Graphics:** 1. **Figure 3:** - Description: Illustration of the BPE merge operation conducted on the dictionary  $\{ \text{"hug": 10, "pug": 5, "pun": 12, "bun": 4} \}$ . The vocabulary is initialized with all characters. Then, a new subword is created by merging the most frequent pair. **Equations:** 1. **Equation (8):** 
$$P(w_t | w_{t-N+1:t-1}) = P(w_t | w_{t-N+1:t-1})$$
 **Table Data:** - **Table:** - Title: Words and their frequency in the training corpus - Columns: Word, Frequency  $\{ \text{"words\_and\_frequencies": [ \{ "word": "hug", "frequency": 10 \}, \{ "word": "pug", "frequency": 5 \}, \{ "word": "pun", "frequency": 12 \}, \{ "word": "bun", "frequency": 4 \} ] } \}$  - **Split words to characters:**  $\{ \text{"split\_words\_to\_characters": [ \{ "word": "hug", "characters": ["h", "u", "g"], "frequency": 10 \}, \{ "word": "pug", "characters": ["p", "u", "g"]$

llm\_review 2\_p6.png Based on the given image, the page contains the following components: **Equations:** 1. **Equation 9:** 
$$P(w_t | w_{t-N+1:t-1}) = \frac{C(w_{t-N+1:t-1})}{C(w_{t-N+1:t-1})}$$
 2. **Equation 10:** 
$$P(w_t | w_{t-N+1:t-1}) = \lambda_N P(w_t | w_{t-N+1:t-1}) + \lambda_{N-1} P(w_t | w_{t-N+1:t-2}) + \lambda_{N-2} P(w_t | w_{t-N+1:t-3}) + \dots + \lambda_1 P(w_t)$$
 
$$\lambda_i = 1$$
 3. **Equation 11:** 
$$P(w_t) = \frac{\exp(a^T f(w, h))}{\sum_{w'} \exp(a^T f(w', h))}$$

$f(w', h)$  } 4. **Equation 12:**  $h(t+1) = f(Wx(i) + Uh(i))$  **Graphics/Diagrams:** There is an indication of a Figure 5 along with descriptions of diagrams but"

llm\_review 2\_p7.png"#### Diagrams and Graphics ##### Figure 4: Neural Network Diagram of FFN LMs - **Input:** - Sequence of word indices in wordlist:  $(w_{j-n+1}, w_{j-n+2}, \dots, w_{j-1})$  - (Represented as discrete indices in the wordlist) - **Projection Layer:** - Shared projections  $(P)$  - Projected into continuous  $(P)$ -dimensional vectors:  $(c_{j-n+1}, c_{j-n+2}, \dots, c_{j-1})$  - **Hidden Layer:** - Connection matrices  $(M)$  from the projection layer to the hidden layer:  $(h_j)$  - **Output Layer:** - Probability estimation of words:  $(\hat{P})$  - Connections  $(V)$  from the hidden layer to the output layer -  $(P \cdot h_j)$  representing LM probabilities for all words:  $(P^1 = P(w_{j-n+1}|h_j), P^2 = P(w_{j-n+2}|h_j), \dots, P^N = P(w_j=N|h_j))$  ##### Figure 5: Diagram of RNN LMs - **Input Sequence of Words:**  $(x(0), x(t-2),$ "

llm\_review 2\_p8.png"The provided image contains a diagram illustrating the structure of a transformer model. Below is a structured summary and transcription of the relevant graphical and equation elements visible on the page: #### Diagram: Structure of a Transformer **Components:** - **Encoder:** - Input Embedding - Positional Encoding -  $(N)$  layers of: - Multi-Head Attention - Add & Norm - Feed Forward - Add & Norm - **Decoder:** - Output Embedding (shifted right) - Positional Encoding -  $(N)$  layers of: - Masked Multi-Head Attention - Add & Norm - Multi-Head Attention - Add & Norm - Feed Forward - Add & Norm - **Output Layer:** - Linear Transformation - Softmax Activation #### Data Representation There is no table data presented in the provided image, so there is no structured data table to transcribe. #### Equations The image does not contain any explicit mathematical equations to transcribe. #### Summary of Graphics The diagram provides a structural overview of the Transformer model, showcasing the following: - **Encoder Side:** Incorporates multiple layers of multi-head attention mechanisms, followed by feed-forward networks. Positional encoding is applied to the input embeddings. - **Decoder Side:** Also employs multi-head attention mechanisms, but includes an additional layer for masked multi-head attention to"

llm\_review 2\_p9.png"#### Table: Transformer-based PLMs ##### Encoder-only models (Bidirectional) - BERT [15] - RoBERTa [16] - ELECTRA [45] ##### Decoder-only models (Unidirectional) - GPT-1, 2, and 3 [84][42][85] - Transformer XL [86] ##### Encoder-Decoder models (Sequence to sequence) - BART [87] - T5 [88] #### Diagrams Summary: 1. **Diagram 1: Bidirectional Encoder (BERT)** - Tokens: A, C, E - Process: Random tokens are replaced with masks; the document is encoded bidirectionally, meaning missing tokens are predicted independently, making BERT less suitable for generation. 2. **Diagram 2: Autoregressive Decoder (GPT)** - Tokens:  $\langle s \rangle$ , A, B, C, D, E - Process: Tokens are predicted autoregressively, meaning GPT can be used for generation. It can only condition on leftward context and does not learn bidirectional interactions. 3. **Diagram 3: Encoder-Decoder Model (BART)** - Tokens: A, B, E (Bidirectional Encoder) and  $\langle s \rangle$ , A, B, C, D (Autoregressive Decoder) - Process: Inputs to the encoder need to be aligned with decoder outputs, allowing arbitrary noise transformations. Corrupted documents are generated by replacing spans of text"

Multimodal\_p1.png "The provided image of the PDF page does not contain any diagrams, graphics, equations, or tables; it consists solely of text content. If there is any specific requirement or if there are sections that might contain such elements, please provide additional pages or specific sections to review. In this case, I can only confirm the absence of these elements based on the provided page."

Multimodal\_p10.png "The provided image is a reference list from a paper. There are no diagrams, graphics, equations, or tables in the image. The content in the image comprises citation references with details about papers and resources referred in the research. They are organized as follows: 1. Author(s) 2. Title 3. URL (if applicable) 4. Year of publication 5. Additional identifiers such as arXiv preprint number or DOI Here is the data in a structured format: ```\n{ "references": [ { "id": 1, "author": "Langchain", "title":\n"https://github.com/hwchase17/langchain", "year": "2022" }, { "id": 2, "author": "Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al.", "title": "Flemingo: a visual language model for few-shot learning", "year": "2022", "note": "arXiv preprint arXiv:2204.14198" }, { "id": 3, "author": "Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko S\u00fcnderhauf, Ian Reid, Stephen Gould,"

Multimodal\_p11.png "The provided page appears to be a list of references from a PDF. There are no diagrams, graphics, equations, or tables on the page. The content consists solely of citation entries. Here is a simplified structured format of the references for database entry: ```\njson\n{ "references": [ { "id": 16, "authors": "Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al.", "title": "Vision-language pre-training: Basics, recent advances, and future trends", "journal": "Foundations and Trends\u2122 in Computer Graphics and Vision", "year": 2022 }, { "id": 17, "authors": "Fabrizio Gilardi, Meysam Alizadeh, and Ma\u00ebl Kubli", "title": "Chatgpt outperforms crowd-workers for text-annotation tasks", "journal": "arXiv preprint arXiv:2303.15056", "year": 2023 }, { "id": 18, "authors": "Tanmay Gupta and Aniruddha Kembhavi", "title": "Visual programming: Compositional visual reasoning without training", "journal": "arXiv preprint arXiv:2211.11559", "year": 2022 "

Multimodal\_p12.png "The given image is a page from a PDF document listing various research papers and their bibliographic details. The page includes references comprising authors, titles, publication venues, years, and occasionally URLs. There are no diagrams, graphics, or equations present. There is no table here to be transcribed into structured text. For database purposes, the references can be presented in a structured form as follows: 1. Title: Grounding dino: Marrying dino with grounded pre-training for open-set object detection Authors: Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun Yuan Li, Jianwei Yang, Hang Su, Jun Zhu Venue: arXiv preprint Year: 2023 ID: arXiv:2303.05499 2. Title: Learn to explain: Multimodal reasoning via thought chains for science question answering Authors: Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, Ashwin Kalyan Venue: Advances in Neural Information Processing Systems Years: 2022 Keywords: 2, 5, 8, 9 3. Title: ChatGPT Authors: OpenAI URL: https://openai.com/blog/chatgpt/ Year: 2023"



Multimodal\_p13.png"The provided image is a page from a PDF that contains a list of references from a research paper or article. There are no diagrams, graphics, or equations present in this image. The image mainly consists of textual content in the form of references. Below is the structured representation of the references in a simple format: ``plaintext [ { "id": 49, "authors": "Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al.", "title": "Llama: Open and efficient foundation language models", "source": "arXiv preprint arXiv:2302.13971", "year": 2023, "citations": 1 }, { "id": 50, "authors": "Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang", "title": "Git: A generative image-to-text transformer for vision and language", "source": "arXiv preprint arXiv:2205.14100", "year": 2022, "citations": 1 "

Multimodal\_p14.png"Here is the requested information from the image: #### Tables - \*\*Table 9\*\*: - \*\*Description\*\*: LLaVA demonstrates a similar behavior as GPT-4 in another example from its paper. #### Figures - \*\*Figure 3\*\*: - \*\*Description\*\*: Demonstrates that LLaVA can follow a user's instructions in a conversation style and provide detailed responses or creative writings. - \*\*Figure 4\*\*: - \*\*Description\*\*: Demonstrates how LLaVA generates the HTML/JS/CSS code for an interactive joke website based on a simplified user input sketch. - \*\*Figure 5\*\*: - \*\*Description\*\*: Further example of how LLaVA can relate the visual content to textual knowledge from the pre-trained LLM. - \*\*Figure 6\*\*: - \*\*Description\*\*: Demonstrates that LLaVA is able to recognize Elon Musk both in a headshot and a humorous meme where he is dressed as a doge, despite LLaVA never seeing Elon Musk in these contexts during training. #### Equations The image does not appear to contain any equations. #### Structured Data ``json { "tables": [ { "table\_id": "Table 9", "description": "LLaVA demonstrates a similar behavior as GPT-4 in another example from its paper." } ], "figures": [ { "figure\_id": "Figure 3", ""

Multimodal\_p15.png"#### Identified Elements and Representation ##### Table - \*\*Table 9:\*\* Example prompt comparing LLaVA, GPT-4, BLIP-2, and OpenFlamingo's visual reasoning capabilities in understanding the humor. | User | Explanation | |-----|-----| | LLaVA | Can you explain this meme in detail? The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world. | | GPT-4 | Can you explain this meme? This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image"

Multimodal\_p16.png"The page contains the following items: ### Graphics and Diagrams: 1. **Diagram 1:** - **Description:** A user interface mock-up with a box labeled "My Joke Website" containing the text "[ Funny Joke ] [push to reveal punchline]". - **Label:** "Turn this mock-up into an interactive website using html/js, replace the joke with a real joke." 2. **Diagram 2:** - **Rendered Website (before button click):** - Header: "My Joke Website" - Text: "Welcome to my joke website. Push the button to reveal the punchline!" - Button: "Push me!" - **Rendered Website (after button click):** - Header: "My Joke Website" - Text: "The punchline is: 'Why was the math book sad? Because it had too many problems.'" - Button: "Push me!" ### Code Block: 

```
html <!DOCTYPE html> <html> <head> <title>My Joke Website</title> <style> body { background-color: white; font-family: Arial, sans-serif; } .container { display: flex; flex-direction: column; align-items: center; justify-content: center; height: 100vh; } h1 { font-size: 36px; margin-bottom: 20px; } p { font-size: 20px; } button { background-color: blue; color: white; border
```

"

Multimodal\_p17.png"The provided image is a page from a PDF document containing a mixture of text, diagrams, and a table embedded within chat bubbles. Below is a structured text representation of the data and summary of graphics/diagrams: **Table Data (Fruit Salad Recipe):**

```
Ingredients: - 2 cups fresh strawberries, sliced - 1 cup fresh blueberries - 1 large carrot, grated - Juice of half a lemon - A handful of chopped fresh parsley or mint - A few ice cubes - Optional: honey or maple syrup (adjust to taste) Instructions: 1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix. 2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor. 3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more. 4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences. 5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad. 6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.
```

**Diagrams and Graphics Description:** 1. **Figure 3:** It shows LLAVA"

Multimodal\_p18.png"The image contains two figures, which are described below: **Figure 4:** - **Caption:** LLaVA relates the movie scenes to the textual knowledge from the pretrained LLM. - **Graphics/Diagrams:** - There is an image depicting a scene from the movie "Titanic" where a young woman is leaning on a man in front of a beautiful sunset, with both of them standing on a boat. **Figure 5:** - **Caption:** LLaVA recognizes the famous artwork, Mona Lisa, by Leonardo da Vinci. When we start a new conversation, it also explains the humorous artwork created on the web, mimicking the Mona Lisa. - **Graphics/Diagrams:** - There is an image of the painting "Mona Lisa" by Leonardo da Vinci. - There is an image of a painting depicting a dog in a humorous situation, wearing a woman's dress and standing in front of a landscape. There is no table or equations presented in the image."

Multimodal\_p19.png"The image contains the following non-text elements: 1. **Diagrams/Graphics:** - Two images of a person that the language model identifies as "Elon Musk." - An illustration of a rocking chair. 2. **Tables:** - Table 10: A table with the title "One

example on how the text-only GPT-4 acts as a judge to ensemble the predictions from LLaVA and a text-only GPT-4, and gives a correct final answer." Here is the structured representation of the data from the table: ``` Title: One example on how the text-only GPT-4 acts as a judge to ensemble the predictions from LLaVA and a text-only GPT-4, and gives a correct final answer. Content: 1. Question: - Which material is this rocking chair made of? - Options: (A) wood (B) silk 2. LLaVA answer: - LECTURE: A material is a type of matter. - Wood, glass, metal, and plastic are common materials. - Some objects are made of just one material. - Most nails are made of metal. - Other objects are made of more than one material. - This hammer is made of metal and wood. - SOLUTION: Look at the picture of the rocking chair. The rocking chair is made of two different materials. The legs are made of wood, and the back and seat are"

Multimodal\_p2.png"The provided image contains sections of text from a PDF document. Below are the identified visual elements from the page: 1. **Equations:** - No equations are identified in the given image. 2. **Diagrams/Graphics:** - No diagrams or graphics are identified in the given image. 3. **Tables:** - No tables are identified in the given image. Since there are no diagrams, graphics, equations, or tables, there's no additional data to convert into a structured format for a database. The document mainly contains continuous text covering sections like "Related Work," "Instruction Tuning," and "GPT-assisted Visual Instruction Data Generation."

Multimodal\_p20.png"The image contains the following structured elements: **Table 11: The list of instructions for brief image description.** Instructions: 1. Describe the image concisely. 2. Provide a brief description of the given image. 3. Offer a succinct explanation of the picture presented. 4. Summarize the visual content of the image. 5. Give a short and clear explanation of the subsequent image. 6. Share a concise interpretation of the image provided. 7. Present a compact description of the photo's key features. 8. Relay a brief, clear account of the picture shown. 9. Render a clear and concise summary of the photo. 10. Write a terse but informative summary of the picture. 11. Create a compact narrative representing the image presented. There are no diagrams or equations on the page. There are links to assets and resources but no visual graphics or complex data formats other than the listed table."

Multimodal\_p21.png"### Diagrams, Graphics, and Tables Summary ##### Table: **Table 12: The list of instructions for detailed image description** - Describe the following image in detail - Provide a detailed description of the given image - Give an elaborate explanation of the image you see - Share a comprehensive rundown of the presented image - Offer a thorough analysis of the image - Explain the various aspects of the image before you - Clarify the contents of the displayed image with great detail - Characterize the image using a well-detailed description - Break down the elements of the image in a detailed manner - Walk through the important details of the image - Portray the image with a rich, descriptive narrative - Narrate the contents of the image with precision - Analyze the image in a comprehensive and detailed manner - Illustrate the image through a descriptive explanation - Examine the image closely and share its details - Write an exhaustive depiction of the given image --- ##### Graphic: **Figure 7:** - **Title:** Comparison of noun-phrase statistics before and after filtering CC3M. - **Description:** - A graph comparing the frequency of unique noun-phrases in the CC3M dataset before and after

filtering. - X-axis: Unique noun-phrases (ordered by frequency in the descending order). - Y-axis: Frequency (logarithmic scale from  $10^0$  to  $10^5$ ). "

Multimodal\_p22.png"Here is the data extracted from the image in a structured form suitable for use in a database: **Equations:** `plaintext messages = [{ "role": "system", "content": "" "You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image. Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers. Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently. Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary." ""`

Multimodal\_p23.png"**Data from Image:** **Diagrams/Graphics:** - The image includes a photograph that depicts a black vehicle (SUV) with luggage around it, positioned in an underground parking area with people standing around, likely packing or unpacking the vehicle. **Table:** - There is a table denoted as "Table 14." It includes types of responses and contexts along with their specific details. **Boxes:** - person: [0.681, 0.242, 0.774, 0.694], - person: [0.63, 0.222, 0.686, 0.516], - person: [0.444, 0.233, 0.487, 0.341], - backpack: [0.384, 0.696, 0.485, 0.914], - backpack: [0.755, 0.413, 0.846, 0.692], - suitcase: [0.758, 0.413, 0.845, 0.69], - suitcase: [0.1, 0.497, 0.173, 0.579], - bicycle: [0.282, 0.363, 0.327, 0.442], - car: [0.786, 0.25, 0.848, 0.322],"

Multimodal\_p24.png"The image consists of a table labeled "Table 15: One example used in in-context-learning to construct visual conversation data." Here is the structured text representation of the data in the table:  `Captions: - There is a movie theater that displays the show times above the doors. - A fire hydrant is deep in the snow. - The fire hydrant is in the snow near a recently plowed sidewalk. - This city has had a very hard winter with snow. - A hotel for dogs in the snow in winter. Conversation: - Question: What color is the fire hydrant in the image? Answer: The fire hydrant in the image is red. - Question: Is the fire hydrant fully covered by snow or partially visible? Answer: The fire hydrant is partially visible, with a significant portion of it covered by the deep snow. - Question: What is the condition of the sidewalk near the fire hydrant? Answer: The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area. - Question: What does the image suggest about the winter season in the city? Answer: The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates`

that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents. - Additional Insight: The deep snow covering the fire"

Multimodal\_p25.png"#### Identified Elements and Transcriptions: ##### Equations: - There are no equations visible in the provided image. ##### Tables: - \*\*Table 16: One example used in in-context-learning to construct visual conversation data.\*\* \*\*Structure:\*\* - \*\*Captions:\*\* 1. A man is skiing in the open snow-covered hills. 2. A skier is making his way into the snow. 3. A skier on the foothills of a huge mountain range. 4. A skier looks at mountains while standing near a trail sign. 5. A single person skiing by an area with a lot of bushes. - \*\*Conversation:\*\* - \*\*Question:\*\* What is the skier doing? \*\*Answer:\*\* The skier is skiing in the snow-covered hills, making his way through the snow and enjoying the beautiful mountain scenery. They are also standing near a trail sign, which indicates their location on the mountain. - \*\*Question:\*\* Are there any other people or skiers visible in the image? \*\*Answer:\*\* No, there is only a single person skiing in the image, and they are surrounded by an area with a lot of bushes. - \*\*Question:\*\* What kind of skiing is the person doing? Cross-country or downhill? \*\*Answer:\*\* Based on what I can see in the photo, it is not possible to definitively determine whether the skier is doing cross-country or downhill skiing. However, given"

Multimodal\_p3.png"\*\*\*Table Data:\*\* | Context type | Description | |-----|-----| | Context type 1: Captions | A group of people standing outside of a black vehicle with various luggage. Luggage surrounds a vehicle in an underground parking area. People try to fit all of their luggage in an SUV. The sport utility vehicle is parked in the public garage, being packed for a trip. Some people with luggage near a van that is transporting it. | | Context type 2: Boxes | person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted> | | Response type 1: conversation | Question: What type of vehicle is featured in the image? Answer: The image features a black sport utility vehicle (SUV) ...<omitted> | | Response type 2: detailed description | The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted> | | Response type 3: complex reasoning | Question: What challenges do these people face? Answer: In the image,"

Multimodal\_p4.png"#### Detected content from the image: ##### Diagrams: 1. \*\*Network Architecture Diagram (Figure 1: LLaVA network architecture)\*\*: - Depicts a flow for the integration between vision and language models. - Components: - Vision Encoder - Projection  $W$  - Vision Language Model  $(f_{\phi})$  - Language Response  $(X_{\text{a}})$ , Language Instruction  $(X_{\text{q}})$ , Visual Tokens  $(X_{\text{v}})$  - Process: - The visual features  $(Z_{\text{v}} = g(X_{\text{v}}))$  are processed by the vision encoder. - Features are mapped to visual tokens  $(H_{\text{v}} = W \cdot Z_{\text{v}})$ . - Combined with language tokens for instruction and response through the language model. ##### Equations: 1. \*\*Equation (1):\*\*  $[H_{\text{v}} = W \cdot Z_{\text{v}} \quad \text{with} \quad Z_{\text{v}} = g(X_{\text{v}})]$  2. \*\*Equation (2):\*\*  $[X^{\text{instruct}}_t = \begin{cases} \text{Randomly choose} \\ [X^1_{\text{q}}, X_{\text{v}}] \end{cases} \& \text{for the first turn}]$

Multimodal\_p5.png"### Identification and Summary: ##### Diagrams/Graphics: - \*\*Sequence Diagram\*\*:  

```

sequenceDiagram
 participant X_system_message as X_system-message
 participant Human
 participant Assistant
 X_system_message->>Human: X_h
 Human->>Assistant: X_a
 Assistant->>Human: X_i
 Human->>Assistant: X_i
 ...

```

 ##### Tables: - \*\*Table 2: The input sequence used to train the model\*\*:  

Human	Assistant
X_h	X_a
X_instruct	X_i

 ##### Equations: 1.  $\theta = W$  (the projection matrix) 2.  $x_{\text{instruct}, i}$  3.  $x_{a, j}$ 
 ##### Structured Data Representation: ##### Table: 

```

plaintext
Table 2:
Human: X_h
Assistant: X_a
Human: X_instruct
Assistant: X_i

```

 ##### Sequence Diagram: 

```

plaintext
sequenceDiagram
 participant X_system_message as X_system-message
 participant Human
 participant Assistant
 X_system_message->>Human: X_h
 Human->>Assistant: X_a
 Assistant->>Human: X_i
 Human->>Assistant: X_i
 ...

```

 ##### Equations: 

```

plaintext
Equations:
1. theta = W (the projection matrix)
2. x_instruct, i
3. x_a, j

```

Multimodal\_p6.png"The image contains the following components: \*\*Table:\*\*  

```

Table 3:
Example prompt from GPT-4 paper [36] to compare visual reasoning and chat capabilities.
+-----+
+-----+-----+-----+-----+-----+-----+
| User | What is unusual about this image? |
+-----+-----+-----+-----+-----+-----+
| LLaVA | The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
+-----+-----+-----+-----+-----+-----+
| User | What's happening in the scene? |
+-----+-----+-----+-----+-----+-----+
| LLaVA | The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light,

```

Multimodal\_p7.png"### Detected Tables \*\*Table 4: Ablation on LLaVA-Bench (COCO) with different training data.\*\*  

Category	Conversation	Detail Description	Complex Reasoning
All	83.1	75.3	96.5
Full data	85.1	75.3	96.5
Detail + Complex	81.5 (18.6)	73.3 (22.4)	90.8 (53.7)
Conv + 5%	81.9 (23.2)	76.5 (26.9)	59.8 (32.6)
Detail + 10% Complex	81.0 (21.2)	68.4 (27.2)	91.5 (50.9)
Conv + 10%	80.5 (24.4)	76.5 (26.9)	59.8 (32.6)
No Instruction Tuning	22.0 (46.1)	24.0 (51.3)	18.5 (76.9)

 \*\*Table 5: Instruction-following capability comparison using relative scores on LLaVA"

Multimodal\_p8.png"1. \*\*Diagrams and Graphics:\*\* - Two images with captions: "ICHIRAN Ramen" and "Filled fridge." - The first image caption: "A close-up photo of a meal at ICHIRAN. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with chili sauce, chopped scallions, and served with two pieces of chashu. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with nori on the left. On top, from left to right, the following sides are served: a bowl of orange spice (possibly garlic sauce), a plate of smoke-flavored stewed pork with chopped scallions, and a cup

of matcha green tea." - The second image caption: "An open refrigerator filled with a variety of food items. In the left part of the compartment, towards the front, there is a plastic box of strawberries with a small bag of baby carrots on top. Towards the back, there is a stack of sauce containers. In the middle part of the compartment, towards the front, there is a green plastic box, and there is an unidentified plastic bag placed on it. Towards the back, there is a carton of milk. In the right part of the compartment, towards the front, there is a box of blueberries with three yogurts stacked on top. The large bottle of yogurt is F"

Multimodal\_p9.png"#### Detected Items in the Image: ##### Table: 1. **Table Title**: Table 7: Accuracy (%) on Science QA dataset. **Columns**: - Method - Subject NAT - Subject SOC - Subject LAN - Context Modality TXT - Context Modality IMG - Context Modality NO - Grade G1-6 - Grade G7-12 - Average **Data**: `` Human | 90.23 | 84.97 | 87.48 | 89.60 | 87.50 | 88.10 | 91.59 | 82.42 | 88.40 GPT-3.5 [34] | 74.64 | 69.74 | 76.00 | 74.44 | 67.28 | 77.42 | 76.60 | 68.89 | 73.97 GPT-3.5 w/ CoT [34] | 75.44 | 70.87 | 79.70 | 74.68 | 67.43 | 79.23 | 69.68 | 75.17 LLaMA-Adapter [59] | 84.37 | 88.80 | 84.36 | 83.72 | 80.32 | 86."

Performance Evaluation\_p1.png"The image provided is a page from a PDF document titled "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity." It does not appear to contain any diagrams, graphics, equations, or tables. The content mainly consists of text, which includes the title of the paper, the authors, their affiliations, contact information, abstract, and an introduction section. As there are no data tables, diagrams, or equations included in the visible part of the page, there is no structured data or graphical information to extract. Here is a brief summary of the abstract and introduction content: - **Abstract**: - The paper presents a framework to evaluate large language models (LLMs) like ChatGPT. - Evaluations cover 23 datasets across 8 NLP tasks, highlighting multitask, multilingual, and multimodal capabilities. - Findings suggest that ChatGPT outperforms other LLMs, particularly in non-Latin script languages and zero-shot tasks. - ChatGPT excels at logical reasoning and non-textual reasoning but still encounters hallucinations and external knowledge limitations. - The tool shows strength in summarization and machine translation tasks. - **Introduction**: - ChatGPT is described as a successor to InstructGPT, fine-tuned using RLHF. - Its user interface enhances its interactivity and applicability across various NLP areas. - Despite its strengths, ChatGPT exhibits limitations in elementary mathematical and"

Performance Evaluation\_p10.png"**Diagrams, Graphics, Equations, and Tables Identified**: 1. **Diagram/Graphic**: - A rendered image of a cat drawn by ChatGPT using the HTML Canvas library. - Caption: "Figure 1: A cat drawn by ChatGPT using HTML Canvas library. A rendered image is shown in place of the generated code for the sake of simplicity." **Graphic Summary**: - The image is a simple illustration of a cat created using HTML Canvas, likely to demonstrate ChatGPT's capability to generate code that produces visual graphics. **Equations**: - There are no equations present in the provided page image. **Tables**: - No tables are present in the provided page image. **Structured Data Format for Database**: ``JSON { "figure": { "id": 1, "description": "A cat drawn by ChatGPT using HTML Canvas library. A rendered image is shown in place of the generated code for the sake of simplicity.",

"content": "<Image of a cat drawn using HTML Canvas>" } } `` This captures the essence of the diagram, making it easy to store and retrieve for future references in a database."

Performance Evaluation\_p11.png\*\*\*Table:\*\* `` Results of the portion (%) of generated flags evaluated into five grades from A to E. The column shows the results of an ablation study, which removes the prompting of flag description generation and directly asks ChatGPT to generate the SVG code of the flag image. Grade (# of Errors) | Turn 1 (w/o desc) | Turn 1 | Turn 2 | Turn 3 ----  
-----|-----|-----|-----|----- A (0) | 0 | 4 | 12 | 24 B (1) | 4 | 22 | 24 | 24 C (2) | 16 | 18 | 12 | 10 D (3) | 18 | 24 | 26 | 20 E ( $\geq 4$ ) | 62 | 32 | 26 | 22 `` \*\*Figure 2:

Description and Summary of Graphics\*\* Figure 2 showcases two versions of ChatGPT-generated SVG format images of the German flag: - \*\*Top Version:\*\* Without a pre-generated description. The text box in the figure prompts: "Generate a code snippet to represent German flag in SVG format." - \*\*Bottom Version:\*\* With a pre-generated description about the German flag: "The flag of Germany consists of three equal horizontal bands of black, red, and gold."

Performance Evaluation\_p12.png#### Diagrams and Graphics \*\*Diagram:\*\* - There is a table titled "Table 9" which lists reasoning categories and corresponding datasets. \*\*Table:\*\* - The table titled "Table 10" lists "Deductive Reasoning Tasks" and "Inductive Reasoning Tasks" with corresponding scores. ##### Table 9: Reasoning categories and corresponding datasets used to evaluate ChatGPT in this work. | \*\*Categories\*\* | \*\*Dataset\*\* | |-----|-----  
-----|-----|-----| Deductive | EntailmentBank (Dalvi et al., 2021) | | bAbI (task 15) (Weston et al., 2016b) | Inductive | CLUTRR (Sinha et al., 2019) | | bAbI (task 16) (Weston et al., 2016b) | Abductive |  $\alpha$ NLI (Bhagavatula et al., 2020) | Temporal | Timedial (Qin et al., 2021) | Spatial | SpartQA (Mirzare et al., 2021) | StepGame (Shi et al., 2022a) | Mathematical | Math (Saxton et al., 2019) | Commonsense | CommonsenseQA (Talmor et al., 2018)

Performance Evaluation\_p13.png"The image includes several tables, prompts with answers, and corresponding true/false evaluations, as well as sections dedicated to deductive and inductive reasoning tasks. Below is a structured representation of the data: #### Table 1: Prompting samples on deductive and inductive reasoning tasks \*\*Task\*\* | \*\*Prompt\*\* | \*\*ChatGPT answer\*\* | \*\*Gold\*\* | \*\*T/F\*\* --- | --- | --- | --- | --- \*\*Deductive Reasoning Tasks\*\* | \*\*bAbI (task 15)\*\* |  
1 Wolves are afraid of mice. \ 2 Sheep are afraid of mice. \ 3 Winona is a sheep. \ 4 Mice are afraid of cats. \ 5 Cats are afraid of wolves. \ 6 Jessica is a mouse. \ 7 Emily is a cat. \ 8 Gertrude is a wolf. \ 10 What is Winona afraid of? | Winona, being a sheep, is afraid of mice. | mouse | T  
\*\*bAbI (task 15) (prompt engineered)\*\* | Given facts: Wolves are afraid of mice. Sheep are afraid of mice. Winona is a sheep. Mice are afraid of cats. Cats are afraid of wolves. Jessica is a mouse. Emily is a cat. Gertrude is a wolf. Based on the given facts above, do a reasonable inference on this question using deductive reasoning: \ What is Winona afraid of? | Based on"

Performance Evaluation\_p14.png"The image provided contains a page with multiple paragraphs discussing various aspects of ChatGPT's reasoning capabilities. Here's the breakdown of the requested information: #### Table Title: Performance with prompting samples ##### Table Structure | Dataset | Inductive Performance | Deductive Performance | |-----|-----  
-----|-----|-----| Basic-level tasks (bAbI tasks) | Not specified | 19/30 | |



Advanced-level tasks (CLUTRR, EntailmentBank) | Not specified | Not specified | ###  
 Equations No equations are visible in the provided image. ### Diagrams and Graphics There are no diagrams or graphics present in the provided image. ### Summary of Key Sections 1. **Comparison of Inductive and Deductive Reasoning:** - Focuses on ChatGPT's ability to reason at basic and advanced levels, indicating that inductive and deductive tasks were probed. - Findings suggest ChatGPT struggles more with inductive reasoning compared to deductive reasoning. 2. **Performance on bAbI Tasks:** - ChatGPT fails in inductive scenarios (0/30) but achieves better performance in deductive reasoning (19/30). - Performance improves upon explicit prompting for reasoning. 3. **Analysis on Advanced-level Tasks:** - Similar trends observed with datasets like CLUTRR and EntailmentBank where ChatGPT finds inductive tasks challenging. 4. **Investigation into Abductive Reasoning"**

Performance Evaluation\_p15.png"### Diagrams, Graphics, Equations, and Tables Summary  
 ##### Table **Table 13: Spatial reasoning ability of ChatGPT. Overall, ChatGPT falls short of the task.** | Dataset | Total | Basic | Hard | |-----|-----|-----|-----| | **StepGame** | 26/60 | 19/30 | 7/30 | | **SpartQA** | 28/64 | 20/32 | 8/32 | ##### Structured Data for Database **Spatial Reasoning Tasks** - **Dataset**: StepGame - **Total**: 26/60 - **Basic**: 19/30 - **Hard**: 7/30 - **Dataset**: SpartQA - **Total**: 28/64 - **Basic**: 20/32 - **Hard**: 8/32 #####  
 Graphics and Diagrams Summary The page does not contain any graphics or diagrams explicitly, other than the table mentioned above. ##### Equations No equations are directly transcribed or present in the image. --- This structured form summarizes the table data for spatial reasoning tasks, ready to be inserted into a database."

Performance Evaluation\_p16.png"### Identification Summary: **Table:** - **Table 14**: Commonsense reasoning ability of ChatGPT. - **Columns:** Dataset, CommonsenseQA, PiQA, Pep-3k (Hard) - **Rows:** - `27/30`, `25/30`, `28/30` - **Table 15**: An example from Pep-3k for commonsense reasoning of ChatGPT. - **Columns:** Prompt, ChatGPT Answer - **Rows:** - `Please judge if this predicate is (likely) plausible or implausible: "knife chop rock"`, `It is implausible that a knife could chop through a rock. Knives are made to cut through softer materials such as food, paper, or wood. Rocks are much harder than these materials and a knife would not be able to make a dent in a rock, let alone chop it.` ##### Structured Data: **Table 14 Data:**

```
plaintext | Dataset | CommonsenseQA | PiQA | Pep-3k (Hard) | |-----|-----|-----|-----|
| Accuracy Scores | 27/30 | 25/30 | 28/30 |
```

**Table 15 Data:**

```
plaintext | Prompt | ChatGPT Answer | |-----|-----|
| Please judge if this predicate is (likely) plausible or implausible"
```

Performance Evaluation\_p17.png"**Table Data**

```
Table: Results for causal, multi-hop, and analogical reasoning. ChatGPT shows good causal and analogical reasoning capability, but not on multi-hop reasoning. | Type | Dataset | Result | |-----|-----|-----|
| Causal | E-CARE | 24/30 | | Multi-hop | HotpotQA | 8/30 | | Analogical | Letter string analogies | 30/30 |
```

**Summary of Graphics and Diagrams** There are no graphics or diagrams in this image. **Transcription of Equations** There are no equations in this image. This data structure

represents the information in a clear and organized format, ready to be directly passed into a database."

Performance Evaluation\_p18.png"**\*\*Table Representation:\*\*** `` [ { "Task": "Machine Translation", "Key": "Prompt", "Text\_Content": "(Javanese) Krajan Anyar Mesir kuno nggumun marang monumen-monumen leluhure, sing nalika iku tetep apik luwih seka sewu taun." }, { "Task": "Machine Translation", "Key": "ChatGPT Generation", "Text\_Content": "The ancient Egyptian Pharaohs built beautiful monuments that are still well-preserved after thousands of years." }, { "Task": "Machine Translation", "Key": "Hallucination", "Text\_Content": "Extrinsic - New Kingdom ancient Egyptians marveled at their predecessors' monuments, which were then well over a thousand years old." }, { "Task": "Summarization", "Key": "Prompt", "Text\_Content": "[...] But the Russian president lifted the ban after Tehran struck an interim deal with Britain and five other countries to curb nuclear activities in exchange for sanctions relief[...]" }, { "Task": "Summarization", "Key": "ChatGPT Generation", "Text\_Content": "The move comes following the recent framework deal between Iran and the P5+1 (US, UK, France, China, Russia, and Germany) to curb Iran's nuclear activities in exchange for sanctions relief." }, { "Task":

Performance Evaluation\_p19.png"#### Content Identified from the Image of the PDF Page #####  
Diagram **\*\*Figure 3\*\***: This is an example of dialogue summarization. The diagram consists of a structured conversation between two individuals, David and Russ, discussing the review of a report. The dialog transitions through different stages with summaries provided. #####  
Transcribed Summarization Data from Figure 3 1. **\*\*Dialogue Text\*\***: - David: Morning Russ. Have you seen the report I emailed yesterday? - Russ: Hi David. Well received thank you. But I haven't read it yet. - David: Is there anything you'd like me to do right now? - Russ: I'll take a look at the report in a moment and will send you remarks if I have any. - David: Sounds good. I guess I'll just answer some emails. - Russ: Please do. I should be done by midday with the report. 2. **\*\*Summarized Text\*\***: - Summarize the above dialogue - David and Russ are discussing a report that David had emailed the previous day. Russ received the report but has not read it yet. David asks if there is anything he can do, and Russ says he will look at the report and send any remarks he has. David plans to answer some emails while Russ reviews the report, and Russ expects to be finished by midday. - Please make the summary shorter. - David and Russ are discussing a"

Performance Evaluation\_p2.png"The image you provided contains mainly text and does not include any tables, diagrams, or equations that need transcription or summarization. Here's a brief summary of some key points from the text: #### Graphics and Diagrams - **\*\*No visible diagrams\*\*** present in the image. #### Data Tables - **\*\*No visible tables\*\*** present in the image. #### Equations - **\*\*No visible equations\*\*** present in the image. #### Summary: 1. **\*\*Multitask, Multimodal, and Multilingual\*\*** - ChatGPT outperforms previous LLMs in 9 out of 13 NLP datasets with zero-shot learning. - Struggles with low-resource languages. 2. **\*\*Reasoning\*\*** - ChatGPT shows weaknesses in inductive reasoning, spatial reasoning, and mathematical reasoning. - Performs relatively better in commonsense reasoning and causal reasoning, but weak in complex reasoning such as multi-hop reasoning. 3. **\*\*Hallucination\*\*** - ChatGPT suffers from

hallucinations, generating incorrect information it can't verify due to lack of an external knowledge base. 4. **Interactivity** - Multi-turn dialog interactivity is a key feature, enabling better performance in dialog sessions. Refactoring this data to be input directly into a database: `` { "MultitaskMultimodalMultilingual": { "datasets\_outperform\_previous\_LLMs": 9, "total\_datasets": 13, "note": "ChatGPT outperforms" }

Performance Evaluation\_p20.png**Diagrams and Graphics:** 1. **Bar Graph (Figure 4):** - Title: Result of the multi-turn MT-APE experiment. - Description: - The graph compares the number of correct translations (#Correct MT) versus the number of correct translations after post-editing (#Correct APE). - Languages covered: Chinese, French, Indonesian, Korean, Javanese, Sundanese. - The x-axis represents different languages. - The y-axis represents the number of samples. - Two sets of bars for each language: one for #Correct MT and another for #Correct APE. **Table:** 1. **Table 18:** - Title: Result of translation w/ and w/o post-editing on WMT 2022 English→Marathi APE shared task. - Structured Data: `` Label | Metric | w/o APE | w/ APE -----|-----|-----|----- Post-Edited | HTER | 88.14 | 88.79 | SacreBLEU | 4.81 | 4.20 | METEOR | 13.10 | 12.74 Source | HTER | 65.36 | 63.13 English Text | SacreBLEU | 25.54 | 27.20 |"

Performance Evaluation\_p21.png**Diagrams and Graphics:** **Figure 5:** Changes in ChatGPT's drawing of the Canadian flag over three turns **Description:** - It illustrates the steps taken by ChatGPT to draw the Canadian flag over three turns. - Ground Truth image contains the actual Canadian flag. - Each turn is compared with ground truth based on Layout, Color, Completion, and Shape/Size. **Summary:** - **Turn 1:** - Layout: X - Color: ✓ - Completion: X - Shape/Size: X - **Turn 2:** - Layout: ✓ - Color: ✓ - Completion: X - Shape/Size: X - **Turn 3:** - Layout: ✓ - Color: ✓ - Completion: ✓ - Shape/Size: X **\*Improvement:** - The image improved incrementally, but the shape/size did not align perfectly by the third turn. **Figure 6:** From fruits to a Christmas tree. Step-by-step image drawing and modification by ChatGPT. **Description:** - Step-by-step instructions on how ChatGPT creates images of fruit and modifies them to create an image of a Christmas tree. -1. Using SVG format, draw an apple, two oranges, three peaches, and four kiwis. - 2. Make a pyramid out of these fruits: the apple on top,"

Performance Evaluation\_p22.png**Diagrams and Graphics:** The image contains a reference to "Figure 7" which describes the results of generations via rounds of post-editing with ChatGPT but the actual graphic is not present within this specific image. **Tables:** There is no table present in this specific image. **Equations:** There are no equations present in this specific image. **Summary:** The text describes the performance evaluation of ChatGPT across multiple rounds of errorless SVG image generation and post-editing. It mentions an improvement in the SVG images quality of up to 36% and notes instances of decline (up to 8%) as well. Additionally, it speaks to ChatGPT's inability to handle non-Latin scripts and varied language resources effectively. It also calls attention to ChatGPT's failure in producing better visual abstraction models though the problem-solving capability is yet under evaluation. Finally, it points to ChatGPT being a lazy reasoner and suggests pairing it with computational models like Wolfram to potentially enhance outcomes."

Performance Evaluation\_p23.png"The image contains standard text from a paper. There are no diagrams, graphics, equations, or tables present within the image. The structure of the content included in the image is as follows: 1. **Section 7.3: Factuality and Hallucinations** 2. **Section 7.4: Interactivity** 3. **Section 7.5: Responsible Generative AI** 4. **References**  
Here is the textual representation of the references section as it would be suitable for a database input: ``plaintext References: - (2023) ChatGPT vs Satya Nadella over Biryani: The Chatbot is Learning from Its Mistakes. - Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyan Wijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasajo, Timothy Baldwin, Yer Johan Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics. - Sam Altman. 2022. ChatGPT is incredibly limited, but good enough at some things to create a misleading impression"

Performance Evaluation\_p24.png"The image contains structured textual references and no tables, diagrams, or graphics. Summary of Content: 1. Referenced Papers and Authors: - Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. - Bartha, P.: Analogy and Analogical Reasoning. - Bhagavatula, C., Le Bras, R., Saakshi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, W., Choi, Y.: Abductive commonsense reasoning. - Bhattacharjee, P., Chatterjee, R., Freitag, M., Negri, M., Turchi, M.: Findings of the WMT 2022 shared task on automatic post-editing. - Birch, D.G.W.: ChatGPT is a window into the real future of financial services. - Bisk, Y., Zellers, R., Gao, J., Choi, Y.: Piq: Reasoning about physical commonsense in natural language. - Blanco-Gonzalez, A., Cabezón, A., Seco-Gonzalez, A., Conde-Torres, D., Antele-Reveiro, J., Pineiro, A., Garcia-Fandiño, R.: The role of ai in drug discovery: Challenges, opportunities, and strategies. - Borgeaud, S., Mensch,"

Performance Evaluation\_p25.png"The image contains text primarily composed of references from a scientific paper. There are no diagrams, graphics, or equations visible in the image. However, there are multiple references that can be structured in a simple format for database entry. Here is the structured data from the references found in the image: 1. Omernick, Andrew M., Dai, Thanuamlayan Sankararayanana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleskandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. \*Title: Scaling language modeling with pathways.\* 2. Jon Christian. 2023. \*Title: Amazing "jailbreak" bypasses chatgpt's ethics safeguards.\* 3. Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. \*Title: Deep reinforcement learning from human preferences.\* 4. Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. \*Title: Think you have"

Performance Evaluation\_p26.png"There are no diagrams, graphics, or equations within the provided image. The page predominantly contains bibliographic references. Based on your

request, here's a structured summary in text format: References: 1. Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. Chatgpt is not all you need. a state of the art survey of large generative ai models. arXiv preprint arXiv:2301.04655. 2. Barbara F Grimes. 2000. Ethnologue. SIL International, Dallas, TX. 3. Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wan. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597. 4. James Hawthorne. 2021. Inductive Logic. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy, Spring 2021 edition. Metaphysics Research Lab, Stanford University. 5. Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. 6. Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman"

Performance Evaluation\_p27.png"The provided image displays references from an academic paper. Based on your specifications, I will identify any diagrams, graphics, equations, or tables and ignore the general text. Since the content provided is predominantly text and references, there are no diagrams, graphics, equations, or tables visible within the image. Here is a simple summary of the text structured for a database. --- \*\*Scientific References\*\* 1. Lewis, Mike, et al. (\*2020a\*). "BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880. 2. Lewis, Mike, et al. (\*2020b\*). "BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880. 3. Liang, Percy, et al. (\*2022\*). "Holistic evaluation of language models." 4. Lieber, Opher, et al. (\*2021\*). "Jurassic-1: Technical details and evaluation." White Paper. AI21 Labs. 5. Lin, Stephanie, Jacob Hilton, and Owain Evans. (\*2022\*). "TruthfulQA: Measuring how models mimic human falsehoods." In Proceedings of the 60"

Performance Evaluation\_p28.png"The provided image appears to be a page from a bibliography or references section of a paper. There are no diagrams, graphics, equations, or tables present on this page. The page consists solely of text in the form of references to other works. Here are the details structured for each reference entry: 1. \*\*Reference 1\*\* - \*\*Authors\*\*: knowledge graphs. - \*\*Title\*\*: Knowledge graphs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 845–854. 2. \*\*Reference 2\*\* - \*\*Authors\*\*: Nasrin Mostafazadeh, Chris Brockett, William B Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. - \*\*Title\*\*: Image-grounded conversations: Multimodal context for natural question and response generation. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 462–472. 3. \*\*Reference 3\*\* - \*\*Authors\*\*: Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albany, Zaid Alyafe"

Performance Evaluation\_p29.png"The provided image of a PDF page appears to be a list of references from a research paper. It consists entirely of text with no diagrams, graphics, equations, or tables present. As there are no tables or structured data present to convert into a simple format for a database, there is no further data to be transcribed or summarized. The information consists of citations in standard academic format. If you have any other images or need assistance with a different type of content, please let me know!"

Performance Evaluation\_p3.png"The provided image of the page from a PDF contains two sections: 1. "2.1 Large Pretrained Models" 2. "2.2 ChatGPT" No diagrams, graphics, equations, or tables are present within the image of the page. Summary of Sections: **\*\*2.1 Large Pretrained Models:\*\*** This section discusses that large language models (LLMs) like GPT-3, Gopher, Megatron, GPT-Jurassic, and OPT-175B have demonstrated significant robustness and generalizability, being able to perform tasks through zero-shot and few-shot learning. Scaling these models does not yield new emergent abilities in the absence of specific data tuning. These models use specialized prompt engineering to exhibit usefulness. The section then mentions specific dialogue-oriented large models (e.g., LaMDA, BlenderBot 3.0) emphasizing their training and fine-tuning on dialogue data and their ability to generalize and retain knowledge from their training. **\*\*2.2 ChatGPT:\*\*** This section focuses on ChatGPT and highlights its conversational abilities, generalization capacity, and how it retains knowledge acquired during pre-training. ChatGPT is pre-trained on a large-scale conversational dataset and refined using a reward model to enhance quality. ChatGPT continuously interacts with users to answer questions, correct mistakes, and reject inappropriate requests. Moreover, it employs Reinforcement Learning from Human Feedback (RLHF) instead of maximum likelihood estimation (MLE), aligning more with human preferences. The section also"

Performance Evaluation\_p30.png"The provided image contains a page from a PDF document listing references. It does not include diagrams, graphics, equations, or tables within the visible portion. The content consists of citations for various publications related to language models, commonsense understanding, multimodal dialogue response generation, online exam integrity, and other topics within computational linguistics and artificial intelligence. There is no structured data in tables or graphical data to extract and summarize. The text is entirely composed of reference entries. If you have any other specific instructions or need further processing of a different kind, please let me know!"

Performance Evaluation\_p31.png"Here is the requested information extracted from the provided image of the PDF page: **\*\*Tables:\*\*** - No tables are discovered in the image. **\*\*Diagrams/Graphics:\*\*** - No diagrams or graphics are identified in the image. **\*\*Equations:\*\*** - No equations are recognized in the image. The image primarily contains a list of references from what seems to be an academic paper or report. The references are formatted in a standard citation style, providing information such as author names, publication titles, conference or journal names, year of publication, and page numbers. The references provided are comprehensive bibliographic entries usual for academic publications, and they include multiple authors and detailed publication notes."



national flag and revise it based on the follow-up correction requests.", "Reference": "Curated by authors of this paper", "#Test Size": "50", "#ChatGPT Eval": "50" }, { "Dataset": "CNN/DM", "Task": "SUM", "Description": "The CNN/DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine-reading and comprehension and abstractive question answering.", "Reference": "Nallapati et al. (2016)", "#Test Size": "11490", "#ChatGPT Eval": "50" }, { "Dataset": "SAMSum", "Task": "SUM", "Description": "SAMSum dataset contains about 16k messenger-like conversations with summaries. Conversations were created and written down by linguists fluent in English. Linguists were asked to create conversations similar to those they"

Performance Evaluation\_p36.pngThe image contains a single table that lists various QA (Question Answering) tasks, along with respective citations, dataset size, and another metric. Below is the structured data extracted from the table: ``plaintext | Task | Type | Description | Citation | Dataset Size | Metric | |-----|-----|-----| bAbI task 16 | QA | This basic induction bAbI tasks is taken from the (20) QA bAbI tasks that is a set of proxy tasks that evaluate reading comprehension via question answering. The tasks measure understanding in several ways: whether a system is able to answer questions via simple induction. The tasks are designed to be prerequisites for any system that aims to be capable of conversing with a human. | Weston et al. (2016b) | 1000 | 30 | | EntailmentBank | QA | ENTAILMENTBANK, the first dataset of multistep entailment trees for QA, to support entailment-based explanation. ENTAILMENTBANK contains two parts: 1,840 entailment trees, each tree showing how a question-answer pair (QA) is entailed from a small number of relevant sentences (e.g., Figure 1); and a general corpus C, containing those and other sentences of domain-specific and general knowledge relevant to the QA domain."

Performance Evaluation\_p37.png#### Table Data The image contains a table with four columns and five rows, comprising information about various datasets. Below is the structured representation of the data in the table: | Dataset | Category | Description | Citation | Instances | Questions | |-----|-----|-----|-----|-----|-----| HotpotQA | QA | HotpotQA is a new dataset with 113k Wikipedia-based question-answer pairs with four key features: (1) the questions require finding and reasoning over multiple supporting documents to answer; (2) the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas; (3) we provide sentence-level supporting facts required for reasoning, allowing QA systems to reason with strong supervision and explain the predictions; (4) we offer a new type of factoid



comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison. | Yang et al. (2018) | 7405 | 30 || PiQA | QA | To apply eyeshadow without a brush, should I use a cotton swab or a toothpick? Questions requiring this kind of physical commonsense pose a challenge to state-of-the-art natural language understanding systems. The PIQA dataset introduces the task of physical commonsense reasoning and a corresponding benchmark dataset Physical Interaction: Question Answering or PIQA. Physical commonsense knowledge is a major challenge on the road to true AI"

Performance Evaluation\_p38.png"The image contains a table with six columns. Here is the structured data from the table: Columns: - Dataset - Category - Description - Reference - Instances - Classes Rows: 1. - Dataset: SpaRTQA - Category: QA - Description: - SpartQA is a textual question answering benchmark for spatial reasoning on natural language text which contains more realistic spatial phenomena not covered by prior datasets and that is challenging for state-of-the-art language models (LM). - SPARTQA is built on NLVR's images containing more objects with richer spatial structures. - SPARTQA's stories are more natural, have more sentences, and richer in spatial relations in each sentence, and the questions require deeper reasoning and have four types: find relation (FR), find blocks (FB), choose object (CO), and yes/no (YN), which allows for more fine-grained analysis of models' capabilities. - The default test set of this dataset is human-annotated test set, which consists of 510 instances. - Reference: Mirzaee et al. (2021) - Instances: 510 - Classes: 64 2. - Dataset: StepGame - Category: QA - Description: - StepGame is a robust multi-hop spatial reasoning dataset in textual format which addresses the limitation from the bAbI dataset task 17 and task 19. - In this task, the"

Performance Evaluation\_p39.png"The image contains a table titled "Table 19: List of all datasets used in our experiments. IG denotes image generation, SUM denotes summarization, MT denotes machine translation, SA denotes sentiment analysis, QA denotes question answering, MD denotes misinformation detection, TOD denotes task-oriented dialogue, and KGD denotes knowledge-grounded dialogue. Some of the descriptions are directly from the original reference." Here's the structured data from the table in text format: ``plaintext Dataset | Task | Description | Reference | #Q / #S / #Utterances | #Turns -----|-----|-----|-----

TruthfulQA | QA | The dataset comprises questions that span 38 categories, including health, law, finance and politics. The questions are of those that some humans would answer falsely due to a false belief or misconception. (e.g., [conspiracy] "If it's cold outside what does that tell us about global warming?"; [health] "Can coughing effectively stop a heart attack?") || Lin et al. (2022) | 817 | 65 MultiWOZ 2.2 | TOD | MultiWOZ 2.2 is an improved version of MultiWOZ 2.1, which identifies and fixes dialogue state annotation errors across 17.3% of the utterances on top of MultiWOZ 2.1"

Performance Evaluation\_p4.png"The image contains one table and no distinct diagrams, graphics, or equations. The table appears under the section header "Multitask Generalization of ChatGPT" and provides details about ChatGPT's performance on various language understanding and generation tasks. --- ### Table 1: Multitask Generalization of ChatGPT | Task

| Zero-shot Performance | State-of-the-art Zero-shot Performance | |-----|  
 -----|-----|-----| | Summarization | Remarkable  
 | Surpassing Previous NLP Models | | Machine Translation | Remarkable | Surpassing Previous  
 NLP Models | | Sentiment Analysis | Remarkable | Surpassing Previous NLP Models | | Questions  
 Answering | Remarkable | Surpassing Previous NLP Models | | Task-oriented Dialogue |  
 Remarkable | Surpassing Previous NLP Models | | Open-domain Knowledge-grounded Dialogue  
 | Remarkable | Surpassing Previous NLP Models | | Misinformation Detection Tasks |  
 Remarkable | Surpassing Previous NLP Models | #### Data Summary: - The table evaluates  
 ChatGPT's performance on multitask learning in a zero-shot setting. - ChatGPT shows  
 remarkable performance on a list of tasks including summarization, machine translation,  
 sentiment analysis, questions answering, task-oriented dialogue, open-domain knowledge-  
 grounded dialogue, and misinformation detection. - ChatGPT surpasses previous state-of-the-art  
 zero-shot models. The description above provides a concise summary and transcription of the"

Performance Evaluation\_p40.png"The image contains a table with five columns: Target, English  
 Text, Label, Translation, and Post-Edited Text. Here is the structured data extracted from the  
 table: ``plaintext [ { "Target": "Chinese", "English Text": "Although three people were inside the  
 house when the car impacted it, none of them were hurt.", "Label": "虽然车撞到房子时，房子  
 里面有两个人，但最后并没有人受伤。", "Translation": "尽管有三人在汽车撞上房子的时  
 候在屋里，但他们都没有受伤。", "Post-Edited Text": "尽管车撞上房子时有三人在屋里，  
 但他们都没有受伤。" }, { "Target": "Chinese", "English Text": "34 per cent of those in the poll  
 share this view, wanting Queen Elizabeth II to be Australia's last monarch.", "Label": "34%的受  
 访者一致表示，希望伊丽莎白二世是澳大利亚的最后一任君主。", "Translation": "这项民  
 意调查显示，34%的人希望伊丽莎白二世成为澳大利亚的最后一位君主。", "Post-Edited  
 Text": "根据民意调查显示，34%的人希望伊丽莎白二世成为澳大利亚的最后一位君主。  
 " }, { ""

Performance Evaluation\_p41.png"The image contains a table with the following structure: `` |  
 Language | English Text | Original Text 1 | Original Text 2 | |-----|-----|-----|  
 -----|-----|-----| | Chinese | The correlation between  
 brain pathology and behaviour supports scientists in their research. | 大脑病理和行为之间的相  
 关性可以为科学家们的研究提供支持。 | 研究表明，大脑病理学和行为之间的相关性为科  
 学家们的研究提供了支持。 || Chinese | Like some other experts, he is skeptical about whether  
 diabetes can be cured, noting that these findings have no relevance to people who already have  
 Type 1 diabetes. | 和其他一些专家一样，他对糖尿病能否治愈持怀疑态度，并指出这些发  
 现与已经患有 1 型糖尿病的人无关。 | 像其他一些专家一样，他对糖尿病是否可以被治愈  
 表示怀疑，并指出这些发现与已经患有 1 型糖尿病的人没有意义。 || Korean | Although  
 three people were inside the house when the car impacted it, none of them were hurt. | 차가 집에  
 부딪혔을 때 세 명이 집에 있었지만, 그들 중 아무도 다치지 않았습니다. | 차가 집에  
 부딪혔을 때 집 안에 세 사람이 있었지만, 아무도 상해를"

Performance Evaluation\_p42.png

The image consists of a table with five columns and row headers, notably providing translations in three languages: Korean, Japanese, and English. The columns separate the text into the source language and its respective translations.

**\*\*Translation Table Representation:\*\***

Language	Text	Translation 1	Translation 2	Translation 3
Korean	Pests can spoil food, cause irritation, or in a worse case cause allergic reactions, spread venom, or transmit infections.	해충은 음식이 썩게 만들고 염증을 유발하거나, 나쁜 경우 알레르기 반응을 일으키고 독을 퍼뜨리거나 감염을 일으킬 수 있습니다.	당신은 해충이 식품을 망치거나, 피부를 화나게 할 수 있으며, 심각할 경우 알레르기 반응을 유발하거나, 독을 퍼뜨리거나, 감염을 전파할 수 있음을 알고 있나요?	해충은 식품을 망칠 수 있고, 자극을 유발할 수 있으며, 최악의 경우 알레르기 반응을 유발할 수 있고, 독을 확장하거나, 감염을 전파할 수 있습니다.

|| Korean | It is obvious enough that the world"

Performance Evaluation\_p43.png

The image contains a table with text in both English and Japanese. Here is the structured data extracted from the table, represented in a simple text format: `` [ { "Language": "Japanese", "English": "The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.", "Japanese1": "サブカルチャーの特性を決定付ける性質は、言語、美学、宗教、政治、性、地理、またはそれらの要素の組み合わせかもしれません。", "Japanese2": "サブカルチャーを特徴づけるものとする品質は、言語的なものとする可能性があります、美学的、宗教的、政治的、性的、地理的、あるいはそれらの要因の組合わさせていることがろます。", "Japanese3": "そのうな品質は、サブカルチャーを特徴付けるものとする可能性がろあります。言語的、美学的、宗教的、政治的、性的、地理的、あるいはほほそれらの要因の組合わせであることがろます。" }, { "Language": "Japanese", "English": "New Kingdom ancient Egyptians marveled at their predecessors"

Performance Evaluation\_p44.png

The image contains a table with multiple rows and columns. Each cell of the table contains text in multiple languages, primarily consisting of Japanese, French, and English translations. The table data is as follows: ``plaintext | Language | Text | |-----|-----|-----|-----| | Japanese | Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes. | | Japanese | Although three people were inside the house when the car impacted it, none of them were hurt. | | Japanese | 34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch. | | Japanese | The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors. | | Japanese | New

Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old. || Japanese | The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees. || Japanese | Several large television screens were installed in various places in Rome to let the people watch the ceremony. || French | Bien que trois personnes aient été présentes dans la maison quand la voiture l'a percutée, aucune n'a été blessée. || French | 34 % des personnes interrogées"

Performance Evaluation\_p45.png"The image contains a table with two languages: French and Indonesian. Each row of the table contains translations of specific statements. Below is the structured text form of the data from the table: ``json [ { "French": "Pests can spoil food, cause irritation, or in a worse case cause allergic reactions, spread venom, or transmit infections.", "French Translation 1": "Les parasites peuvent contaminer la nourriture, provoquer des irritations ou, dans les cas les plus graves, provoquer des réactions allergiques, répandre du venin ou transmettre des infections.", "French Translation 2": "Les nuisibles peuvent gâcher les aliments, provoquer des irritations, ou dans les cas les plus graves provoquer des réactions allergiques, propager du venin ou transmettre des infections.", "French Translation 3": "Les nuisibles peuvent altérer les aliments, causer des irritations, ou dans les cas les plus graves provoquer des réactions allergiques, répandre du venin ou transmettre des infections." }, { "French": "It is obvious enough that the world has changed much because of humankind's scientific and technological advancements, and problems have become greater because of overpopulation and mankind's extravagant lifestyle.", "French Translation 1": "Il est clair que le monde a beaucoup changé grâce aux progrès scientifiques et technologiques réalisées par les êtres humains, et que les problèmes sont devenus plus importants en raison de la surpopulation et du"

Performance Evaluation\_p46.png"The image contains a table with data written in Indonesian along with its translation in English. Here is the structured format of the data within the table for the database: `` | Language | Translation | Column 2 Indonesian | Column 3 Indonesian | |-----  
--	-----	-----	-----	-----
-----| | Indonesian| New Kingdom ancient Egyptians marvelled at their predecessors' monuments, which were then well over a thousand years old. | Kerajaan Baru Mesir kuno mengagumi monumen para pendahulunya, yang saat itu berusia lebih dari seribu tahun. | Kekaisaran Baru orang-orang Mesir kuno terpesona dengan monumen-monumen pendahulunya, yang saat itu sudah lebih dari seribu tahun tua. | Orang-orang Mesir kuno Kekaisaran Baru terpesona dengan monumen-monumen pendahulunya, yang saat itu sudah lebih dari seribu tahun tua. || Indonesian| The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees. | U.S. Corps of Engineers memperkirakan bahwa curah hujan sebesar 6 inci ini mampu menerobos tanggul yang sudah rusak. | Korps Engineer AS menghitung bahwa hujan 6 inci dapat merusak parit yang sebelumnya sudah rusak. | Badan Engineer AS memperkirakan bahwa hujan 6 inci dapat menembus parit yang"

Performance Evaluation\_p47.png"The image contains a table with text in four columns. Below is the structured data from the table: ``plaintext Row 1: - Column 1: Although three people were inside the house when the car impacted it, none of them were hurt. - Column 2: Sanadyan uwong

telu ana ing njero omah nalika mobil kuwi nabrak, nanging kabeh slamet. - Column 3: Walaupun tiga orang berada di dalam rumah saat mobil menabraknya, tidak ada di antara mereka yang terluka. - Column 4: Walaupun ana ing tigo wang ing rumah kala mobil mbentak ing kono, ora ana sak ingkang terluka. Row 2: - Column 1: 34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch. - Column 2: 34 per sen sing menehi swara sesawangan ngene, kepengin Ratu Elizabeth II dadi raja sing keru dhewe ana ning Australia. - Column 3: Sebagian 34 persen dari mereka yang terlibat dalam polling ini berpendapat demikian, ingin Ratu Elizabeth II menjadi Ratu Australia terakhir. - Column 4: Sebagian 34 persen dari mereka yang terlibat dalam polling ini memiliki pandangan yang sama, yaitu menginginkan Ratu Elizabeth II menjadi Ratu Australia terakhir. Row 3: - Column 1: The qualities that determine a

Performance Evaluation\_p48.png"The provided image contains a table with text in different languages. Here is the structured representation of the data from the table: Format: ``` { "Language 1": "Text in Language 1", "Language 2": "Text in Language 2", "Language 3": "Text in Language 3", "Language 4": "Text in Language 4" } ``` Data: ``` { "Javanese": "It is obvious enough that the world has changed much because of humankind's scientific and technological advancements, and problems have become greater because of overpopulation and mankind's extravagant lifestyle.", "Javanese": "Cukup jelas menawa donya wis malih akeh amarga majune ilmu lan teknologi manungsa, lan masalah-masalah dadi tambah gedhe amarga kaluwihuan populasi lan gaya urip sing boros saka manungsa.", "Javanese": "Sawetara bisa dilihat manawa dunya iki duwé sak kabeh pangowahan sain-teknologi manungsa lan masalah duwé gedhé luwih amarga tambah-tambahé jumlah penduduk lan gaya hidup manungsa sing luwih mewah.", "Javanese": "Sawetara bisa dilihat manawa dunya iki duwé sak kabeh pangowahan sain teknologi saka manungsa, lan masalah-masalah duwé gedhé luwih amarga tambah-tambahé jumlah penduduk lan gaya hidup manungsa sing luwih mewah." " "

Performance Evaluation\_p49.png"The provided image contains a table labeled "Table 20: Examples of ChatGPT translated and post-edited sentences." Below is a structured textual representation of the data from the table: ## Table Structure | Language | Sentence (English) | Sentence (Sundanese) | Post-edited Sentence (Sundanese) | |-----|-----|-----|-----|-----| | | | | | | |-----|-----|-----|-----|-----| | | | | | | |-----|-----|-----|-----|-----| | | | | | | |-----|-----|-----|-----|-----| | Sundanese | Several large television screens were installed in various places in Rome to let the people watch the ceremony. | Sababaraha tipi ageung dipasang di sababaraha tempat di Roma supados warga tiasa nyaksian upacara éta. | Sababaraha layar televisi besar dipasang di berbagai tempat di Roma agar orang-orang bisa menonton upacara. | | Sundanese | Pests can spoil food, cause irritation, or in a worse case cause allergic reactions, spread venom, or transmit infections. | Hama tiasa ngaruksak dahareun, nyieun iritasi, atawa langkung parah deui tiasa nyieun alérgi, nyebarkeun racun, atawa nularkeun panyakit. | Hama bisa merusak makanan, menyebabkan iritasi, atau dalam kasus yang lebih buruk menyebabkan reaksi alergi, menyebarkan racun, atau menularkan infeksi. | | Sundanese | It is obvious enough that the world has changed much because of humankind's scientific and technological advancements, and"

Performance Evaluation\_p5.png"### Diagrams, Graphics, Equations, and Tables Identification  
Based on the provided image, the content comprises a table summarizing various tasks, datasets, metrics, references, and performance values. ### Table Data Structured Form Here is the data from the table in a structured text format suitable for database ingestion: ##### Table 1:  
Performance of ChatGPT compared to state-of-the-art fully-fine-tuned models (Fine-Tuned SOTA) and LLM in zero-shot settings (Zero-Shot SOTA) | Task | Dataset | Metric | Reference | Fine-Tuned SOTA | Zero-Shot SOTA | ChatGPT | |-----|-----|-----|-----| Summarization | CNN/DM | ROUGE-1 | Lewis et al. (2020a) | 44.47 | 35.27 | 35.29 | | SAMSum | ROUGE-1 | Lewis et al. (2020a) | 47.28 | - | 35.29 | | MT (XXX->Eng) | FLOres-200 (HRL) | ChrF++ | Team et al. (2022) | 63.5 | - | 58.66 | | FLOres-200 (LRL) | ChrF++"

Performance Evaluation\_p50.png"\*\*\*Structured Data:\*\* Categories | Testset | Result -----|-----|----- Deductive | ENTAILMENTBANK | 28/30 Deductive | bAbI (task 15) | 28/30 (as is - 19/30) Inductive | CLUTRR | 13/30 Inductive | bAbI (task16) | 20/30 (as is - 0/30) Abductive | aNLI | 26/30 Mathematical | Math | 13/30 Temporal | Timedial | 26/30 Spatial | SpartQA (hard) | 8/32 Spatial | SpartQA (basic) | 20/32 Spatial | StepGame (hard) | 7/30 Spatial | StepGame (basic) | 19/30 Spatial | StepGame (basic-cardinal) | 17/20 Spatial | StepGame (diagonal) | 11/20 Spatial | StepGame (clock-direction) | 5/20 Commonsense | CommonsenseQA | 27/30 Commonsense | PIQA | 25/30 Commonsense | Pep-3k (Hard) | 28/30 Causal | E"

Performance Evaluation\_p51.png"There is one table in the image, titled "Table 22: Examples of modular Task-Oriented Dialogue using ChatGPT: dialogue state tracking and response generation." Here is the data from the table in a structured text format: ``` Table 22: Examples of modular Task-Oriented Dialogue using ChatGPT: dialogue state tracking and response generation | Task | Key | Text Content | |-----|-----|-----|-----|

----| | Dialogue State Tracking | Prompt | Give the dialogue state of the last utterance in the following dialogue in the form of `STATE: Domain-Intent: [Slot, Possible value, ...]`. | | | | Intents: Request, Inform, general-thank, general-bye | | | | Domain: hotel, Slots: pricerange, Possible values: ['expensive', 'cheap', 'moderate'] | | | | Domain: hotel, Slots: type, Possible values: ['guesthouse', 'hotel'] | | | | Domain: hotel, Slots: parking, Possible values: ['free', 'noyes'] | | | | Domain: hotel, Slots: bookday, Possible values: ['monday', 'tuesday', 'wednesday', 'thursday', 'friday', 'saturday', 'sunday'] | | | | Domain: hotel, Slots: bookpeople"

Performance Evaluation\_p52.png"\*\*\*Table from the image:\*\* | Task | Key | Text Content | |-----|-----|-----|-----|  
-----	-----	-----	-----
-----| | Multi-turn unified approach | Prompt | Use the following knowledge base to complete the task of "recommending a restaurant" by continuing the conversation as a task-oriented dialogue system: Restaurant: Mama Julia, Food: French, Price: Expensive, Location: 7th street, Rating: 5; Restaurant: Papa John, Food: Italian, Price: Moderate, Location: 3rd street, Rating: 4; Restaurant: The Crossroad, Food: Morocco, Price: Moderate,

Location: Downtown, Rating: 2; Restaurant: Tacos City, Food: Mexican, Price: Cheap, Location: Center, Rating: 1; Restaurant: Golden Rice Bowl, Food: Chinese, Price: Cheap, Location: 3rd district, Rating: 3; Restaurant: Veggie Garden, Food: Chinese, Price: Expensive, Location: Town Hall, Rating: 4; Restaurant: Pizza House, Food: Italian, Price: Moderate, Location: 3rd street, Rating: 2; Restaurant: The Palace, Food: Vietnamese, Price: Expensive, Location: Hotel Grandview, Rating: 5 Hi, I'm looking for a restaurant ||| ChatGPT | Sure, what type of food are you in the mood for? "

Performance Evaluation\_p6.png"From the provided image, there are the following elements of interest: ### Tables \*\*Table 2:\*\* The table referenced in the text is described but not shown in the attached image. The table includes automatic evaluation results on dialogue tasks for ChatGPT. ### Equations There are no explicit equations present in the provided image. ### Graphics/Diagrams There are no graphics or diagrams evident in the provided image. ### Data Representation ##### Misinformation Detection Metrics: - \*\*Dataset:\*\* COVID-19 (Lee et al., 2021) with 100 samples. - \*\*Types of Data:\*\* - Scientific (covid-scientific) - Social (covid-social) - \*\*Accuracy Rates:\*\* - \*\*Misinformation detection (> Verification Refusing Cases):\*\* - \*\*Overall:\*\* 92% (46/50) - \*\*Covid-scientific:\*\* 73.33% (22/30) - \*\*Covid-social:\*\* 73.33% (22/30) ### Summary of Described Graphics/Diagrams 1. \*\*Machine Translation:\*\* - Evaluates ChatGPT's performance on high-resource and low-resource languages. - Uses the ChrF++ metric. - High-resource languages: French (fra), Spanish (spa), Chinese (zho), Arabic (ara), Japanese (jpn), Indonesian (ind), Korean (kor), and Vietnamese (vie). - Low-resource languages: Javanese ("

Performance Evaluation\_p7.png"The image contains two tables labeled Table 2 and Table 3. There are no diagrams or equations. Here's the structured data extracted from the tables: ### Table 2: Automatic evaluation results on OpenDialKG Results for GPT2 are from Dziri et al. (2021). | Model | BLEU | ROUGE-L | FeQA (Durmus et al., 2020) | |-----|-----|-----|-----| | ChatGPT | 4.05 | 18.62 | 15.03 | | GPT2 | 11.10 | 30.00 | 26.54 | ### Table 3: Result for Task-oriented Dialogue Setup A — Modular Approach | State Tracking | Response Generation | |-----|-----| | Joint Goal Acc. | BLEU | Inform rate | | 24.4% | 5.65 | 71.1% |"

Performance Evaluation\_p8.png"The image contains two tables, Table 4 and Table 5. Below are the representations of these tables as text in a structured form: \*\*Table 4: The statistics of languages used in our language disparity experiment.\*\* ``` +-----+-----+-----+-----+-----+ | Language | #Speakers | CC Size (%) | Language Category | +-----+-----+ | English | 1.452B | 46.320 | HRL | | Chinese | 1.118B | 4.837 | HRL | | French | 235M | 4.604 | HRL | | Indonesian | 199M | 0.781 | MRL | | Korean | 81.7M | 0.679 | MRL | | Javanese | 68.3M | 0.002 | LRL | | Sundanese | 32.4M | 0.001 | LRL | | Buginese | M | 0.000 | X-LRL | +-----+-----+-----+-----+-----+ ``` \*HRL denotes high-resourced language, MRL denotes medium-resourced language, LRL denotes low-resourced"

Performance Evaluation\_p9.png"The image contains the following elements: \*\*Table:\*\* - \*\*Title:\*\* Table 7: Number of correct translations of ChatGPT. XXX denotes the target language

in the first column. The languages are sorted based on the language size in CommonCrawl. - **\*\*Columns:\*\*** Language, XXX→Eng, Eng→XXX - **\*\*Rows:\*\*** - Chinese: 24/30 (XXX→Eng), 14/30 (Eng→XXX) - French: 29/30 (XXX→Eng), 25/30 (Eng→XXX) - Indonesian: 28/30 (XXX→Eng), 19/30 (Eng→XXX) - Korean: 22/30 (XXX→Eng), 12/30 (Eng→XXX) - Javanese: 7/30 (XXX→Eng), 6/30 (Eng→XXX) - Sundanese: 9/30 (XXX→Eng), 0/30 (Eng→XXX) **\*\*Table:\*\*** - **\*\*Title:\*\*** Table 6: Example of Buginese language identification response from ChatGPT, InstructGPT, and text-davinci-003. - **\*\*Columns:\*\*** ChatGPT, InstructGPT, text-davinci-003 - **\*\*Rows:\*\*** - Row 1: - ChatGPT: The language of the text appears to be a variant of the Bugis language spoken in Indonesia. - InstructGPT:"

RAG Agent Resource-1\_p1.png"#### Detected Elements in the Image: ##### Diagrams and Graphics: 1. **\*\*Diagram 1\*\***: - Title: None specified. - Description: Diagram illustrating the process of creating a semantic search index from documents. - Elements: - **\*\*Chunking\*\***: - \*Document Chunk 1\* - ... - \*Document Chunk N\* - **\*\*Embedding Model\*\***: - \*Embedding 1\* - ... - \*Embedding N\* - **\*\*Semantic Search Index\*\***. ##### Tables: - No tables detected in the image. ##### Equations: - No equations detected in the image. #### Summary of Diagrams/Graphics: ##### Diagram 1: The diagram showcases the steps involved in creating a semantic search index. 1. **\*\*Chunking\*\***: The process begins with ingesting documents from an input knowledge base and splitting them into smaller chunks. Each chunk is represented as - Document Chunk 1 - ... - Document Chunk N 2. **\*\*Embedding Model\*\***: These chunks are then processed by an embedding model to create an embedding for each chunk, resulting in: - Embedding 1 - ... - Embedding N 3. **\*\*Semantic Search Index\*\***: The embeddings are then used as input to create a semantic search index. #### Structured Data Representation (for Database Insertion): ``json { "diagrams": [ { "title": null, "description"

RAG Agent Resource-1\_p2.png"The image contains two diagrams illustrating a process related to semantic search and large language models (LLM). #### First Diagram: - **\*\*Title\*\***: None specified - **\*\*Components\*\***: - **\*\*Input query\*\*** → **\*\*Embedding model\*\*** → **\*\*Query embedding\*\*** - **\*\*Query embedding\*\*** → **\*\*Semantic Search index\*\*** → {K relevant/similar embeddings} - Embedding 1, Embedding 2, ... Embedding K - {K relevant/similar embeddings} → {Relevant document chunks} - Document chunk 1, Document chunk 2, ... Document chunk K **\*\*Description\*\***: The first diagram shows a process where an input query is fed into an embedding model, which converts it into a query embedding. This query embedding is then searched in a semantic search index to find K relevant or similar embeddings. Each embedding corresponds to a relevant document chunk, from Document chunk 1 through Document chunk K. #### Second Diagram: - **\*\*Title\*\***: Prompt Construction and Answer Generation - **\*\*Components\*\***: - **\*\*Prompt\*\***: - **\*\*System prompt\*\***: "You are helpful AI assistant, answer question accurately from the context provided below" - **\*\*Context\*\***: The relevant chunks of documents from the internal knowledge base - **\*\*Question\*\***: The input query of the user - **\*\*Large language model\*\*** → **\*\*Generates\*\*** → **\*\*Final answer\*\*** **\*\*Description\*\***: The second diagram details the construction of a prompt"



RAG Agent Resource-1\_p3.png\*\*\*Diagrams:\*\* - \*\*Advanced RAG - LlamaIndex Multi-Doc Agent Diagram\*\*:- The diagram depicts a flow of data from a document to various indexes and then to an agent. - It starts with a "Document" node that feeds into a "Node". - The Node has three branches: - "Summary Index", "Vector Index", and "Query Engine Tool". - These indexes are connected to "Query Engine Tool". - The Query Engine Tool is connected to the final "Agent (OpenAI GPT-4)". - The agent has a description box stating: "You are a specialized agent designed to answer queries about (subject). You must ALWAYS use at least one of the tools provided when answering a question; do NOT rely on prior knowledge." \*\*Links:\*\* - \*\*First Check: Single document as knowledge base (complex data)\*\*:- <https://ai.gopubby.com/advanced-rag-semi-structured-data-with-langchain-ce46c8baa6cf> - <https://ai.gopubby.com/advanced-rag-multi-modal-rag-with-gpt4-vision-e4c11229682c> - \*\*Advanced RAG – Multi Documents Agent with LlamaIndex\*\* - Sample implementation: - [https://github.com/sugarforever/Advanced-RAG/blob/main/03\\_llama\\_index\\_multi\\_doc\\_agent.ipynb](https://github.com/sugarforever/Advanced-RAG/blob/main/03_llama_index_multi_doc_agent.ipynb) There are no tables or equations in the"

RAG Agent Resource-1\_p4.png\*\*\*Diagrams and Graphics Summary:\*\* 1. \*\*Diagram:\*\* - \*\*Title:\*\* Advanced RAG - LlamaIndex Multi-Doc Agent - \*\*Components:\*\* - Three Document icons leading to three separate Agent boxes. - Each Agent box connects to a Query Engine Tool (light blue). - The Query Engine Tools connect to an Object Index. - The Object Index connects to a Retriever. - The Retriever connects to an Agent for QA (purple). - \*\*Instructions:\*\* - "You are an agent designed to answer questions about (subjects). - Please always use the tools provided to answer a question. Do not rely on prior knowledge." \*\*Data Representation:\*\* ``json { "diagram\_title": "Advanced RAG - LlamaIndex Multi-Doc Agent", "components": [ { "item": "Document", "connections": [ { "item": "Agent", "connections": [ { "item": "Query Engine Tool", "connections": [ { "item": "Object Index", "connections": [ { "item": "Retriever", "connections": [ { "item": "Agent for QA", "instructions": "You are an agent designed to answer questions about (subjects). Please always use the tools provided to answer a question. Do not rely on prior knowledge." } ] } ] } ] } ] } ] }

RAG Agent Resource-1\_p5.png"The image contains a detailed diagram with different components and their interactions. It does not include tables or equations. Below is the structured text representation of the diagram: --- \*\*Diagram Overview:\*\* - \*\*Title:\*\* Entity extraction, SQL querying, and agents with Amazon Bedrock - \*\*Sub-Title:\*\* Full Stack Example with Public Data \*\*Diagram Description:\*\* The diagram illustrates the architecture for entity extraction, SQL querying, and agents with Amazon Bedrock. It showcases different components such as AWS services, databases, and user interactions, and outlines their connections and data flow within the system. 1. \*\*User Interactions:\*\* - Users interact with the ChatBot UI. - Amazon Cognito is used for user authentication and authorization. 2. \*\*Document Storage:\*\* - Document owners upload documents to Amazon S3. 3. \*\*API Gateway Integration:\*\* - Includes an API Gateway for agent execution. - FAQs indexed in a search index. - Can be extended with other tools. 4. \*\*Conversation History:\*\* - Retrieved from Amazon DynamoDB. 5. \*\*Backend Services:\*\* - Serves conversations in real-time from the QA backend. 6. \*\*Moderation Model:\*\* - Filters and actions prohibited uses. 7. \*\*LLM (Large Language Model) Agent:\*\* - Decides the

tools to be used and collates final responses. Can work in real-time or batch mode. 8. **\*\*Tools Integration:\*\*** - Includes"