

The Interactive Deep Learning Enterprise (No-IDLE) meets ChatGPT

DANIEL SONNTAG, THIAGO S. GOUVÊA, MICHAEL BARZ, ALIKI ANAGNOSTOPOULOU, SITING LIANG, SARA-JANE BITTNER, and FRANZISKA SCHEURER, German Research Center for Artificial Intelligence (DFKI), Germany

This DFKI technical report presents the anatomy of the No-IDLE meets ChatGPT prototype system (funded by the German Federal Ministry of Education and Research) that provides not only basic and fundamental research in interactive machine learning, but also reveals deeper insights into how to leverage the opportunities arising from large language models and technologies for the No-IDLE project. No-IDLE's goals and scientific challenges centre around the desire to increase the reach of interactive deep learning solutions for non-experts in machine learning. No-IDLE aims to enhance the interaction between humans and machines for the purpose of updating deep learning models, integrating cutting-edge human-computer interaction techniques and advanced deep learning approaches. Considering the recent advances in LLMs and their multimodal capabilities, the overall objective of "No-IDLE meets ChatGPT" should be well motivated. One of the key innovations described in this technical report is a methodology including benchmark studies for interactive machine learning combined with LLMs which will become central when we start interacting with semi-intelligent machines based on optimisation methods like automatic prompt engineering or natural language inference. Our main research question is how ChatGPT and other variants can help improve the accuracy of (semi-) automatic subtasks in image retrieval, captioning, and person/scene recognition.

1 INTRODUCTION

With the convergence of AI and machine learning (ML), IML is where the human-computer interaction (HCI) community meets the ML community, with contributions from related fields such as cognitive science, computer graphics, design or the arts, and natural language processing (NLP), data mining, knowledge representation, and reasoning. The goal of the main project No-IDLE [Sonntag et al. 2024], which stands for Interactive Deep Learning Enterprise, is to improve the interaction between humans and machines to update deep learning (DL) models, by leveraging both state-of-the-art human-computer-interaction and DL approaches. While HCI constitutes a key approach, we will address the Interactive Machine Learning (IML) problem from multiple angles. Informed by emerging directions in both research and commercialisation of IML systems [Oviatt et al. 2019; Zacharias et al. 2018], we will deploy our expertise in multimodal-multisensor interfaces (MMI) and NLP, while also tapping on the broader interdisciplinary community, to deliver on the mission to improve interaction between humans and machines.

This proposal describes an add-on to the No-IDLE project. In NO-IDLE MEETS CHATGPT, we bring No-IDLE, ChatGPT and other large language models and technologies together to apply them to interactive photobook creation (No-IDLE use case), to test and evaluate basic and fundamental research of ChatGPT inclusion. Consider the natural language input "On the first day, we took the bus from the airport to Vancouver" (see figure 1). As a response, the system creates a single page with suitable photos, i.e., from getting on the bus at the airport, a photo of the skyline of Vancouver from inside the bus, and one with aunt Mary who was waiting for them at the bus stop. Our main research question is how ChatGPT and other variants can help improve the accuracy of (semi-) automatic subtasks in image retrieval, captioning, and person/scene recognition.

Authors' address: Daniel Sonntag, daniel.sonntag@dfki.de; Thiago S. Gouvêa, thiago.gouvea@dfki.de; Michael Barz, michael.barz@dfki.de; Aliko Anagnostopoulou, aliki.anagnostopoulou@dfki.de; Siting Liang, siting.liang@dfki.de; Sara-Jane Bittner, sara-jane.bittner@dfki.de; Franziska Scheurer, franziska.scheurer@dfki.de, German Research Center for Artificial Intelligence (DFKI), Oldenburg & Saarbrücken, Germany.

Large language models (LLMs) have gained a lot of attention recently, both in industry and research, especially with the release of various GPT models, such as GPT-3 [Brown et al. 2020] and ChatGPT. LLMs have billions of parameters and are mostly trained on large amounts of (generic) data. Given an input text prompt, these models generate an output answer, ranging from a single word to multi-paragraph texts. Recently, GPT-4 [OpenAI 2023] was released, which, in contrast to its predecessors, is multimodal, meaning that it can take visual input in addition to textual prompts, which fits very well to the research and AI transfer questions of the photobook use case. Currently, there are early works on various LLMs and their inclusion into No-IDLE-like complex (interactive) AI systems for better commonsense reasoning and scene understanding, especially for GPT-4 [Bubeck et al. 2023].

Automatic prompting refers to an automated method to create LLM prompts for a diverse set of tasks. Here, we aim to explore, first and foremost, automatic prompting behaviour in the photobook use case context based on the most recent literature relevant to interactive deep learning. Specifically for prompting, different patterns are addressed by [Ouyang et al. 2022], who train language models to follow instructions with human feedback, and [White et al. 2023], who propose a prompt pattern catalog to enhance prompt engineering with ChatGPT. [Moslem et al. 2023] experiment with prompts for adaptive machine translation. [Feng et al. 2023] investigate if text generation models can act as clinical knowledge bases, while [Sallam 2023] reviews the utility of ChatGPT in healthcare education, research, and practice. According to [Schick et al. 2023], language models can teach themselves to use external tools via APIs. [Wu et al. 2023] present an interactive system for text transformation, which results in better in-context learning capabilities of LLMs. In addition to ChatGPT, other large language models and their usage in No-IDLE should be investigated. ChatGPT, Bard, and Claude are able to respond to a breadth of user queries, including to provide sample code. Training the most advanced LLMs demands immense computational resources and often relies on extensive, proprietary datasets. We will obtain and use academic licences for additional models similar to ChatGPT and report on their utility in the No-IDLE photobook context. We will explore open source models and include research tasks on [natural language inference \(NLI\)](#) (textual entailment) and other forms of commonsense reasoning with LLMs for improving the applications tasks, namely image retrieval, captioning, and person/scene recognition. Related tasks include how does ChatGPT behave in the photobook use case for cross-modal vision-language retrieval, in user-model interaction, and visual storytelling to create a photobook interactively.

2 USECASE: INTERACTIVE PHOTOBOK CREATION

The research questions raised in No-IDLE MEETS CHATGPT will be investigated in the context of the use case of the main project, No-IDLE: the interactive creation of a photobook. We briefly recap the use case and outline how No-IDLE MEETS CHATGPT can extend it. Consider the following scenario:

Family Smith (a family of four) takes many photos from all kinds of events and occasions and regularly likes to create personal photobooks and calendars for themselves and as gifts for family members and friends. Selecting the right photos, arranging them, and writing captions is fun but very time-consuming, and while they appreciate it as a means of their personal expression and creativity, they would like to speed up the process, especially with respect to the more tedious parts like selecting among similar photos or finding a basic arrangement. At the same time, they would like to maintain control and a personal connection to the results. Each family member has their own personal taste: Some are more inclined to funny situations and photos of people, others prefer scenic views and interesting lighting and their personal style of arrangement, some like to put the photos simply side by side, others like to make use of interesting frames, clip art, and creative arrangements. In addition, the goal and target audience influence their choices. For example, they enjoy creating diary-style photobooks of their travels for personal archiving, while crafting

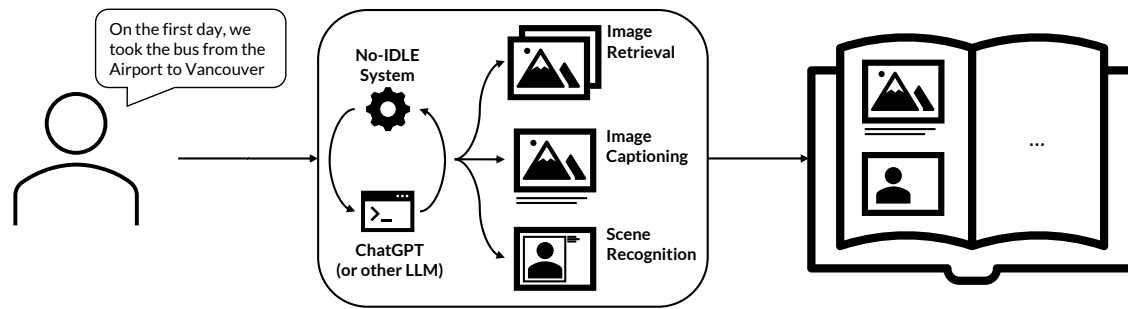


Fig. 1. Deep learning models are used to create photobook pages from natural language input. These modules include, for instance, image retrieval, image captioning, and scene recognition. In this extended use case, we explore LLMs, mainly ChatGPT, to assist in these tasks.

image-based stories of the same trip to share or gift to others. When making photobooks or calendars for holidays or birthday presents, they tend to select photos that not only fit the occasion but, if possible, also include the recipient.

Thankfully, they find out about the AI software that integrates techniques developed within No-IDLE and No-IDLE MEETS CHATGPT. Using these, a photobook can be created by providing a set of images and by sequentially describing the occasion in natural language, be it a holiday trip or a wedding party. They can also describe the style and purpose of the photobook to guide the creation process. To give an example, imagine that they plan to create a photobook about their last family trip to Canada. They start off by telling the system: “This will be a photobook for aunt Mary about our last trip to Canada. We would like to add some dramatic touch to it”. In return, the photobook creation tool suggests a suitable caption and basic style for the photobook. If not suitable, they can edit the caption or adapt the style, e.g., by selecting another frame type for captions or another font family. They would continue by describing how they perceived their vacation to the photobook tool just like they would describe it to another human: “On the first day, we took the bus from the airport to Vancouver” (see figure 1). As a response, the system creates a single page with suitable photos, i.e., from getting on the bus at the airport, a photo of the skyline of Vancouver from inside the bus, and one with aunt Mary, who was waiting for them at the bus stop. Since this is the first time family Smith is using this tool, the automatic caption generation module is uncertain whether its output is suitable and, hence, actively asks for feedback.

Over the past decades, researchers have studied similar scenarios [Sandhaus et al. 2008] and proposed partial solutions for certain sub-tasks. For instance, different methods ranging from semantic modelling [Sandhaus and Boll 2011] and meta data analysis [Boll et al. 2006, 2007] to deep learning solutions [Withöft et al. 2022] have been investigated for retrieving and filtering photos according to general criteria or personal preferences [Maszuhn et al. 2021]. Some of these works have also looked at data from social media activity to learn about user preferences or events [Rabbath et al. 2011a,b]. Other works have looked at the presentation layer, for instance, at how to create aesthetic layouts [Sandhaus et al. 2011] or how to design novel augmented reality interaction techniques to allow users to easily annotate their photos [Henze and Boll 2011]. However, integrated solutions for a complete system are still missing, which highlights both the relevance but also the challenge of the presented scenario. While the goal of this project is not to develop a market-ready photobook application software, we are certain that we will be able to implement the use case as an AI testbed to extend the current state-of-the-art in interactive deep learning. We propose a unique and integrated approach that draws on our expertise from machine learning, NLP, multimodal interaction, and HCI research.

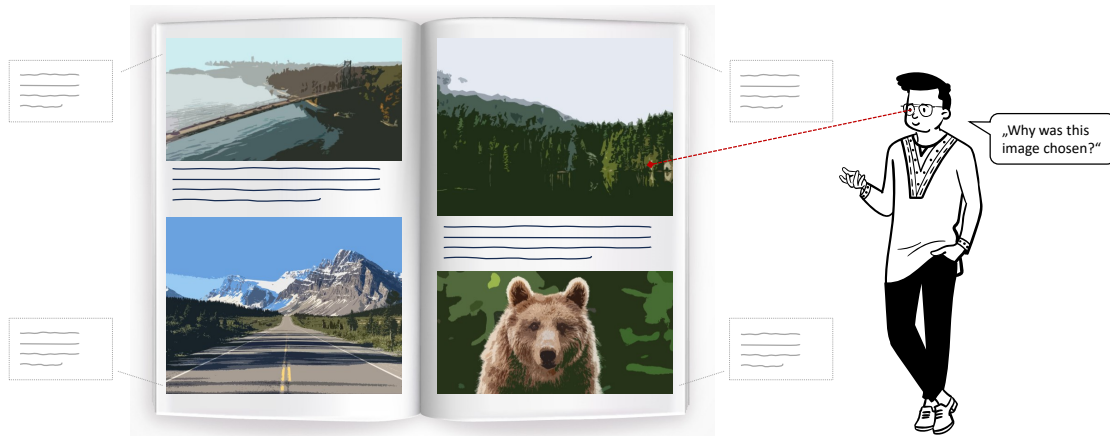


Fig. 2. Example of a multimodal user input to our photobook application. A double page for the photobook has been created by our system with a certain layout and one caption per page that took the meta information of each image into account.

In the following, we refer to ChatGPT powered by the language model GTP-3.5 and GPT-4.0 as ChatGPT-3 and ChatGPT-4, respectively, if the distinction is necessary. In NO-IDLE MEETS CHATGPT, we aim at extending the original use case using LLMs like ChatGPT. This can be done in multiple steps along the photobook creation process, mainly to assist in image retrieval, image captioning, and scene recognition (figure 1, middle part). For example, retrieving images and generating individual captions could benefit from using prompts to ChatGPT. Image captions could also be personalised to the user and the event in this way. One example would be to generate a prompt to an LLM that asks the model to substitute subjects and objects in a caption with named entities (see figure 9). Implicit personalisation can also be achieved by considering the user’s gaze signal to infer important entities that should be highlighted in a generated story (visual storytelling). Furthermore, prompts could be used to generate more consistent stories, e.g., by providing shared context information for image captioning or visual storytelling, or by taking multiple images and their meta information into account to create a single story description (see figure 2). LLMs can provide world knowledge information, in order to further personalise the captions. Last but not least, the layout could be adapted by a multimodal prompt to ChatGPT-4. An integral part, however, would be to explain why a particular image has been chosen, or a particular text has been generated by ChatGPT. An example dialogue is shown in figure 3.

3 GOALS AND SCIENTIFIC CHALLENGES OF NO-IDLE MEETS CHATGPT

Our primary goal is to share our exploration of ChatGPT’s capabilities and limitations in support of our assessment that a technological leap has been achieved that can be used in multimodal multi-sensor AI applications, intelligent user interface (IUI) frameworks in particular. We believe that especially ChatGPT-4’s intelligence signals a true paradigm shift in the field of IUI that can be used at the frontend or the backend of a complex multimodal interaction and multimedia information system such as the one developed in No-IDLE.

NO-IDLE MEETS CHATGPT’s goals and scientific challenges centre around the desire to increase the reach of GPT-4 solutions in highly individualised tasks like photobook creation. In addition, to fully automate tasks in practical applications such as our use case of interactive photobook creation can be extremely difficult and even undesirable. As

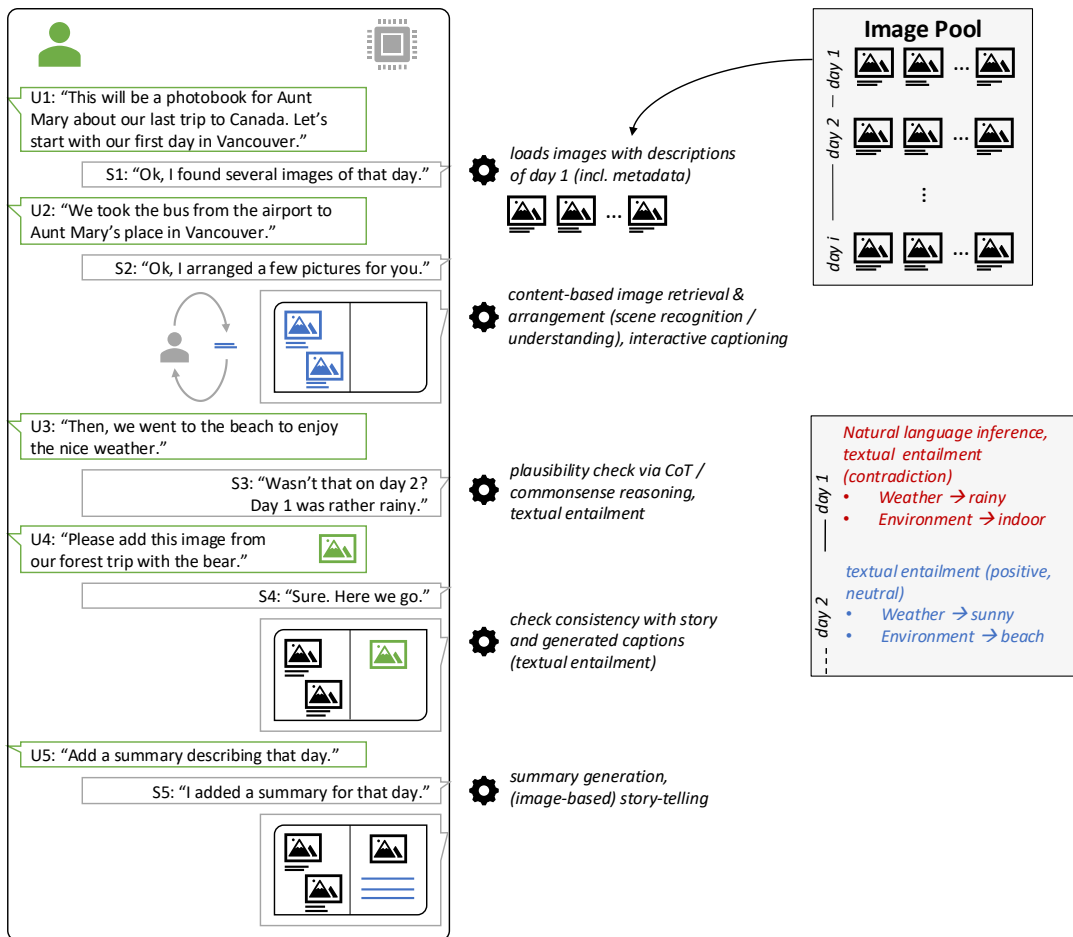


Fig. 3. This example dialogue shows the integration of the NLP components around prompt engineering for incorporating LLMs. The gears represent an integrated prompt creation and LLM usage process along the technical tasks of captioning, natural language inference, scene recognition, and visual storytelling. For example, the U4/S4 dialogue pair results in a textual entailment task to check for consistencies between the user-provided information, the automatic captioning for the provided image, and the captions of the other images and/or the story so far.

a consequence, our goals are to find a computational and design methodology to gracefully combine a new automated service based on ChatGPT with direct dialogical user input or with a direct manipulation system. We declare our scientific goals in the context of our photobook application. However, the technologies developed shall be beneficial for other domains as well such as healthcare, smart manufacturing, or environmental management. They can be summarised as follows:

- (1) Intuitively, fairness, accountability, and transparency in ML are desired goals, but they have been difficult to study from a scientific perspective for recent developments, in DL in particular. Actually, the best known approaches emphasise big data scalability instead of understanding or reasoning. We plan to address this

fundamental problem in NO-IDLE MEETS CHATGPT in a controlled way. While using ChatGPT-4 in a very restricted way as an internal service, this may allow us to deal with the problem of hallucination.

- We will follow the new "LLM-Human-in-the-Loop pipeline" approach [Chen et al. 2023b] and motivate to establish a new benchmark dataset that provides annotation of the fine-grained reasoning steps to automatically measure the reasoning consistency while reducing hallucinations. Hence, the application and evaluation part of our system is strengthened by a ChatGPT-4-powered database construction part (see section 4.1.2)
 - We provide in-depth research on the applicability of Chain-of-Thought (CoT) based on an extensive list of subtask benchmarks. We evaluate the possibility control of this problem by (1) the "LLM-Human-in-the-Loop pipeline", (2) by strengthening the CoT self-critic by providing a possibility for it to reflect on its own by our supervised textual entailment task that can identify hallucinations as contradictions, and (3) by exploring new effective methods against LLM hallucination, namely deductive verification of CoT reasoning.
- (2) Define and declare the role of humans in embedded ChatGPT applications: (1) realising the importance of studying users; (2) reducing the need for supervision by ML practitioners; (3) explore interactivity in a tight coupling between the system and the user; (4) handle human ambiguity and confusion and instill trust and confidence through feedback and explanations; provide a way for users to understand why the system had made a particular prediction, and adjust the (DL) learner's reasoning (and ChatGPT's for that matter) if its prediction was wrong.
- We evaluate the controlling of the commonsense reasoning abilities of LLMs in the context of the photobook creation tool.
 - We implement two user-centric HCI tasks (section 3.3 on page 24) that focus on developing new forms of interaction with LLM-enhanced applications based on HCI principles such as direct manipulation.
 - The question is whether these tasks capture distinct aspects of the photobook use case, namely image retrieval and storytelling, and encourage users to leverage the accuracy, creativity, and commonsense reasoning abilities of LLMs while exerting control of the created content.
- (3) Evaluating LLMs in our application context: We propose an integrated evaluation approach to assess commonsense reasoning capabilities by NLI tasks.
- We aim to probe ChatGPT's responses and behaviours, to verify its consistency, coherence, and correctness, and to uncover its limitations and biases. How can LLMs best contribute to image retrieval, image captioning, scene recognition, and storytelling.
 - Fine-grained scientific goals and challenges include: (1) how effectively can LLMs perform NLI tasks to facilitate entailment assessment and ensure logical consistency between descriptions generated by LLMs from different perspectives in the self-verification loop?; (2) how grounded and hallucination-free are (storytelling) descriptions generated by an LLM? (3) can an LLM extend self-verification to user input verification and help identify inconsistencies in it?

3.1 Natural Language Processing (NLP)

3.1.1 ChatGPT preliminaries. As mentioned earlier, LLMs have attracted significant attention in both industry and research. With billions of parameters, they generate outputs based on text prompts. GPT-4 [OpenAI 2023], unlike its predecessors, is multimodal, processing both text and visual inputs. LLMs can solve purely text-related tasks, and

ChatGPT-3: Prompt pattern "Fact Check List"
Q: When generating a caption, please provide a list of objects at the end, which are relevant for your answer. <i>In the image, there is a table with a carrot, an orange, a banana, and a knife.</i>
A: Caption: "A colourful assortment of fresh fruits and a sharp knife arranged neatly on a wooden table." List of objects: 1. Table 2. Carrot 3. Orange 4. Banana 5. Knife

Table 1. Fact check list for image captioning improvement. Based on [White et al. 2023].

their users can explore fields such as coding, medicine, vision, and more. ChatGPT-4, in particular, often exceeds the capabilities of earlier models. For this reason, it is sometimes viewed as an early precursor of an artificial general intelligence (AGI) system [Bubeck et al. 2023]. There are still many limitations and problems, such as hallucinating, i.e., inventing untrue content, or making mistakes in simple tasks [Cobbe et al. 2021; Maynez et al. 2020].

3.1.2 Prompt patterns. As a guiding thread, we will adhere to prompt patterns identified by [Ouyang et al. 2022] and [White et al. 2023] that we adapt to our purposes. Identifying such patterns is an important aspect of prompt engineering. It can help optimise results obtained from the LLM by paying attention to the user's exact purposes with their respective questions. For example, the pattern "Fact Check List" by [White et al. 2023] can help identify potential error sources when used in image captioning. Similarly to the approach of Chain-of-Thought prompting [Wei et al. 2022b], it breaks the answer down into the individual components that are of importance, as seen in table 1. The user asks for a list of objects that lead to ChatGPT-3's answer. In a more comprehensive and complex example, such a list can help narrow down troubleshooting attempts to one specific aspect if one of the objects in the list does not appear in the image, for example. With ChatGPT-4, the output is similar, but can also be written in a more poetic tone ("a fresh medley of fruit [...] artfully displayed [...] for a healthy preparation"), for example.

In the following tables 2 to 6, we show a selection of other prompt patterns identified by [White et al. 2023] and [Ouyang et al. 2022] for this project, with examples from medicine and the XAI domain.

Prompt pattern "Classification"
a) List and categorise all objects present in the image.
b) List and categorise all lesions in the medical image.
c) Caption and categorise all the input images according to ...

Table 2. Categorisation tasks for medical image captioning. Based on [Ouyang et al. 2022].

3.1.3 Comparison between ChatGPT-3 and ChatGPT-4. The currently available free version of ChatGPT is based on the LLM GPT-3.5. ChatGPT-4 is a more advanced model than ChatGPT-3.5, which shows in several ways. GPT-3.5 has 175 billion parameters, whereas the number of parameters of GPT-4, while not publicly known, is estimated to be between 1.7 trillion and more than 100 trillion [Howarth 2024; Patel and Wong 2023; Young 2021]. The token limit of ChatGPT-3.5 is 4,096, and that of ChatGPT-4 goes up to 32,768. For ChatGPT-4, we use a token limit of approximately

Prompt pattern "QA, closed"
Which of the following diseases is frequent among teachers?
a) Burnout
b) Constipation
c) Dementia

Table 3. Q&A in the medical domain. Based on [Ouyang et al. 2022].

Prompt pattern "Reflection"
Explain the reasoning behind the captions generated.

Table 4. Reflecting on given answers and providing explanations for them. Based on [White et al. 2023].

Prompt pattern "Chat"
Act as a medical advisor. I am a patient asking questions about my health. The format should be a chat between the advisor and the patient.

Table 5. Chat with detailed instructions. Based on [Ouyang et al. 2022].

Prompt pattern "Cognitive Verifier"
When I ask a medical question, please ask one to three questions that help you create better output. Combine all the answers for the final answer to the original question.

Table 6. Prompting ChatGPT to ask questions for better output. Based on [White et al. 2023].

8,000 in our experiments, for which we have a commercial licence. Due to its more advanced features, ChatGPT-4 can perform significantly better in various areas and for several tasks. For example, a lower risk of hallucinations and a significant improvement in multilingual performance have been found. One of the most noticeable advancements is the multimodal ability of ChatGPT-4, which supports text input and images as multimodal input (see, for example, [Koubaa 2023; OpenAI 2023]). A constraint of both models is missing data timeliness, as the data reaches no further than January 2022 for ChatGPT-3.5, and April 2023 for ChatGPT-4. Other text data quality dimensions are yet to be explored [Sonntag 2004].

GPT models make use of various chat roles during interaction with the user. Besides the user and assistant role, which corresponds to the human user and the chat assistant, there is an additional *system* role, allowing the system developer to give internal instructions for the conversation. As a first empirical result towards defining the NLP work packages in this proposal, we compared ChatGPT-3.5 and ChatGPT-4 in the integrated image captioning task of the photobook use case (table 7, with the same user prompt and varying system instructions).

3.1.4 Prompt engineering and automatic prompting. Prompt engineering is the task of creating an input prompt such that the LLM has the best chance of generating the desired output. It is an empirical science that studies how different prompting strategies can be used to optimise output LLM performance. Although a variety of approaches exist, the question is, first, how effective prompt engineering can be done (see, for example, [Zhou et al. 2023a]), and (2) how the process can be optimised. As pointed out by [White et al. 2023], prompt patterns are a knowledge transfer method

User: "Here is a caption for an image: a woman is petting a black and white cat. Please rewrite it for a photobook, given the following information: - Cat name: Zizel - Person name: Aliko - Date: Dec 2021"		
System instruction	ChatGPT-3.5 answer	ChatGPT-4 answer
You are a helpful assistant.	Zizel, the beloved companion of Aliko, receives gentle affection in the cozy embrace of friendship. December 2021.	Zizel, the black and white feline, enjoys tender strokes from Aliko, December 2021.
You are an unhelpful assistant.	A woman named Aliko embraces the enchanting feline charm of Zizel during a heartwarming encounter in the cozy winter of Dec 2021. But hey, who needs words when you have a picture?	Aliko with some random black and white animal named Zizel in Dec 2021.
You are an assistant who writes funny texts.	Aliko and Zizel: Discovering the Secrets of Feline Mischief - Dec 2021	Meet the lovely Aliko, caught in the act of showering affections on the striking, black and white feline, Zizel. A moment caught in time in December 2021, capturing their purr-fectly beautiful bond.
You are an assistant who only gives happy answers.	A joyous moment captured in Dec 2021, as Aliko shares a heartwarming connection with her beloved cat, Zizel.	Joyful Moment Captured in Dec 2021: Aliko Lovingly Stroking Her Adorable Black and White Feline, Zizel.
You are an assistant who only gives sad answers.	Moments of fleeting joy captured in sorrow: Aliko caresses Zizel, her only solace in a heart-wrenching December 2021.	Aliko, captivated by a profound solitude, tenderly strokes Zizel, the black and white cat in December 2021. Their shared silence screams a thousand words of longing and unspoken goodbyes.

Table 7. ChatGPT-3.5 vs. ChatGPT-4's performance on modifying a caption with an instruction. In this case, the user input stays the same, while we change the system instruction (as seen in the first column). However, please note that "capturing their purr-fectly beautiful bond", and "Their shared silence screams a thousand words of longing and unspoken goodbyes." are hallucinations we aim to control with multimodal input or filtering mechanisms.

analogous to software patterns since they provide reusable solutions to common problems faced in a particular context, i.e., output generation and interaction when working with LLMs.

Most prompts are comprised of the same few (optional) components: Input Data: This is the actual data that the LLM is expected to process (e.g., the sentence being translated or classified, the document being summarised, etc.). Instruction: Instead of showing concrete exemplars of correct behaviour in the prompt, we could textually describe what to do via an instruction; for example: "Write the words in the following sentence in reverse order." The instruction in table 3 is seen in the first line, while the choices below are considered the input data. Exemplars: One of the best ways to demonstrate the correct behaviour to an LLM is to provide a few concrete examples of input-output pairs inside the prompt. Indicators: Providing input to an LLM in a fixed and predictable structure is helpful, so we might separate different parts of our prompt by using indicators; for example, indicating question/answer or input/output pairs, such as in table 10. Context: Beyond the components described above, we may want to provide extra "context" or information to the LLM in some way (see figure 9). Another interesting technique tangentially related to instruction prompting is role prompting, which assigns a "role" or persona to the LLM. This role is assigned within the prompt via a textual snippet such as: "You are a doctor." or "You are a musical expert." (as seen in table 5). This context helps understand the question better and tailor the answer (also cf. relevance to personalisation).

Define ChatGPT role	Learn from ChatGPT	Chained Prompting
<ul style="list-style-type: none"> - Act as a Linux terminal - Act as "position" Interviewer - Act as a JavaScript Console - Act as an Excel Sheet - Act as an English Teacher - Act as a Plagiarism Checker - Act as an Advertiser - Act as a Relationship Coach - Act as a Recruiter 	<ul style="list-style-type: none"> - Explain clearly - Explain uniquely - Explain detailed - Explain like I'm 5 - Explain with examples - Explain to 5th graders - Explain like Noam Chomsky - Explain detailed with examples - Explain to high school students 	<p>Write an article about ChatGPT. First give me the outline, which consists of a headline, a teaser, and several subheadings. [Output]</p> <p>Now write 5 different subheadings. [Output]</p> <p>Add 5 keywords for each subheading. [Output]</p>

Table 8. Prompt patterns/templates for roles from various user blogs

Prompts for Marketers	Prompts for Designers	Prompts for Developers
<ul style="list-style-type: none"> - Can you provide me with some ideas for blog posts about [topic]? - Write a product description for my [product or service or company]. - Suggest inexpensive ways I can promote my [company] without using social media. - How can I obtain high-quality backlinks to raise the SEO of [website name]? 	<ul style="list-style-type: none"> - Generate examples of UI design requirements for a [mobile app]. - How can I design a [law firm website] in a way that conveys [trust and authority]? - Create a text-based excel sheet to input your copy suggestions. Assume you have 3 members in your UX writing team. 	<ul style="list-style-type: none"> - Develop an architecture and code for a <description> website with JavaScript. - Help me find mistakes in the following code <paste code below>. - I want to implement a sticky header on my website. Can you provide an example using CSS and JavaScript? - Please continue writing this code for JavaScript <insert code below>.

Table 9. Prompt patterns/templates for user groups from various user blogs

Tables 8 and 9 list other prompt patterns/templates from various user blogs to modify ChatGPT's output to be investigated in the photobook use case. In this project, we use DFKI IML's image captioning and use ChatGPT-4, as well as other LLMs, as an additional image captioning component and as a visual storytelling component. In the context of implementing mixed initiative interaction, we plan to explore the automatic prompting behaviour for ChatGPT-3 and ChatGPT-4. Especially the multimodal abilities of the latter are of interest for the photobook use case. We also plan to evaluate its concept understanding skills and its consistency, coherence, and correctness.

Automatic prompting refers to an automated method to create prompts for a diverse set of tasks. We will focus on the optimisation of prompts and calling them to ChatGPT's API. We want to optimise the actual textual/character-based prompt that is generated from human input (AutoPrompt Method)¹. Then we want to generate and execute the prompt completely automatically as a module subtask of a more complex AI system [Loedeman et al. 2022; Wu et al. 2022]. In our use case, these modules include, for instance, image retrieval, image captioning, and scene recognition.

In this extended use case of automatic prompting, we explore LLMs, mainly ChatGPT-3 and ChatGPT-4, but also open-source models like Llama 3 and LLaVA, to assist in these tasks. The main point is (also cf. figure 1) that the encompassing AI system automatically addresses ChatGPT via an API. In this process, the encompassing AI system must automatically generate the prompt. Since ChatGPT-4 is multimodal, meaning that it can process visual input along

¹This refers to the seminal paper [Pryzant et al. 2023] using natural language "gradients" and beam search algorithms.

with textual one, the prompt engineering process also becomes multimodal. A detailed categorisation of intelligent user interface key components, which serves to explain embedded ChatGPT function calls, can be seen in figure 4. The red dots indicate where ChatGPT can be used as an internal service via a function call (through a commercial API).^{2 3}

To implement the ChatGPT function calls in the No-IDLE photobook system, in this project we will use the template provided by [Zhou et al. 2023a] to generate prompts for (visual) storytelling from descriptions in isolation (from the VIST dataset, [Huang et al. 2016]).

In one example we use to illustrate the task, we only provide the model with one input-output pair (table 10). The generated prompts show that ChatGPT-3 generalises better, even with incomplete input. ChatGPT-4, however, provides a correct prompt but is way too detailed and tailored only for this specific input-output pair. One resulting task is to optimise the prompt behaviour for both cases.

ChatGPT-3 vs. ChatGPT-4
<p>Q: I gave a friend an instruction and one input. The friend read the instruction and wrote an output for the input. Here is the input-output pair:</p> <p>Input: A black frisbee is sitting on top of a roof. A man playing soccer outside of a white house with a red door. The boy is throwing a soccer ball by the red door. A soccer ball is over a roof by a frisbee in a rain gutter. Two balls and a frisbee are on top of a roof.</p> <p>Output: A discus got stuck up on the roof. Why not try getting it down with a soccer ball? Up the soccer ball goes. It didn't work so we tried a volley ball. Now the discus, soccer ball, and volleyball are all stuck on the roof.</p> <p>Formulate the instruction.</p> <p>ChatGPT-3: Write a description of a scene using the given input sentence.</p> <p>ChatGPT-4: The instruction is to reimagine and rewrite the given sentences using different terms, whilst still maintaining their key elements. For instance, replace 'frisbee' with 'discus', 'house' with 'place' or 'spot', 'playing soccer' with 'kicking the soccer ball'. Additionally, continue the story logically based on the impressions given in the input.</p>

Table 10. Automatic prompt generation with ChatGPT-3 vs. ChatGPT-4 for a visual storytelling task with captions in isolation input.

3.1.5 LLMs in the photobook use case. In No-IDLE, we rely on multiple deep learning components for our photobook creation support, such as image classification, text-to-image retrieval, and face recognition, as seen in figure 6. Image captioning, namely generating a description for a given image, is one of the system's central components.

In No-IDLE MEETS CHATGPT, we investigate the following questions: how does ChatGPT behave in the photobook use case for (1) cross-modal retrieval, (2) user-model interaction, (3) visual storytelling, and (4) commonsense reasoning.

²Here is a strong relationship to Systems AI, see:

<https://www.dfki.de/en/web/research/research-departments/foundations-of-systems-ai>

³Previous works have examined automatic prompting for language models. [Shin et al. 2020] propose a method for eliciting knowledge from language models with automatically generated prompts, in which customised prompts for specific tasks are created. [Zhang et al. 2023] automatise Chain-of-Thought-prompting in LLMs by enabling models to generate reasoning chains. [Zhou et al. 2023a] propose an approach for automatic instruction generation and selection, enabling LLMs to engineer prompts of human-level quality. [Hao et al. 2022] optimise prompts for text-to-image generation to obtain more aesthetically pleasing images. Their framework adapts original user input to model-preferred prompts. [Zhou et al. 2023b] demonstrate that the performance of automated prompting is heavily dependent on the amount of data available. The approach of [Singh et al. 2023] explains data patterns via the algorithm iPrompt: Auto-prompting is extended to generate a semantically meaningful natural language prompt that explains a key characteristic of given data examples.

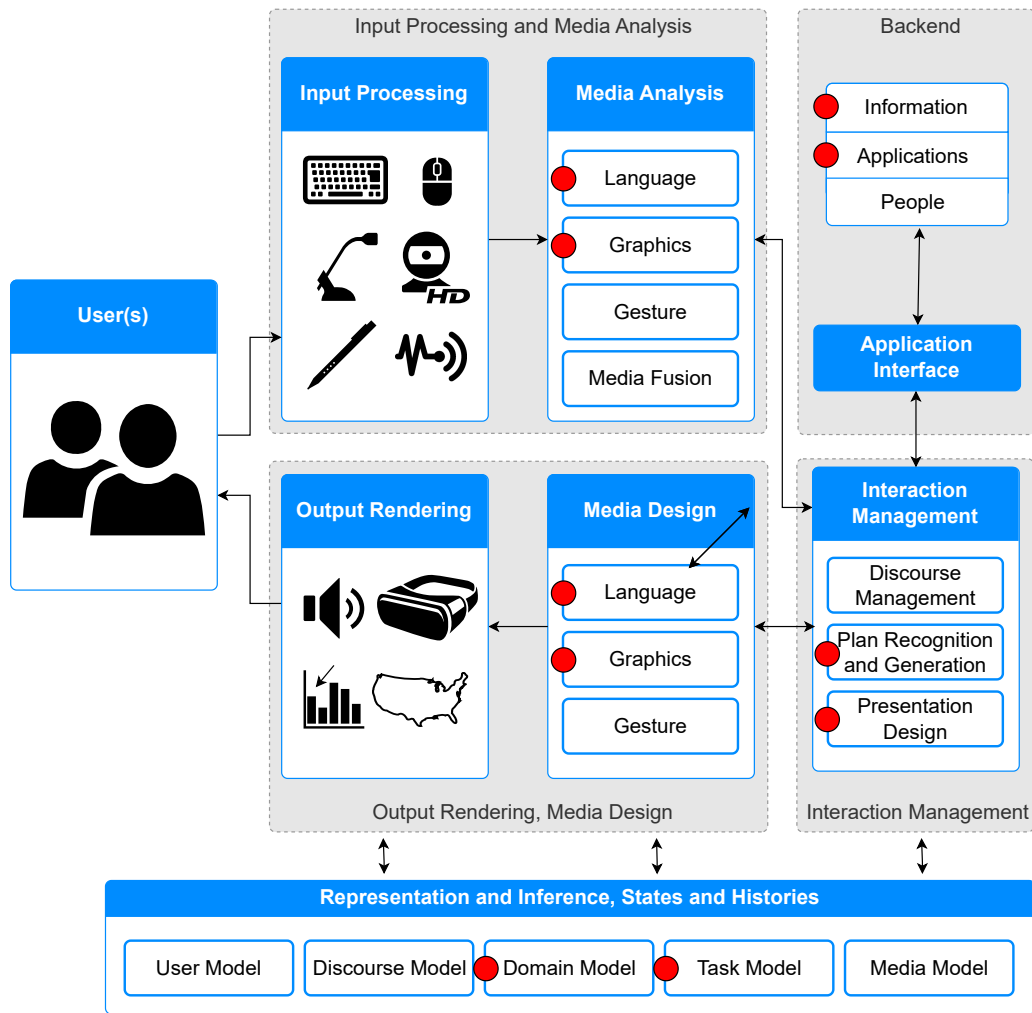


Fig. 4. Categorisation of intelligent user interface key components from [Zacharias et al. 2018], based on the conceptual architecture in [Wahlster and Maybury 1998] and DFKI's Smartweb system [Sonntag et al. 2007]. The red dots indicate where ChatGPT can be integrated.

- (1) In the first step of creating a photo page, our own modules or ChatGPT can be used to obtain relevant information (for example, location), which can then be used for text-to-image retrieval. We can also get salient objects (figure 5) and personalised captions (figure 6).
- (2) Additionally, the dialogue component of LLMs such as ChatGPT-3 and ChatGPT-4 can be used for interaction with the photobook, for example, by generating questions that help the user contextualise and control the caption.
- (3) LLMs can also generate more elaborate contextualised descriptions of (collections of) photos via visual storytelling.
- (4) Last, LLMs can be used as a commonsense reasoning component.

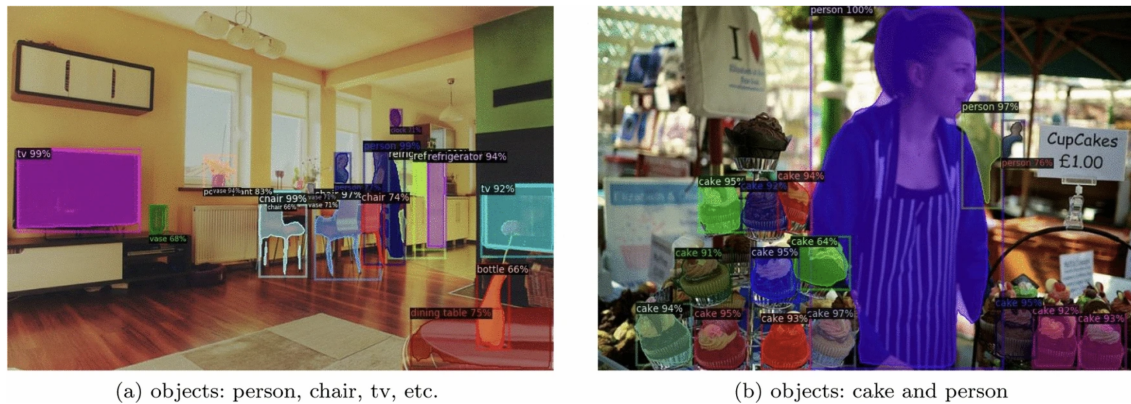


Fig. 5. Object-specific salient regions highlighted with corresponding bounding boxes as bottom-up features. From [Biswas et al. 2020].

Image captioning and visual storytelling. Recent image captioning approaches include large-scale vision-language pre-training and generalised models that work for various computer vision and vision-language tasks [Li et al. 2022a; Zou et al. 2022]. Interactive image captioning as envisioned has, in contrast, not gained as much attention as other machine learning tasks. [Jia and Li 2020] involve the human-in-the-loop by providing incomplete sequences as input, in addition to each image, during inference time. [Biswas et al. 2020] extend the Show, Attend, and Tell architecture [Xu et al. 2015] by combining explainable high-level and low-level features and beam search during decoding time (see figure 5).

Various components are activated in response to user commands. For example, when a user says, "On the first day, we took the bus from the airport to Vancouver," it initiates both text-to-image retrieval and image captioning components. This is modelled by the media analysis component. Models used in these cases must be adapted incrementally [Anagnostopoulou et al. 2023; Hartmann et al. 2022a].

In contrast to image captions that can be found in general-purpose datasets such as MS COCO [Lin et al. 2014] or Flickr30k [Plummer et al. 2015], the captions generated by our captioning component should be (1) entity-aware (e.g., instead of generic descriptions of objects or concepts, the captions contain proper names for named entities), (2) stylised, and (3) controllable. Existing models for entity-aware captioning usually generate a template caption with place-holders for named entities, which is then filled with information retrieved from associated text or knowledge bases [Biten et al. 2019; Lu et al. 2018]. Ramnath et al. [Ramnath et al. 2014] propose an approach for personalised template-filling with information such as geolocation, time stamp, detected landmarks, and recognised faces, which we plan to extend to incorporate finer-grained location information specified by the user. To generate stylised captions, we will explore caption generation reflecting sentiment [Mathews et al. 2016], specific styles [Gan et al. 2017; Guo et al. 2019], and taking into account a user's active vocabulary [Chunseong Park et al. 2017]. In the refinement phase, when additional captions are generated for newly retrieved images, the user should be able to exert fine-grained control over the concepts to be included in the caption, e.g., by actively modifying an abstract scene graph representation based on which the caption is generated [Chen et al. 2020].

(Multimodal) LLMs offer a promising solution for generating entity-aware and controllable captions. In a recent preliminary study, we examined the potential of LLMs and multimodal models (LMMs) to support journalistic practices

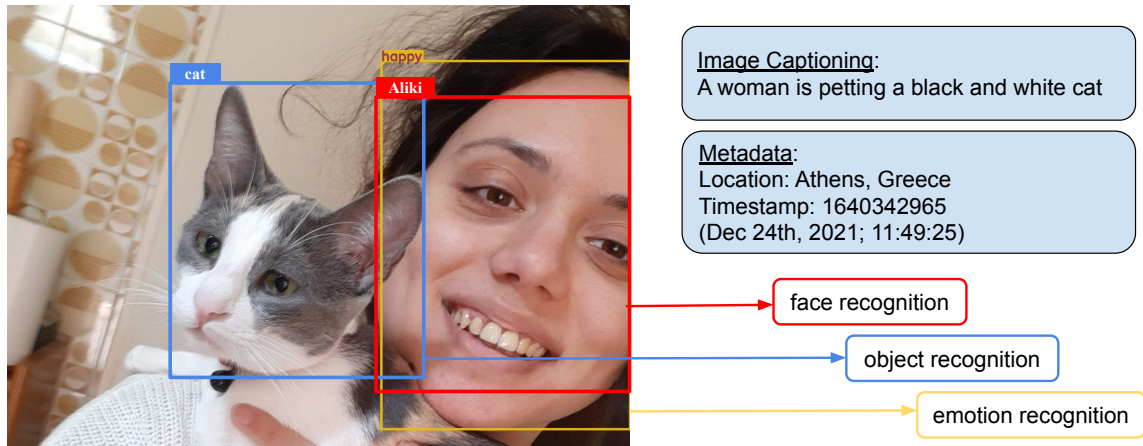


Fig. 6. Modules for personalised and contextualised image captioning: image captions, metadata, person and emotion, as well as specialised object recognition is required for the generation of a personalised caption. A sample input to the LLM is seen in table 7.

by generating contextualised captions for images accompanying news articles, a task closely related to the No-IDLE MEETS CHATGPT use case. We compared the performance of LLMs with a two-stage pipeline comprising separate captioning and contextualisation models [Anagnostopoulou et al. 2024].

Since we are experimenting with multimodal LLMs, automatic prompt engineering can be rendered multimodally, e.g., containing images. This ability should be explored in our photobook use case.

When the (deep learning) components lack information or pose a high uncertainty towards certain decisions, especially in the early steps of the system adaptation, the agent can ask clarification questions, thus asking for contextualising feedback ("Can you tell me more about this day?", "Who is present in this image?"), collecting it, if it is provided by the user (for example, in the case of hallucinations or incorrect information), or providing model explanations if asked ("Why was this image chosen?"). The collected feedback can then be used to improve the generated captions. Since this feedback will be in natural language, it is yet to be investigated how LLMs, and more specifically prompt engineering, can be used to render it into a format that can be used to update the LLM components themselves.

In contrast to generating captions for images in isolation, the visual storytelling component generates a sequence of captions that form a coherent story for a retrieved sequence of images [Huang et al. 2016; Jung et al. 2020; Wang et al. 2020]. Like captioning, the visual story component must be entity-aware and controllable. To this end, we will investigate to what extent approaches for adapting DFKI's captioning model can be transferred to the visual storytelling task of ChatGPT.

3.2 Natural language inference as a main evaluation task

NLI is a task in NLP that involves determining the logical relationship between two given pieces of text, typically referred to as the premise and the hypothesis [MacCartney and Manning 2008]. NLI requires models to understand the semantics and contextual nuances of language.

Here we describe the natural language inference (NLI) questions we are examining and explore our claims. Inspired by recent work on commonsense and NLI as a subtask, for example on legal textual entailment ([Nguyen et al. 2023b], negation detection assessment of GPTs ([Nguyen et al. 2023c] and evaluations of abductive reasoning support ([Nguyen

Statement	Premise with evidence annotations (boldface)	Label
The primary trial and the secondary trial both used MRI for their interventions.	Single arm of healthy postmenopausal women to have two breast MRI (baseline and post-treatment) . Letrozole of 12.5 mg/day is given for three successive days just prior to the second MRI. Healthy women will be screened for Magnetic Resonance Imaging (MRI) contraindications, and then undergo contrast injection, and SWIFT acquisition.	entailment
More than 1/3 of patients in cohort 1 of the primary trial experienced an adverse event.	Adverse Events 1: Total: 69/258 (26.74%) , Anaemia 3/258 (1.16%), Febrile neutropenia 13/258 (5.04%) Adverse Events 2: Total: 64/224 (28.57%), Anaemia 2/224 (0.89%), Febrile neutropenia 4/224 (1.79%)	contradiction

Table 11. Examples from the NLI4CT dataset [Jullien et al. 2023]. Label indicates the inference relation (entailment or contradiction) between a clinical trial report (premise) and a natural language statement ($premise \rightarrow statement$). Supporting facts that justify the predicted relation, shown in boldface, are part of the supervised dataset.

et al. 2023a)), we list the specific questions that we are exploring and evaluating. Finally, we explain how these research questions are related to the larger issues raised in the introduction.

3.2.1 *NLI in the medical domain: an ongoing project.* Our starting point for evaluation is a medical use case we implement in the context of a joint project in partnership with Accenture⁴. In the *Natural Language Inference for Clinical Trial Data* (NLI4CT) case [Jullien et al. 2023], automatic prompting (AutoPrompt) methods are applied to enhance a model’s ability to reason over complex medical information. The NLI4CT data collection, which is comprised of clinical trial reports (CTR), is used for the following two NLI tasks of interest for commonsense reasoning (table 11):

- **Textual entailment:** The statements may make claims about a single CTR premise or several CTR premises. The task is to determine the inference relation (entailment vs contradiction) between a statement and one or more CTR premises.
- **Evidence retrieval:** Given a CTR premise, and a statement, the output a set of supporting facts, extracted from the premise, necessary to justify the label predicted in the textual entailment task (e.g., entailment or contradiction).

In the current AutoPrompt project, the evaluation is as follows: The entailment prediction and evidence retrieval results are evaluated using standard accuracy metrics, including precision, recall, and F1-score. In No-IDLE MEETS CHATGPT, we transfer this task to the photobook use case. As stated earlier, we aim to explore, first and foremost, automatic prompting behaviour in the photobook use case context based on the most recent literature relevant to interactive deep learning. Specifically for prompting, different patterns are addressed by [Ouyang et al. 2022], who train language models to follow instructions with human feedback, and [White et al. 2023], who propose a prompt pattern catalog to enhance prompt engineering with ChatGPT. This transfer task is explained next.

3.2.2 *NLI in the photobook case: AutoPrompt transfer task.* The investigation into AutoPrompt methods (automated techniques for prompt generation) for No-IDLE MEETS CHATGPT is aligned with the challenges posed by NLI4CT tasks of textual entailment and evidence retrieval, where multi-hop NLI is essential. Current research in the medical

⁴<https://iml.dfki.de/news/autoprompt-aims-to-improve-chatgpts-analysis-of-clinical-data/>

domain focuses on developing methods that are suited for capturing the complex nature of clinical trial data and the nuanced inferences required within the framework of multi-hop reasoning. In No-IDLE MEETS CHATGPT, we propose a corresponding approach that integrates the same four key subtasks to be implemented: *textual context analysis*, *multi-hop reasoning*, *external knowledge incorporation*, and *user-machine interaction*. This approach aims to provide a comprehensive and context-aware methodology for navigating and reasoning over complex data in the medical domain in the Accenture project and will be transferred and generalised to the photobook use case. The key tasks to explore and evaluate centre around extending the investigation of AutoPrompt methods from clinical trial data analysis to the photobook use case. The main goal task for the model in NLI4CT is to verify claims by reasoning from reliable clinical trial sources. Instead of verifying medical claims, we create a photobook that requires the model to create stories and captions for the visual content and user preferences, such as specific visual cues, emotions and themes. This requires the ability to combine images content and text into a coherent narrative. Given the special requirements of photobook creation, particularly for tasks like image captioning and storytelling, the following adaptations in each key task aim to ensure the generation of visual narratives that are contextually rich, logically consistent, and aligned with user preferences:

Image and textual content analysis: Special prompts should guide the model better to extract relevant information, identify entities, and discern the logical relationships within each piece of text. The process of prompt engineering is complemented by the construction of a task and domain-specific knowledge graph, where entities serve as nodes and relationships as edges. The synergy between prompts and graph-based reasoning enhances coherence, facilitating multi-hop exploration. No-IDLE MEETS CHATGPT task: Reflect narrative structure in the photobook context. Extend entity recognition to include visual elements, thematic coherence in one still image and across different images. For example, special prompts guide multimodal models to capture specific visual cues, including object recognition and attributes prediction, and establish relationship between visual elements.

Multi-hop reasoning: AutoPrompt methods incorporate multi-hop reasoning across interconnected pieces of clinical trial information. No-IDLE MEETS CHATGPT task: In the case of photobook creation, special prompts facilitate the multi-hop reasoning algorithms to traverse the relationships between visual elements, considering connections not only within individual images but also across images, themes, and contextual relevance, to achieve a coherent image-based narrative.

External knowledge: AutoPrompt methods integrate path reasoning and external knowledge retrieval to enhance the model's understanding of medical concepts and relationships beyond the immediate textual context. In commonsense reasoning, the model excels at inferring implicit relationships and contextual nuances. In numerical reasoning, prompts should assist the model to interpret and operate quantitative information. No-IDLE MEETS CHATGPT task: Connecting to external sources of knowledge also benefits visual understanding and storytelling. Special prompts guide the model to incorporate world knowledge, commonsense reasoning, and cultural and historical references contributing to a deeper, richer and more informed narrative. Special prompts also play an important role in guiding the model to cross-reference information with external knowledge ensuring accuracy and reliability in the generations.

User-machine interaction: AutoPrompt methods "encourage" the model to interactively verify its findings and rectify errors based on user feedback. User-machine interaction also adds a dynamic element, allowing the model generation process to be more responsive to the needs and preferences of the end-users involved in (clinical trial) data analysis. No-IDLE MEETS CHATGPT task: foster user-machine interaction to make the photobook creation process more responsive to spontaneous preferences. Enable users to influence the choice of images (themes, emotional tone) and

interactive generation of visual narratives. This interaction mode enables users to actively shape the creative direction of the photobook, facilitating the construction of narratives in a collaborative and personalised manner.

We now specify our research questions. With experimental setups 1 and 2, we ground them in the current state-of-the-art of the entailment use case, explain how they can be achieved and describe how the achievement of goals can be measured. User-machine interaction aspects are explained in section 3.3.

3.2.3 Experimental Setup 1. While the medical use case relies on the NLI4CT dataset, a systematic evaluation in the photobook context requires utilising a series of well-established datasets. Our starting point for first experiments are the OK-VQA [Marino et al. 2019] and Visual Genome [Krishna et al. 2017] datasets and benchmarks: These evaluation datasets and benchmarks are crucial to ensuring that the AutoPrompt methods are thoroughly assessed and meet the specific challenges associated with the evaluation of LLM approaches for the photobook creation process.

A thoughtful and strategic prompt approach is necessary *to enhance the understanding of the scene* by the LLMs and to improve the generation of textual descriptions based on the extracted visual elements. Recent advances in prompting strategies, particularly the CoT [Wei et al. 2022b] approach and its variants, i.e., CoT-ST [Wang et al. 2022], ToT [Yao et al. 2023], CoVe [Dhuliawala et al. 2023], have shown significant improvement in the reasoning abilities of LLMs. However, multi-step prompting and prediction increase the likelihood of errors [Chu et al. 2023]. Self-verification becomes an indispensable component in addressing these concerns [Dhuliawala et al. 2023; Weng et al. 2023]. In self-verification, the LLMs reverse steps by checking the logical consistency and coherence between the latest generated response and previous responses. As stressed earlier, **NLI** is a task in NLP that involves determining the logical relationship between two given pieces of text, typically referred to as the premise and the hypothesis [MacCartney and Manning 2008]. **Framing self-verification as an NLI task enables the LLMs to evaluate its own outputs by treating them as hypotheses to be validated against earlier premises or facts.**

The overarching research question guiding our experiment is: **How effectively can LLMs conduct NLI subtasks to assess the logical consistency between captions or visual narratives they generate?**

To address this research question, we draw inspiration from the existing **NLI** benchmarks [Bowman et al. 2015; Camburu et al. 2018; Do et al. 2020; Kayser et al. 2021] and identify four key subtasks in a NLI pipeline:

- (1) **Textual Analysis** Identify the key terms and concepts in both premise and hypothesis descriptions.
- (2) **Reasoning** The model should identify and present textual evidence for making assessment decision, drawing from the comparison between two descriptions.
- (3) **Entailment Classification** Determining the logical relationship between descriptions, entailment, contradiction or neutral.
- (4) **Rationale Summary** Given the extracted evidence and entailment assessment decision, provide a comprehensible rationale as feedback on the misalignment or how to refine the second caption.

Details of the experimental setup: in our experimental setup, LLMs undertake **NLI** tasks; we can systematically evaluate the relationship between the two self-generated narratives. Through explaining the logical coherence between pairs of textual descriptions based on the same scene or image originating from different perspectives (e.g., person, location, emotion), the model autonomously identifies and rectifies inconsistencies, refining its generated content over successive iterations. An illustrative example is shown in table 12 and table 13. Table 13 demonstrates how the **NLI** assessment made by ChatGPT-3.5 following our strategy, comparing the hypothesis description with premise image caption. The hypothesis description is generated to extend the perspective of *"roles and relationships"*. In our design plan, the model is expected to demonstrate its inference results regarding the entailment relationship between two

Premise (validated)	Hypothesis	Label	Explanation
A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.	A young mother is playing with her daughter in a swing.	neutral	Child does not imply daughter and woman does not imply mother.

Table 12. In an evaluation instance derived from e-SNLI [Camburu et al. 2018], the premise comprises a verified image caption, while the hypothesis extends into a detailed description concerning the roles and relationships of the individuals. The annotated label for the entailment relationship between these two captions is neutral, signifying the absence of explicit information about the people in the provided imagery, as the human annotator justifies the *neutral* label in the explanation.

Textual Analysis	<p>Validated Caption (Premise): Child in a yellow plastic safety swing. Laughing. Dark-haired woman in pink and coral pants standing behind her</p> <p>Second Caption (Hypothesis): Young mother. Playing with her daughter. In a swing.</p>
Reasoning	<ol style="list-style-type: none"> 1. The first description portrays a child in a swing with a woman standing behind, but the specific relationship between them is not explicitly mentioned. 2. The second description introduces the concept of a young mother playing with her daughter in a swing. While the roles align, it's not explicitly stated that the woman in the first description is the same as the young mother in the second description. 3. The lack of specific details connecting the woman in the first description to the role of a mother in the second description leaves room for ambiguity. 4. The woman in the first description could be a caregiver, family member, or someone else standing behind the child. 5. While there are overlapping elements like the swing and the presence of a woman, the precise relationship remains unclear, and the descriptions may represent different perspectives or interpretations of the same scene.
Entailment Classification	Neutral
Rationales Summary	The second description does not necessarily entail the first, as the specific identity of the woman in the first description as the young mother of the child is not explicitly confirmed in the given context.

Table 13. In this example, ChatGPT-3.5 demonstrates its reasoning capability in determining the relationship between the premise and hypothesis presented in table 12. In the context of NLI, the model is guided through four key steps when evaluating the second generated caption: textual analysis, reasoning, entailment classification, and rationale summarisation. The rationale from the entailment classification is then used as feedback on how to refine the second caption.

visual descriptions. The results, particularly the rationales for "contradiction" or "neutral" outcomes, should be used to guide the refinement of captions for the next steps. This involves correcting the misalignment or requesting additional details from the user.

The connection to the photobook use case is straightforward: After the self-verification in responses, the subsequent steps involve analysing the model’s justifications to conclude its reasoning responses. If the model correctly identifies that the successively generated description lacks information or contradicts its premise, it is a valuable indication for the model to inquire about more knowledge or further instructions from the user (cf. user feedback). In a second step of our experimental setup 1, we include a number of additional benchmark datasets to systematically assess the adaptability of the NLI-guided method and the consistency of reasoning capabilities in LLMs across different cases in the photobook scenario:

- **SNLI** [Bowman et al. 2015]: A large NLI benchmark based on the Flickr image captions [Young et al. 2014].
- **e-SNLI** [Camburu et al. 2018]: An NLI benchmark contains part of SNLI samples and a manually annotated explanation given the premise, hypothesis and label.
- **SNLI-VE** [Xie et al. 2019]: A visual NLI benchmark adds the original image besides the image caption as premise.
- **e-SNLI-VE** [Li et al. 2021]: A visual NLI benchmark augments human annotated explanation given the image and language premise, language hypothesis and label.
- **FEVER** [Thorne et al. 2018]: This dataset consists of sentence pairs sourced from Wikipedia, to assess the accuracy of claims by reasoning factual evidence.
- **MultiNLI** [Williams et al. 2018]: The dataset comprises sentence pairs from ten distinct genres of written and spoken English, allowing for a comprehensive evaluation of generalisability across various linguistic contexts.
- **MedNLI** [Romanov and Shivade 2018]: This dataset contains sentence pairs grounded in the medical history of patients and annotated by medical professionals.
- **NLI4CT** [Jullien et al. 2023]: This dataset is derived from breast cancer clinical trials. Tasks are framed as classifying the entailment relationships between a statement and the clinical trial report as premise and extracting evidence to support the classification decision.
- **StrategyQA** [Geva et al. 2021]: Although the tasks in this dataset are designed in a question-answering scenario, the annotations provided in the dataset, including term descriptions and the supporting facts, are valuable for assessing the reasoning and decision-making capability of LLMs. Hence we adapt this dataset to our experiment setup.

3.2.4 Experimental Setup 2. In this second experimental setup, we focus on vision-language (VL) models. Let’s recall some preliminaries from commonsense reasoning. LLM capabilities for commonsense reasoning and their limits have been investigated in literature [Huang and Chang 2023; Wei et al. 2022a; Zhou et al. 2020b]. It remains to see how they can be explored in the context of our work, besides natural language inference as in experimental setup 1, especially for filtering relevant images and creating a compelling story. If, for example, commonsense reasoning can inform query expansion, if the topic is ‘Christmas’, objects that can be found in relevant images would include the local Christmas market, Christmas trees, etc. A second, more detailed example, is shown in figure 7. The textual premise provided by a direct captioning component (as we are implementing in No-IDLE) delivers a textual description like "A person is blowing out candles on a cake." The commonsense reasoning steps are like: *you blow out candles* → *birthday* ← *motivation* ← *Bake a cake*, resulting in "A person is blowing out candles on a birthday cake."

In a third example scenario (figure 8), the conventional caption (textual premise of multiple images) would be "group of runners get prepared to run a race", while the commonsense-enriched caption would include information like the intention behind the action, attributes and effect of actions from the agents: "In order to *win a medal*, a group of runners gets prepared to run a race. As a result *they are congratulated at the finish line*. They are *athletic*."



Fig. 7. "A person is blowing out candles on a birthday cake." This is an example image showing a case when identifying semantic concepts, not explicitly represented in the scene, would help to provide a better description. Commonsense reasoning is needed to relate the elements, namely people, cake and candle, to the concept of a birthday. Image source <https://www.stockfood.com/images/11162229-Young-woman-blowing-candles-on-birthday-cake-studio-shot>.



Fig. 8. The commonsense-enriched caption would include information like the intention behind the action, attributes and effect of actions from the agents: "In order to *win a medal*, a group of runners gets prepared to run a race. As a result *they are congratulated at the finish line*. Example from [Fang et al. 2020] and video contents [realsbstn (YouTube channel) 2009].

In general, an integral part of this work package task will be to evaluate ChatGPT's human-like ability to make presumptions about the type and essence of ordinary situations humans encounter every day. These assumptions include judgements about the nature of physical objects, taxonomic properties, and people's intentions. A very challenging part is to explore the usage for commonsense reasoning in explanations of why a particular image has been chosen or why a particular caption or story has been generated, as explained next.

In addition to experimental setup 1 with a focus on textual premises, we design experiments focussing on the visual content and visual premises of the photobook use case scenario. Our envisioned contextualised captioning workflow is illustrated in figure 9. Our experiments focus mainly on the *visual storytelling* component, with additional explorative tasks for *image retrieval*. Our starting point for these experiments are vision-language (VL) datasets. Our evaluation tasks are guided by the following research questions:

- **How grounded⁵ and hallucination-free are (storytelling) descriptions generated by an LLM?**
- Can an LLM extend self-verification to user input verification and help identify inconsistencies in it?
- How can (multimodal) NLI efficiently be integrated into the image retrieval component?

⁵In image captioning, grounding usually refers to the generation of captions that focus on certain parts of the image [Ma et al. 2020; Zhou et al. 2020a]. In this case, we have an extension of the image captioning task with the generation of stories - hence, grounding refers to the stories being directly related to the semantic content of the images, as present in the captions.

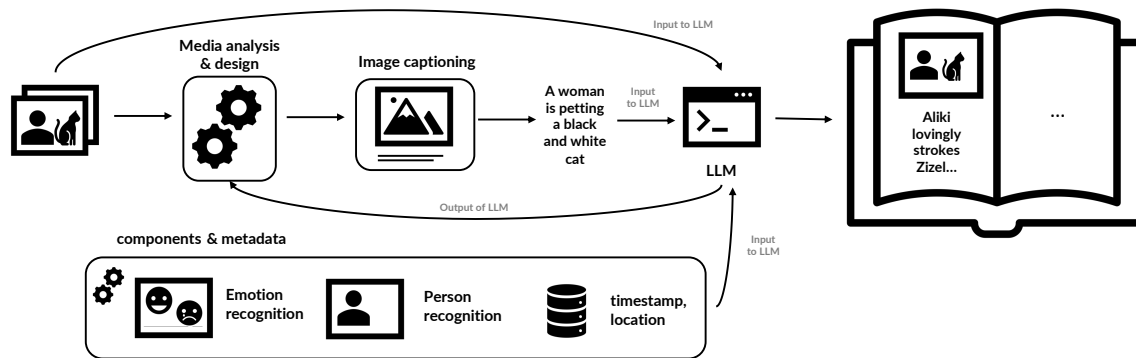


Fig. 9. Contextualised image captioning with an image captioning module, DL components, metadata, and the use of an LLM.

DII – Premise	SIS – Hypothesis	Label
Three tents in the grass near two people.	They had a lot of tents.	contradiction
A little girl in a floatie swims towards her dad in the jacuzzi.	She enjoyed spending some time with grandpa in the pool.	contradiction
A father and daughter spend time together with others outdoors.	My dad took us to the 4th of July picnic.	neutral
Mushrooms are among the food sitting on a white plate.	They are finally served with a side of risotto.	neutral
A woman looks down at her baby while they sit on a blanket outside.	My wife took care of one of the kids.	entailment
Two boys are enjoying cotton candy at a park event in the summer.	The family went to the park for 4th of July.	entailment

Table 14. Manually annotated examples from the VIST dataset to demonstrate how an NLI task based on this dataset would look like. DII: description-in-isolation, SIS: story-in-sequence.

For our main task (**entailment generation and verification**), we use the VIST [Huang et al. 2016] and the Video2Commonsense [Fang et al. 2020] datasets. The VIST dataset contains images, along with their *descriptions in isolation or DII* (simple image captions) and *stories in sequence or SIS* (contextualised visual storytelling). The images are grouped into sequences of five, for which the SIS is generated. As presented in table 14 we consider DIIs to be the premise, and SIS to be the hypothesis. Given the premises or image captions, the task is to use LLMs to generate the 'entailment' or storytelling part. The dataset Video2Commonsense contains video clips which have been annotated with commonsense descriptions about the agents in the video, including commonsense aspects (the effect of the action, as well as the attribute of the agent giving the action). In our case, the LLM can be used to obtain commonsense-enriched descriptions. We use natural language generation metrics as depicted in table 14 and employ NLI to verify whether the LLM generation is entailed in the caption.

This task is directly related to the visual storytelling component of the photobook use case scenario since it involves generating a story-in-sequence rather than plain image or frame captions. Besides the generation component, the NLI, or verification component, ensures that the story generated corresponds to the content of the image, checking for LLM hallucination at this point.







Visual Premise	Textual Premise	Hypothesis	Label	Explanation
	The brown dog with a red bandana is sitting with its paws up on the computer.	A person is sitting at his desk with his cat on his lap, he hates dogs and refuses to own one.	contradiction	A dog isn't a person.
	An Asian tennis player is about to hit the ball.	A ladies golf player teeing up at the 18th hole.	contradiction	Tennis is not golf.
	A girl rides a horse.	A girl is riding a horse through a park.	neutral	The girl is clearly riding a horse but it is not clear that it is in a park.
	A person wearing a hat and concealing clothing is on a bicycle carrying a large amount of products on the back with a large group of bicyclers behind her.	There is currently rush hour traffic.	neutral	There is a lot of traffic, but it does not mean that it is rush hour.
	A beagle walks along the calm shore.	There is an animal on the beach.	entailment	A beagle is an animal, and walking on a beach requires that one actually be on said beach.
	People are admiring a work of art.	People standing, looking at art.	entailment	Looking at art is a restatement of staring up at a work of art.

Table 15. Visual Premises, Textual Premises, Hypotheses, Labels and Explanations. Examples from the e-SNLI-VE dataset [Kayser et al. 2021].

In addition, we define two exploration tasks related to image retrieval & visual storytelling:

- **Visual NLI for image retrieval:** For this use case, we can utilise the e-SNLI-VE dataset [Kayser et al. 2021]: Given an image premise, a hypothesis is formulated, which can be *entailed*, *contradicting*, or *neutral* (see examples in table 15). This is a task that can be approached in two ways: The premise used can be an image, or, even more suitably, a caption or a group of captions to an image. In order to obtain the caption(s), we can either use captions provided by the e-SNLI dataset, or employ a (potentially separate) image captioning system. Congruency between hypotheses and images/captions can be crucial for the image retrieval component of the use case. In the case the user inputs a certain prompt for retrieval (hypothesis) which does not return the desired image(s), an explanation for this behaviour can be provided.
- **Entailment assessment for user feedback:** For this task, we plan to use the VIST dataset, with additional NLI annotations between DIIs and SISs. Given that the hypothesis is provided by user feedback, namely a user's story or feedback for an image, the task is to check if it is congruent with the premise/caption. If there is a contradiction, a feedback loop between the system and the user can be initiated to either rectify the mistake or provide an explanation. Since this task includes the verification of (user-generated) stories, it is connected to the visual storytelling component of the use case.

Additional datasets which can be used to evaluate the LLM's storytelling and commonsense capabilities include the following:

- **VQA-E** [Li et al. 2018], a visual question answering with explanations derived from captions, including specific, abstract, and subjective questions;
- **MS COCO** [Lin et al. 2014], a general use computer vision and image captioning dataset, containing five captions for each image;
- **refCOCO** [Kazemzadeh et al. 2014], which contains referring expressions for distinct objects in the MS COCO dataset;
- **OpenImages** [Kuznetsova et al. 2020] containing, among other annotations such as bounding boxes, object segmentations and image-level classes, so-called "*localized narratives*";
- **Conceptual Captions** [Sharma et al. 2018], including images and captions in different styles (both alt-text scraped from the web and conceptual captions derived from them).

Explainability—Connection to XAINES (BMBF) The experiments described in this setup include the generation of different forms of text, including explanations for the NLI model decisions. In the case of the main task, the dataset itself contains explanations; in the case of the exploratory tasks, the feedback loop initiated can potentially provide the system with an explanation from the user, which can then be integrated as feedback to the model. These explanatory properties of the tasks are connected to the XAINES project. XAINES [Hartmann et al. 2022b] furthermore differentiates between *machine learning* and *domain expert* explanations. In the former case, model explanations help identify system deficiencies, while in the latter case, explanations can contribute to the domain expert's satisfaction and trust. In the context of NO-IDLE MEETS CHATGPT, both machine-learning and domain-expert explanations are relevant: Machine-learning explanations can assist in identifying errors during the photobook generation, while domain-expert explanations can be used to assess if the decision the model took was justified. For the context of our proposed project, domain-expert explanations are relevant, given that the users are "experts" of their own photo collection.

3.3 Human-Computer Interaction (HCI)

A central goal of this project is to leverage LLMs’ abilities to improve user interaction with deep learning-powered systems. In the current section, we propose to explore how the methods described in section 3.2 can be controlled effectively by the user so as to improve usability of the interactive system for photobook creation.

We distinguish between three types of goals: user-, NLP and AI research-, and HCI research goals. The NLP goals are described in section 3.2; the user goals are to compose a narrative by selecting the right images and writing a story about them; lastly, the HCI challenges consist of enabling the user to exert control over the generated content.

With the HCI goals in mind, we explore what opportunities for interaction are offered by the NLP algorithm designs, and how they can serve as endpoints linked to components of user interfaces. A trivial way for a user to control the content generated by LLMs is via the input prompt. Here, our goal is to identify novel forms of interaction that allow the user to produce rapid, directed, incremental, and reversible changes on the generated content, in line with principles of direct manipulation interfaces [Hutchins et al. 1985; Shneiderman 1983].

To facilitate investigation, we propose two user-centric surrogate tasks that capture distinct aspects of the photobook creation use case: content-based image retrieval (the *biased journalist* task), and visual storytelling (*heads-bodies-legs* task). Each task will require the development of a human-in-the-loop system that exploits LLMs abilities to accomplish a different user goal. These two tasks will constitute sandbox environments for developing general principles and concrete modules that will later be integrated into the complete photobook creation tool.

3.3.1 Biased journalist task. The first task addresses image and textual content analysis, as well as content-based image retrieval. In this task, the user’s goal is to retrieve a small set of images from a large pool by providing an input prompt in natural language. A (mnemonic) example application would be a user writing a factual newspaper article with a biased, preconceived spin, who uses the tool to choose illustrative photos that confirm that bias. An analogous application would involve finding holiday photos on a smartphone image archive by asking questions in natural language.

How can AutoPrompt methods help the user accomplish these goals? While the design of user studies with specific interactive systems will depend on the outcome of research in section 3.2, we sketch two scenarios with illustrative algorithm designs for solving the biased-journalist task, as well as corresponding HCI goals.

In the first scenario, a **large multimodal model (LMM)** such as GPT-4V or LLaVA [Li et al. 2023] is used for both content analysis and image retrieval, and the user interacts with the result set. The LMM takes as input both (a) a large pool of images and (b) a natural language retrieval prompt, and outputs a set of images that are related to the input prompt. The user can exert control over the generated content by refining the results, e.g., by simply hand picking images from the result set. Alternatively, the user can use pointing and clicking to indicate which resulting images are a better match to the intention expressed through the input prompt, as well as adjustments to the natural language prompt. Furthermore, NLI methods can be used to revise results and detect occasional logical inconsistencies of choices made by the model, and verification steps can be revised and corrected by the user.

The scenario illustrated above requires running inference with large models on the entire image pool, and might encounter limitations in terms of computational costs for very large pools. A second possible scenario alleviates this problem by first embedding the image pool into a smaller, semantically structured space. In this scenario, natural language captions are generated for each of the images in the pool; both captions and input prompt are then embedded with a sentence embedding model [Li et al. 2022b], and content based image retrieval takes place in this space by selecting points based on their position relative to the embedded input prompt. Images from sentence embeddings could then be presented to the user as actionable, low dimensional point clouds [Kath et al. 2023a,b]. Alternatively,



Fig. 10. In this children's game, any combination of head, body, and legs makes up a coherent (albeit whimsical) full figure. In the *heads-bodies-legs* task, the user is challenged to elicit a coherent story out of a combination of randomly chosen images by operating a direct manipulation interface. (Image source: <https://www.lucykingdesign.com/journal/2013/04/mixies.html>)

ChatGPT-3

Q: Write a story with about 100 words that connects the following elements, rearranging as needed:

- The brown dog with a red bandanna is sitting with its paws up on the computer.
- An Asian tennis player is about to hit the ball.
- A girl rides a horse.

A: In a quaint town, the brown dog with a stylish red bandanna belonged to Emily, a spirited girl with a love for animals. As **she typed away on her computer**, the canine companion sat beside her, **paws up in playful curiosity**. Meanwhile, on the nearby tennis court, an Asian player prepared to serve a powerful shot, determination etched on her face. At the edge of the sports complex, a young girl named Lily enjoyed a horseback ride, the **wind tousling her hair**. Unbeknownst to them, *their worlds intertwined, creating a tapestry of diverse lives woven together in the vibrant fabric of the community.*

Table 16. Example story generated with ChatGPT-3 from the textual premises of three arbitrarily chosen images (in this case, picked from the top of table 15). Text in boldface indicates hallucinations clearly contradicted by the images (see page 22); italics indicate trivial juxtaposition.

captions could be generated conditionally based on the input prompt, offering the user an opportunity for iteration. Using conditional captions should yield better results, but at a higher computational cost.

3.3.2 Heads-bodies-legs task. The second task addresses visual storytelling, leveraging *large language models (LLMs)*' potential to make sense of an arbitrary set of images while avoiding hallucinations. In a well known children's game, a series of human or animal figures are split into three parts (head, body, and legs), and any arbitrary combination of the three parts will generate a coherent (albeit whimsical) full figure. Similarly, the goal of the user in this task is

to generate a coherent story from a set of given, potentially unrelated images. A main research goal in this task is to enable efficient control of LLM hallucinations through user input.

An illustrative, baseline implementation of a tool that solves the heads-bodies-legs task by leveraging LLMs is given in table 16. Starting with an arbitrary set of three images (the first three entries on table 15), the first step is the generation of captions for each image independently. Next, ChatGPT is asked to compose a story connecting them. In this example, while the generated story does show a degree of coherence, two problems are encountered: trivial juxtapositions and hallucinations.

An unsatisfactory feature of the generated story is that coherence is accomplished largely by simple juxtaposition of facts, a trivial strategy (italics in table 16). This is to be expected given the difficulty of connecting such unrelated images. The HCI goal of enabling the user to generate rich, coherent, and interesting stories will require finding an adequate difficulty level; difficulty in this case can be operationalised, e.g. as distance between captions in sentence embedding space. Another problem with the generated text is that it includes a number of glaring contradictions with the input images (hallucinations; compare boldface text in table 16 with input images on page 22). We hypothesise that these hallucinations are boosted by the degree of difficulty of the task, and that they can be tamed with the verification methods described in section 3.2. A first implementation of a user-facing system could include a user-triggered verification loop that tests for entailment between information contained on images and the generated text. The premise and evidence annotations could be displayed to the user in a similar fashion as shown in table 11, and the user could indicate by pointing and clicking which portions of the evidence are most relevant for the generated story.

4 EVALUATION PLAN

In this section we provide details about our general evaluation process and study plan. According to internal and external reviews, we introduce benchmark evaluations of subtasks in addition to user studies.

4.1 Benchmark studies

It is difficult to rate the performance of LLMs on a set of standard benchmark datasets for the end-to-end system because there is currently no benchmark for the entire dialogue-based photobook scenario. In addition, we cannot access detailed information about training data of ChatGPT-4. Since GPT models are not restricted to a certain domain but can refer to any topic, choosing an appropriate evaluation metric constitutes a problem, too. One solution is to use a more specific approach where we (1) evaluate certain subtasks where benchmark exists or (2) generate benchmark data for certain subtasks. In this way, we satisfy the rigorous standards of scientific evaluation to a larger degree, which was not possible before (cf. criticisms of the standard approach to measuring AI systems, new benchmarks must be proposed).

4.1.1 Existing benchmarks. The experimental setups 1 and 2 introduced several benchmarks to start with. We also collected two extensive lists of additional potential benchmarks for consideration. Table 17 shows detailed information of computer vision (CV) and natural language processing (NLP) datasets for LLM evaluation along the subtasks of text-to-image retrieval, interaction with the user, visual storytelling and contextualisation, and commonsense reasoning. Table 18 focuses on vision-language (VL) cross-modal evaluation.

4.1.2 Benchmark generation. While focussing on prompt engineering, automatic prompting, and scene understanding, we will follow the new "LLM-Human-in-the-Loop pipeline" approach [Chen et al. 2023b] and motivate to establish a new benchmark dataset that provides annotation of the fine-grained reasoning steps to automatically measure the reasoning

ds type	task	dataset name	description, # of instances	evaluation	1	2	3	4
CV	image sentiment analysis	Image Sentiment [Hassan et al. 2022]	3679	A/P/R/F1	x		x	
CV	FER (emotion recognition)	RAF-DB [Li and Deng 2019; Li et al. 2017]	29672	A/P/R/F1	x		x	
CV	FER (emotion recognition)	FERplus [Barsoum et al. 2016]	35887	A/P/R/F1	x		x	
CV	FER (emotion recognition)	AffectNet [Mollahosseini et al. 2019]	450000	A/P/R/F1	x		x	
CV	person recognition	DigiFace-1M [Bae et al. 2023]	720K images, 10K identities & 500K images, 100K identities	accuracy		x	x	
CV	person recognition	VGG Face 2 [Cao et al. 2018]	3,310,000 images, 9,131 identities			x	x	
CV	object detection & localisation, facial recognition	CIFAR-10; CIFAR-100 [Krizhevsky and Hinton 2009]	6000 images x 10 classes; 600 images x 100 classes		x	x	x	
CV	image classification	ImageNET(full) [Deng et al. 2009]	100,000 synsets x 1000 images		x			
CV	human action prediction	Kinetics400 Dataset [Kay et al. 2017]	1150		x		x	
CV	scene understanding, scene graph generation	Visual Genome [Krishna et al. 2017]	100K images where each image has an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects		x	x	x	
CV	scene understanding	ScanNet [Dai et al. 2017]	2.5M views in 1513 scenes		x	x	x	
CV	object detection & localisation, semantic segmentation	MS COCO, COCO Caption [Lin et al. 2014]	330K images (>200K labelled), 1.5 million object instances, 80 object categories, 91 stuff categories, 250,000 people with key-points	NLG metrics	x		x	
CV	object detection & localisation, semantic segmentation	OpenImages [Kuznetsova et al. 2020]	9159704		x		x	
NLP	relation extraction	TACRED [Zhang et al. 2017]	119474	P/R/F1		x		x
NLP	topic modelling	AG news [Corso et al. 2005; Gulli 2005]	127600	PMI, C_v	x			
NLP	semantic role labelling	OntoNotes 5.0 [Weischedel et al. 2013]	1445000 tokens	P/R/F1			x	
NLP	fake news detection / hate speech detection	LIAR [Wang 2017]	12836	accuracy		x		
NLP	dialogue / explanation acts	Wired [Wachsmuth and Alshomary 2022]	1550 instances	A/P/R/F1		x		
NLP	commonsense reasoning	CommonsenseQA [Talmor et al. 2019]	12,247 examples	accuracy				x
NLP	commonsense reasoning	StrategyQA [Geva et al. 2021]	2,780 examples	accuracy				x
NLP	commonsense reasoning	AI2 reasoning [Clark et al. 2018]	7,787 multiple-choice questions	accuracy				x

Table 17. Computer vision (CV) and natural language processing (NLP) datasets for LLM evaluation. The numbers on the columns on the right side indicate elements of the photobook use case pipeline each task is relevant for. 1: text-to-image retrieval, 2: interaction with the user, 3: visual storytelling and contextualisation, 4: commonsense reasoning.

type	task	dataset name	description, # of instances	evaluation	1	2	3	4
VL	visual reasoning / compositional QA	GQA [Hudson and Manning 2019]	113000 images, 22M questions	dataset-specific metrics	x	x		
VL	explainable visual QA	VQA-E [Li et al. 2018]	108325 images, 269786 annotations	NLG metrics		x	x	
VL	visual storytelling	VIST [Huang et al. 2016]	81743 images, 20211 sequences	NLG metrics			x	
VL	scene understanding, scene-to-text generation, intention prediction	refCOCO [Kazemzadeh et al. 2014]	130,525 expressions referring to 96,654 distinct objects in 19,894 photographs of real-world scenes	NLG metrics	x		x	
VL	vNLI, fact-based QA	VQA [Antol et al. 2015]	1,105,904 questions; 11,059,040 ground truth answers, At least 3 questions (5.4 questions on average) per image					x
VL	vNLI	e-SNLI-VE [Do et al. 2020]	31783					x
VL	vNLI	VCOPA Dataset [Yeo et al. 2018]	380 questions of commonsense causality with 1,140 images.					x
VL	fact-based QA	Outside Knowledge VQA (OK-VQA) [Marino et al. 2019]	14031 images, 14055 questions, visual reasoning with open knowledge (Wikipedia)					x
VL	fact-based QA	OK-VQA (A-OKVQA) [Schwenk et al. 2022]	Reasoning, knowledge diversity, rationales, answer types					x
VL	fact-based QA	INFOSEEK [Chen et al. 2023a]	Visual information-seeking questions, 9 image class					x
VL	fact-based QA	FBQA [Wang et al. 2018]	fact triplets (<Cat, CapableOf, ClimbingTrees>); 2190 images, 5826 questions, 32 categories of visual concepts (object, action, ...), visual reasoning with given knowledge base					x
VL	fact-based QA	SSS [Jain et al. 2021]	S3VQA: 6765 question-image pairs OKVQAS3: 2640 question-image pairs	accuracy				x
VL	fact-based QA	DAQAR [Malinowski and Fritz 2014]	question-answer pairs + relations and facts annotation number of questions: 12 468	task-specific metric		x		x
VL	fact-based QA	OVEN [Hu et al. 2023]				x		x
VL	fact-based QA	CLEVR [Johnson et al. 2017]	100K images, 999968 questions	accuracy		x		x
VL	fact-based QA	Knowledge-based VQA (KB-VQA) [Garcia-Olano et al. 2022]	visual questions, commonsense questions, KB-knowledge questions	accuracy		x		x
VL	image-to-captions generation	MS COCO, COCO Caption [Lin et al. 2014]	330K images (>200K labelled); 5 captions per image	NLG metrics	x		x	
VL	image-to-captions generation	OpenImages [Kuznetsova et al. 2020]	9159704		x		x	
VL	image-to-captions generation	Conceptual Captions [Sharma et al. 2018]	3369218		x		x	
VL	vNLI, visual commonsense reasoning	Video-and-Language Inference (Violin) [Liu et al. 2020]	95,322 video-hypothesis pairs from 15,887 video clips					x
VL	intention prediction	Video2Commonsense [Fang et al. 2020]	9K videos of human agents performing various actions, annotated with 3 types of commonsense descriptions: intention, effect, attribute		x		x	
VL	vNLI, event extraction	VisualComet [Park et al. 2020]	60K images with place information, 139K events at present, 2.3 events per image, 1.4 million total inference descriptions, 580K before inference descriptions (4.3 per event at present), 580K after inference descriptions (4.3 per event at present), 295K inferences on intent (2.1 per event at present)					x
VL	vNLI, fact-based QA, intention prediction	Visual Commonsense Reasoning (VCR) [Zellers et al. 2019]	290K multiple choice QA problems	accuracy				x

Table 18. Vision-language (VL) datasets for LLM evaluation. The numbers on the columns on the right side indicate elements of the photobook use case pipeline each task is relevant for. 1 text-to-image retrieval, 2: interaction with the user, 3: visual storytelling and contextualisation, 4: commonsense reasoning.

consistency (while reducing hallucinations). Hence, the application and evaluation part of our system is strengthened by a ChatGPT-4-powered database construction part. To address a crucial concern regarding the extent to which reasoning capabilities are fully consistent and grounded, we require a benchmark that encompasses both high-level inference and detailed reasoning chains, which is costly. We tackle this challenge by proposing an instantiation of the "LLM-Human-in-the-Loop" pipeline, which potentially notably reduces cost while simultaneously ensuring the generation of a high-quality dataset. This pipeline functions by incorporating limited human assistance for providing instructions and filtering rules, enabling LLMs to efficiently generate high-quality datasets in a semi-automatic manner, substantially reducing annotation cost. Interestingly, the user input step can be combined with the dialogue sequences with the end user.

We also plan to include the usage of prompts "Let's think step-by-step" [Kojima et al. 2022] and other adaptive prompts following Faithful CoT [Lyu et al. 2023] to enhance benchmark dataset generation. In addition, we add an explorative task of the most recent outcomes in the direction of CoT reasoning, namely its combination with multimodal reasoning [Wang et al. 2024a]. Consider again the picture in table 12: How do you know that it is a girl? How do you know that the other person is a woman? How do you know their relation? Benchmarks should improve the holistic versus component-based recognition. It is a face, how do you know? It has two eyes, a nose, a mouth, and has hair. How do you know that those small circles are the eyes? Because it has the white part of the eye, the iris, the pupil, and the eyelashes.

A dataset for Graph of Thoughts (GoT) might also be in reach [Besta et al. 2023] to enable combining arbitrary LLM thoughts into synergistic outcomes, distilling the essence of whole networks of thoughts, or enhancing thoughts using feedback loops to reduce hallucinations.⁶ A similar idea to detect/reduce hallucinations could be explored by constructing a dataset for exploiting commonsense knowledge about objects for visual activity recognition [Jiang and Riloff 2023] in the context of the VL tasks, including scene understanding, for experimental setup 2.

4.2 User studies

The *biased journalist task* and the *head-bodies-legs task* described in section 3.3 will be undergoing user studies to evaluate on the performance and suitability of the implemented NLI tasks introduced in experimental setup 1 and 2 (see section 3.2) for the user experience. The *biased journalist task* investigates mainly different methods of image retrieval. In comparison, the *heads-bodies-legs task* researches the use of different methods to reduce hallucination and the potential use of user input to extend the self-verification of LLMs.

As described in section 3.2 insights about interaction approaches will be transferred to the No-IDLE MEETS CHATGPT use case from adjacent projects: For example, research into NLI framework in clinical reasoning explores user-machine interaction to interactively verify reasoning conclusions allowing users to directly edit or critique the model's decision-making results. Different feedback approaches such as direct editing and corrective feedback have been implemented. These methods allow users to provide detailed input, which the system uses to refine its reasoning and conclusions. In a mixed-method user study the methods for feedback are explored regarding human factors such as user experience, perceived usability, mental effort, and efficiency. These can be measured through scales such as the PAAS scale for mental load [Paas 1992], system usability scale for perceived usability [Brooke 1986] or in context of a semi-structured interview (SSI) [Blandford 2013]. The insights can then be extracted and applied to user feedback requests in digital photobook creation.

⁶There is also a GitHub implementation available, <https://github.com/spcl/graph-of-thoughts> to be tested in the context of a potential (graph) dataset generation task.

To design the interaction of the human user in the *biased journalist task* and the *head-bodies-legs task*, the design thinking framework (DTF) will be followed during the development [Dam and Siang 2021]. The DTF is especially suitable for novel and sparsely researched areas such as the integration of ChatGPT in an interactive photobook generation scenario. It represents a human-centred and iterative process. The process follows five stages that lead from the emphasis on the user's needs, defining requirements based on the insights, and ideation of solution designs up to rapid prototyping and evaluation.

At the beginning we plan to integrate ChatGPT in smaller studies that first investigate user behaviour and needs regarding interactive photobook creation in a formative and exploratory way. For the *biased journalist task* this refers to investigating how humans would rate the subset of images that the current solution would offer them and how they would rate as well as alter these subsets. Similarly, in the *head-bodies-legs task* the current capabilities of generating a coherent story between images would be investigated with users and how they would rate and alter these results. Additionally, literature regarding both tasks should be included in investigating the current state. For that, methods are applied that fit this goal such as tasks with a think-aloud protocol [Cotton and Gresty 2006] followed by a SSI [Blandford 2013] and a thematic analysis (TA) [Braun and Clarke 2012]. This is used to get deeper insight into the user's behaviour and feature needs to be able to define fitting requirements. Based on the iterative approach, more complex combinations of system features and tasks over time will follow. First qualitative investigations in the emphasise stage will range from 5 to 10 participants as suggested by [Alroobaea and Mayhew 2014].

The final study approach in the evaluation stage has a summative character. We use an experimental setup comparing several conditions for both tasks in a factorial design. Hereby, the choice and methods evaluated in conditions are still to be evolved based on the results of experimental setups 1 and 2 (see section 3.2). However, the study regarding the *biased journalist task* investigates the performance of different methods in image retrieval and entailment assessment. In comparison, the *heads-bodies-legs task* researches the use of different methods to reduce hallucination and the potential use of user input to extend the self-verification of LLMs.

For the evaluation, users will engage in either of the two tasks. Hereby, the specific conditions of each tasks are analysed by quantitative and qualitative methods. For the *biased journalist task* this might mean that the user gets a subset of images selected by the LLM from a large set and will rate this subset for example in satisfaction or accuracy. Within, the method to choose the subset is altered in various conditions to investigate the best performing algorithm. Hereby, the altered method represents the independent variable, which performance is tested. As a further step, the subsets of images could be further iterated based on human-feedback. As a metric, the times of iterations, time for the completion of choosing a subset or the error rate could be measured. Additionally, questionnaires especially for measuring user experience and usability, e.g., SUS [Brooke 1986], PANAS-X [Watson and Clark 1994] and other SDT-based [Ryan and Deci 2000] tools related to motivation and also physical and mental load (e.g., NASA-TLX [Hart 2006; Hart and Staveland 1988]) will be used to get the subjective experience of the users. A set of qualitative methods is applied to supplement the results of the questionnaires with insights into the experience of the user during the task and the ability to detect pain points as well as well-performing aspects. For that, the Think-aloud-Protocol is applied during the task and a SSI with a follow-up reflexive thematic analysis.

Similarly, the *heads-bodies-legs task* is evaluated. Hereby, one possibility for an independent variable is to alter the starting preferences that the algorithm starts the generation with. For example, the user could get a set of keywords by the LLM and choose some, or they get a graphical representation, or as a baseline, get no possibility to input preferences. The generated results could then be rated in subjective questionnaires by their hallucination or accuracy for the user. Additionally, the process of choosing preferences could be investigated in questionnaires or qualitatively as introduced

above. Another independent variable could be the investigation of the verification loop: After receiving the first version of the generated story by the LLM corresponding to the three images, the user can give feedback that is then integrated in the next iteration of the result. Hereby, the hallucination could be investigated based on the number of corrections that the user inputs. Further, similar to the first task, the times of iterations to reach the goal state, completion time or the error rate of the system could be measured. Additionally, subjective questionnaires as introduced could measure the user experience and preferred interaction method of the users. Through qualitative methods, while applying the Think-aloud-Protocol, a better insight into pain points of the interaction design could be identified. The final user studies can then be oriented toward a larger user group up to 20 participants.

5 CONCLUSION

In this DFKI technical report, we presented the anatomy of the NO-IDLE MEETS CHATGPT prototype system (funded by the German Federal Ministry of Education and Research) with the goal to leverage the opportunities arising from large language models and technologies for the No-IDLE project. One of the key innovations described in this technical report is a methodology including benchmark studies for interactive machine learning combined with LLMs. Our main research question, how ChatGPT and other variants can help improve the accuracy of (semi-) automatic subtasks in image retrieval, captioning, and person/scene recognition, has been described. Future work will investigate, first and foremost, coping with hallucination, outdated knowledge, and non-transparent reasoning processes. Retrieval-Augmented Generation (RAG) has emerged as a promising approach by incorporating external knowledge [Gao et al. 2024; Wang et al. 2024b]. Other related topics include interactive 3D image segmentation [Shen et al. 2024], compositional reasoning with LLMs [Lu et al. 2023], and reasoning segmentation via LLMs [Lai et al. 2024].

ACKNOWLEDGMENTS

This work is funded by the German Federal Ministry of Education and Research under grant number 01IW23002.

REFERENCES

- Roobaea Alroobaea and Pam J. Mayhew. 2014. How many participants are really enough for usability studies?. In *2014 Science and Information Conference*. IEEE, 48–56.
- Aliki Anagnostopoulou, Thiago S. Gouvêa, and Daniel Sonntag. 2024. Enhancing Journalism with AI: A Study of Contextualized Image Captioning for News Articles using LLMs and LMMs. In *IJCAI 2024 Workshop on Trustworthy Interactive Decision-Making with Foundation Models*. <https://doi.org/10.48550/arXiv.2408.04331>
- Aliki Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. 2023. Towards Adaptable and Interactive Image Captioning with Data Augmentation and Episodic Memory. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustainNLP)*. Association for Computational Linguistics, Toronto, Canada (Hybrid), 245–256. <https://doi.org/10.18653/v1/2023.sustainlp-1.19>
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. 2023. DigiFace-1M: 1 Million Digital Face Images for Face Recognition. In *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.
- Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (Tokyo, Japan) (ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 279–283. <https://doi.org/10.1145/2993148.2993165>
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *CoRR* abs/2308.09687 (2023). <https://doi.org/10.48550/ARXIV.2308.09687> arXiv:2308.09687
- Rajarshi Biswas, Michael Barz, and Daniel Sonntag. 2020. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI - Künstliche Intelligenz, German Journal on Artificial Intelligence - Organ des Fachbereiches "Künstliche Intelligenz" der Gesellschaft für Informatik e.V. (KI)* 36 (7 2020), 1–14.

- Ali Furkan Biten, Lluís Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! Context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12466–12475.
- Ann E. Blandford. 2013. Semi-structured qualitative studies. Interaction Design Foundation.
- Susanne Boll, Philipp Sandhaus, Ansgar Scherp, and Sabine Thieme. 2006. MetaXa—Context- and Content-Driven Metadata Enhancement for Personal Photo Books. In *Advances in Multimedia Modeling (Lecture Notes in Computer Science)*, Tat-Jen Cham, Jianfei Cai, Chitra Dorai, Deepu Rajan, Tat-Seng Chua, and Liang-Tien Chia (Eds.). Springer, Berlin, Heidelberg, 332–343. https://doi.org/10.1007/978-3-540-69423-6_33
- Susanne Boll, Philipp Sandhaus, Ansgar Scherp, and Utz Westermann. 2007. Semantics, content, and structure of many for the creation of personal photo albums. In *Proceedings of the 15th ACM international conference on Multimedia (MM '07)*. Association for Computing Machinery, New York, NY, USA, 641–650. <https://doi.org/10.1145/1291233.1291385>
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).
- Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- John Brooke. 1986. System usability scale (SUS): A quick-and-dirty method of system evaluation user information. *Reading, UK: Digital equipment co ltd* 43 (1986), 1–7.
- 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. 2020. Language Models are Few-Shot Learners. *CoRR abs/2005.14165* (2020). arXiv:2005.14165 <https://arxiv.org/abs/2005.14165>
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. *Advances in Neural Information Processing Systems* 31 (2018), 9539–9549.
- Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*. IEEE Computer Society, 67–74. <https://doi.org/10.1109/FG.2018.00020>
- Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9962–9971.
- Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023a. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 14948–14968. <https://aclanthology.org/2023.emnlp-main.925>
- Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023b. Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models. arXiv:2309.04461 [cs.CL]
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. arXiv:2309.15402 [cs.CL]
- Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 895–903.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *CoRR abs/1803.05457* (2018). arXiv:1803.05457 <http://arxiv.org/abs/1803.05457>
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
- Gianna M. Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, Allan Ellis and Tatsuya Hagino (Eds.). ACM, 97–106. <https://doi.org/10.1145/1060745.1060764>
- Deborah Cotton and Karen Gresty. 2006. Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology* 37, 1 (2006), 45–54.
- Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2432–2443. <https://doi.org/10.1109/CVPR.2017.261>
- Rikke F. Dam and Teo Y. Siang. 2021. What is design thinking and why is it so popular?
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495* (2023).
- Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations. *arXiv preprint arXiv:2004.03744* (2020).

- Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2Commonsense: Generating Commonsense Descriptions to Enrich Video Captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 840–860. <https://doi.org/10.18653/v1/2020.emnlp-main.61>
- Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard Hovy. 2023. CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 313–327. <https://doi.org/10.18653/v1/2023.eacl-main.24>
- Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3137–3146.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL] <https://arxiv.org/abs/2312.10997>
- Diego García-Olano, Yasumasa Onoe, and Joydeep Ghosh. 2022. Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 705–715. <https://doi.org/10.1145/3487553.3524648>
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* 9 (2021), 346–361. https://doi.org/10.1162/tac1_a_00370
- Antonio Gulli. 2005. The anatomy of a news search engine. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005 - Special interest tracks and posters*, Allan Ellis and Tatsuya Hagino (Eds.). ACM, 880–881. <https://doi.org/10.1145/1062745.1062778>
- Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. MSCap: Multi-Style Image Captioning with Unpaired Stylized Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4204–4213.
- Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611* (2022).
- Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908. Issue: 9.
- Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, Vol. 52. Elsevier, 139–183.
- Mareike Hartmann, Aliko Anagnostopoulou, and Daniel Sonntag. 2022a. Interactive Machine Learning for Image Captioning. *CoRR* abs/2202.13623 (2022). arXiv:2202.13623 <https://arxiv.org/abs/2202.13623>
- Mareike Hartmann, Han Du, Nils Feldhus, Ivana Kruijff-Korbayová, and Daniel Sonntag. 2022b. XAINES: Explaining AI with Narratives. *KI - Künstliche Intelligenz* 36, 3 (Dec. 2022), 287–296. <https://doi.org/10.1007/s13218-022-00780-8>
- Syed Zohaib Hassan, Kashif Ahmad, Steven Hicks, Pål Halvorsen, Ala I. Al-Fuqaha, Nicola Conci, and Michael Riegler. 2022. Visual Sentiment Analysis from Disaster Images in Social Media. *Sensors* 22, 10 (2022), 3628. <https://doi.org/10.3390/S22103628>
- Niels Henze and Susanne Boll. 2011. Who’s That Girl? Handheld Augmented Reality for Printed Photo Books. In *Human-Computer Interaction – INTERACT 2011 (Lecture Notes in Computer Science)*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer, Berlin, Heidelberg, 134–151. https://doi.org/10.1007/978-3-642-23765-2_10
- Josh Howarth. 2024. Number of Parameters in GPT-4 (Latest Data). <https://explodingtopics.com/blog/gpt-parameters>
- Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 12031–12041. <https://doi.org/10.1109/ICCV51070.2023.01108>
- Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 1049–1065. <https://doi.org/10.18653/v1/2023.findings-acl.67>
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual Storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1233–1239. <https://doi.org/10.18653/v1/N16-1147>
- Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 6700–6709. <https://doi.org/10.1109/CVPR.2019.00686>
- Edwin L. Hutchins, James Hollan, and Donald A. Norman. 1985. Direct Manipulation Interfaces. *Hum. Comput. Interact.* 1 (1985), 311–338. <https://api.semanticscholar.org/CorpusID:16355120>
- Aman Jain, Mayank Kothiyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 2491–2498. <https://doi.org/10.1145/3404835.3463259>
- Zhengxiong Jia and Xirong Li. 2020. iCap: Interactive Image Captioning with Predictive Text. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 428–435.

- Tianyu Jiang and Ellen Riloff. 2023. Exploiting Commonsense Knowledge about Objects for Visual Activity Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 7277–7285. <https://doi.org/10.18653/v1/2023.findings-acl.457>
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1988–1997. <https://doi.org/10.1109/CVPR.2017.215>
- Maël Jullien, Marco Valentino, Hannah Frost, Paul O’regan, Donal Landers, and André Freitas. 2023. SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, Toronto, Canada, 2216–2226. <https://doi.org/10.18653/v1/2023.semeval-1.307>
- Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In So Kweon. 2020. Hide-and-Tell: Learning to Bridge Photo Streams for Visual Storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (Apr. 2020), 11213–11220. <https://ojs.aaai.org/index.php/AAAI/article/view/6780>
- Hannes Kath, Bengt Lüers, Thiago S. Gouvêa, and Daniel Sonntag. 2023a. A Virtual Reality Tool for Representing, Visualizing and Updating Deep Learning Models. <https://doi.org/10.48550/arXiv.2305.15353> arXiv:2305.15353 [cs].
- Hannes Kath, Thiago S. Gouvêa, and Daniel Sonntag. 2023b. A Human-in-the-Loop Tool for Annotating Passive Acoustic Monitoring Datasets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 7140–7144. <https://doi.org/10.24963/ijcai.2023/835> 3 citations (Semantic Scholar/DOI) [2023-09-15].
- Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV]
- Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1244–1254.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. 787–798. <https://doi.org/10.3115/v1/D14-1086>
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- Anis Koubaa. 2023. GPT-4 vs. GPT-3.5: A Concise Showdown. (4 2023). <https://doi.org/10.36227/techrxiv.22312330.v2>
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yanniss Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision* 123, 1 (may 2017), 32–73. <https://doi.org/10.1007/s11263-016-0981-7>
- Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto, Toronto, Ontario. <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* (2020).
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. LISA: Reasoning Segmentation via Large Language Model. arXiv:2308.00692 [cs.CV] <https://arxiv.org/abs/2308.00692>
- Chunyun Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890* (2023).
- Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. 2022a. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv preprint arXiv:2205.12005* (2022).
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.
- Qing Li, Qingyi Tao, Shafiq R. Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 11211)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 570–586. https://doi.org/10.1007/978-3-030-01234-2_34
- Ruiqi Li, Xiang Zhao, and Marie-Francine Moens. 2022b. A Brief Overview of Universal Sentence Representation Methods: A Linguistic View. *ACM Computing Surveys (CSUR)* 55 (2022), 1 – 42. <https://api.semanticscholar.org/CorpusID:247753662>
- Shan Li and Weihong Deng. 2019. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Transactions on Image Processing* 28, 1 (2019), 356–370.
- Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2584–2593.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

- Jingzhou Liu, Wenhui Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A Large-Scale Dataset for Video-and-Language Inference. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 10897–10907. <https://doi.org/10.1109/CVPR42600.2020.01091>
- Jochem Loedeman, Maarten C. Stol, Tengda Han, and Yuki M. Asano. 2022. Prompt Generation Networks for Efficient Adaptation of Frozen Vision Transformers. *CoRR* abs/2210.06466 (2022). <https://doi.org/10.48550/arXiv.2210.06466> arXiv:2210.06466
- Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware Image Caption Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4013–4023. <https://doi.org/10.18653/v1/D18-1435>
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. arXiv:2304.09842 [cs.CL] <https://arxiv.org/abs/2304.09842>
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful Chain-of-Thought Reasoning. arXiv:2301.13379 [cs.CL]
- Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zolt Kira. 2020. Learning to Generate Grounded Visual Captions Without Localization Supervision. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII (Lecture Notes in Computer Science, Vol. 12363)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 353–370. https://doi.org/10.1007/978-3-030-58523-5_21
- Bill MacCartney and Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Coling 2008 Organizing Committee, Manchester, UK, 521–528. <https://aclanthology.org/C08-1066>
- Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 1682–1690. <http://papers.nips.cc/paper/5411-a-multi-world-approach-to-question-answering-about-real-world-scenes-based-on-uncertain-input.pdf>
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Matthias Maszuhn, Larbi Abdenebaoui, and Susanne Boll. 2021. A User-Centered Approach for Recognizing Convenience Images in Personal Photo Collections. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–4. <https://doi.org/10.1109/CBMI50038.2021.9461908> ISSN: 1949-3991.
- Alexander Mathews, Lexing Xie, and Xuming He. 2016. SentiCap: Generating Image Descriptions with Sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence* 30, 1 (Mar. 2016). <https://doi.org/10.1609/aaai.v30i1.10475>
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. <https://doi.org/10.18653/v1/2020.acl-main.173>
- Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* 10, 1 (Jan 2019), 18–31. <https://doi.org/10.1109/TAFFC.2017.2740923>
- Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. *ArXiv* abs/2301.13294 (2023).
- Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023a. How well do SOTA legal reasoning models support abductive reasoning?. In *Proceedings of the International Conference on Logic Programming 2023 Workshops co-located with the 39th International Conference on Logic Programming (ICLP 2023), London, United Kingdom, July 9th and 10th, 2023 (CEUR Workshop Proceedings, Vol. 3437)*, Joaquín Arias, Sotiris Batsakis, Wolfgang Faber, Gopal Gupta, Francesco Pacenza, Emmanouel Papadakis, Livio Robaldo, Kilian Rückschloß, Elmer Salazar, Zeynep G. Saribatur, Ilias Tachmazidis, Felix Weitkämper, and Adam Z. Wyner (Eds.). CEUR-WS.org. <https://ceur-ws.org/Vol-3437/paper1LPLR.pdf>
- Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023b. Black-Box Analysis: GPTs Across Time in Legal Textual Entailment Task. arXiv:2309.05501 [cs.CL]
- Ha Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023c. A negation detection assessment of GPTs: analysis with the xNot360 dataset. arXiv:2306.16638 [cs.CL]
- OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- Sharon Oviatt, Björn Schuller, Philip Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. 2019. *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*. Morgan & Claypool.
- Fred GWC Paas. 1992. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *Journal of Educational Psychology* 84, 4 (1992), 429.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. VisualCOMET: Reasoning about the Dynamic Context of a Still Image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

- Dylan Patel and Gerald Wong. 2023. GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. <https://www.semianalysis.com/p/gpt-4-architecture-infrastructure>
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint arXiv:2305.03495* (2023).
- Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2011a. Automatic creation of photo books from stories in social media. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7S, 1 (Nov. 2011), 27:1–27:18. <https://doi.org/10.1145/2037676.2037684>
- Mohamad Rabbath, Philipp Sandhaus, and Susanne Boll. 2011b. Multimedia retrieval in social networks for photo book creation. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR '11)*. Association for Computing Machinery, New York, NY, USA, 1–2. <https://doi.org/10.1145/1991996.1992068>
- Krishnan Ramnath, Simon Baker, Lucy Vanderwende, Motaz El-Saban, Sudipta N Sinha, Anitha Kannan, Noran Hassan, Michel Galley, Yi Yang, Deva Ramanan, et al. 2014. Autocaption: Automatic caption generation for personal photos. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1050–1057.
- realsbstn (YouTube channel). 2009. *Usain Bolt 9.58 100m New World Record Berlin [HQ]*. Youtube. https://www.youtube.com/watch?v=3nbjhcZ9_g
- Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1586–1596. <https://doi.org/10.18653/v1/D18-1187>
- Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.
- Malik Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 11, 6 (Mar 2023), 887. <https://doi.org/10.3390/healthcare11060887>
- Philipp Sandhaus and Susanne Boll. 2011. Semantic analysis and retrieval in personal and social photo collections. *Multimedia Tools and Applications* 51, 1 (Jan. 2011), 5–33. <https://doi.org/10.1007/s11042-010-0673-1>
- Philipp Sandhaus, Mohammad Rabbath, and Susanne Boll. 2011. Employing Aesthetic Principles for Automatic Photo Book Layout. In *Advances in Multimedia Modeling (Lecture Notes in Computer Science)*, Kuo-Tien Lee, Wen-Hsiang Tsai, Hong-Yuan Mark Liao, Tsuhan Chen, Jun-Wei Hsieh, and Chien-Cheng Tseng (Eds.). Springer, Berlin, Heidelberg, 84–95. https://doi.org/10.1007/978-3-642-17832-0_9
- Philipp Sandhaus, Sabine Thieme, and Susanne Boll. 2008. Processes of photo book production. *Multimedia Systems* 14, 6 (Dec. 2008), 351–357. <https://doi.org/10.1007/s00530-008-0136-y>
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761* (2023).
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. *arXiv* (2022).
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2556–2565. <https://doi.org/10.18653/V1/P18-1238>
- Chuyun Shen, Wenhao Li, Yuhang Shi, and Xiangfeng Wang. 2024. Interactive 3D Medical Image Segmentation with SAM 2. *arXiv:2408.02635* [cs.CV] <https://arxiv.org/abs/2408.02635>
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. <http://arxiv.org/abs/2010.15980> arXiv:2010.15980 [cs].
- Ben Shneiderman. 1983. Direct Manipulation: A Step Beyond Programming Languages. *Computer* 16 (1983), 57–69. <https://api.semanticscholar.org/CorpusID:14942172>
- Chandan Singh, John Xavier Morris, Jyoti Aneja, Alexander M Rush, and Jianfeng Gao. 2023. Explaining Patterns in Data with Language Models via Interpretable Autoprompting. <https://openreview.net/forum?id=GvMuB-YsiK6>
- Daniel Sonntag. 2004. Assessing the Quality of Natural Language Text Data. In *34. Jahrestagung der Gesellschaft für Informatik, Informatik verbindet, INFORMATIK 2004, Ulm, Germany, September 20-24, 2004, Band 1 (LNI, Vol. P-50)*, Peter Dadam and Manfred Reichert (Eds.). GI, 259–263. <https://dl.gi.de/handle/20.500.12116/28866>
- Daniel Sonntag, Michael Barz, and Thiago Gouvea. 2024. *A look under the hood of the Interactive Deep Learning Enterprise (No-IDLE)*. Technical Report. German Research Center for Artificial Intelligence.
- Daniel Sonntag, Ralf Engel, Gerd Herzog, Alexander Pfalzgraf, Norbert Pfeifer, Massimo Romanelli, and Norbert Reithinger. 2007. SmartWeb Handheld - Multimodal Interaction with Ontological Knowledge Bases and Semantic Web Services. In *Artificial Intelligence for Human Computing, ICMI 2006 and IJCAI 2007 International Workshops, Banff, Canada, November 3, 2006, Hyderabad, India, January 6, 2007, Revised Selected and Invited Papers*. 272–295. https://doi.org/10.1007/978-3-540-72348-6_14
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4149–4158. <https://doi.org/10.18653/v1/N19-1421>
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL-HLT*.
- Henning Wachsmuth and Milad Alshomary. 2022. “Mama Always Had a Way of Explaining Things So I Could Understand”: A Dialogue Corpus for Learning to Construct Explanations. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 344–354. <https://aclanthology.org/2022.coling-1.27>
- Wolfgang Wahlster and Mark Maybury. 1998. Intelligent User Interfaces: An Introduction. *RUIU* (1998), 1–13.
- Chengrui Wang, Qingqing Long, Meng Xiao, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. 2024b. BioRAG: A RAG-LLM Framework for Biological Question Reasoning. arXiv:2408.01107 [cs.CL] <https://arxiv.org/abs/2408.01107>
- Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 10 (2018), 2413–2427. <https://doi.org/10.1109/TPAMI.2017.2754246>
- Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 9185–9192.
- William Yang Wang. 2017. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, 422–426. <https://doi.org/10.18653/V1/P17-2067>
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *CoRR* abs/2203.11171 (2022). <https://doi.org/10.48550/ARXIV.2203.11171> arXiv:2203.11171
- Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024a. Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning. <https://synthical.com/article/abbcb13f-508f-4490-9eed-60917065b90c>. arXiv:2401.06805 [cs.AI]
- David Watson and Lee Anna Clark. 1994. The PANAS-X: Manual for the positive and negative affect schedule-expanded form. (1994). Publisher: University of Iowa.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent Abilities of Large Language Models. *CoRR* abs/2206.07682 (2022). <https://doi.org/10.48550/arXiv.2206.07682> arXiv:2206.07682
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michèle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0. <https://catalog.ldc.upenn.edu/LDC2013T19>
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2550–2575.
- Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint arXiv:2302.11382* (2023).
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. <https://doi.org/10.18653/v1/N18-1101>
- Ani Withöft, Larbi Abdenebaoui, and Susanne Boll. 2022. ILMICA - Interactive Learning Model of Image Collage Assessment: A Transfer Learning Approach for Aesthetic Principles. In *MultiMedia Modeling (Lecture Notes in Computer Science)*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet (Eds.). Springer International Publishing, Cham, 84–96. https://doi.org/10.1007/978-3-030-98355-0_8
- Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. 2023. ScatterShot: Interactive In-context Example Curation for Text Transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 353–367.
- Zhuofeng Wu, Sinong Wang, Jiatao Gu, Rui Hou, Yuxiao Dong, V.G.Vinod Vydiswaran, and Hao Ma. 2022. IDPG: An Instance-Dependent Prompt Generation Method. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 5507–5521. <https://doi.org/10.18653/v1/2022.naacl-main.403>
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706* (2019).
- Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (Lille, France) (ICML’15)*. JMLR.org, 2048–2057.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601* (2023).

- Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo, and Seung-won Hwang. 2018. Visual Choice of Plausible Alternatives: An Evaluation of Image-based Commonsense Causal Reasoning. In *LREC*.
- Brett Young. 2021. AI Expert Speculates on GPT-4 Architecture. <https://wandb.ai/byoung3/ml-news/reports/AI-Expert-Speculates-on-GPT-4-Architecture---Vmlldzo0NzA0Nzg4>
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2* (2014), 67–78.
- Jan Zacharias, Michael Barz, and Daniel Sonntag. 2018. A survey on deep learning toolkits and libraries for intelligent user interfaces. *arXiv preprint arXiv:1803.04818* (2018).
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 35–45. <https://doi.org/10.18653/v1/D17-1004>
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=5NTt8GFjUHkr>
- Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020b. Evaluating Commonsense in Pre-Trained Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 05 (April 2020), 9733–9740. <https://doi.org/10.1609/aaai.v34i05.6523>
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023a. Large Language Models Are Human-Level Prompt Engineers. <http://arxiv.org/abs/2211.01910> arXiv:2211.01910 [cs].
- Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020a. More Grounded Image Captioning by Distilling Image-Text Matching Model. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 4776–4785. <https://doi.org/10.1109/CVPR42600.2020.00483>
- Yulin Zhou, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Yarin Gal. 2023b. Revisiting Automated Prompting: Are We Actually Doing Better?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, 1822–1832. <https://doi.org/10.18653/v1/2023.acl-short.155>
- Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. 2022. Generalized Decoding for Pixel, Image, and Language. *CoRR abs/2212.11270* (2022). <https://doi.org/10.48550/arXiv.2212.11270> arXiv:2212.11270