Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

EXPLORING THE USE OF GPT-4 FOR X-RAY PHOTOELECTRON SPECTROSCOPY DATA ANALYSIS

Relatore:
Chiar.mo Prof.
ANDREA ASPERTI

Correlatore:
Dott.ssa
PAOLA DE CECCO

Presentata da:
DAVIDE LUCCIOLI

I Sessione
Anno Accademico 2023-2024

Ai miei genitori
Sabina e Luigi,
ai miei nonni.

Sommario

Questa tesi indaga l'applicazione dei Large Language Models (LLM) nel campo dell'analisi dei dati derivati dalla Spettroscopia Fotoelettronica a Raggi X (XPS). In collaborazione con Applied Materials Inc., lo studio utilizza GPT-4, accessibile attraverso ChatGPT, per valutarne le capacità e i limiti nel consultare vaste basi di conoscenza ed estrarre informazioni specifiche dalla letteratura scientifica. L'obiettivo è valutare il potenziale degli LLM nell'assistere gli ingegneri dei materiali nell'interpretazione di dati complessi ottenuti dalla XPS, fondamentale per la produzione di dispositivi a semiconduttore. Il primo capitolo fornisce una panoramica sull'evoluzione degli LLM, discutendo le architetture chiave e gli sviluppi nel campo. Il secondo capitolo introduce la Spettroscopia Fotoelettronica a Raggi X, spiegando i principi della tecnica, la generazione dei fotoelettroni, gli spostamenti chimici e le sfide nell'analisi degli spettri XPS. Il terzo capitolo dettaglia il setup sperimentale, inclusa la selezione degli articoli di ricerca, i metodi di estrazione dei dati e le sfide affrontate durante lo studio. Infine, il quarto capitolo presenta i risultati degli esperimenti, discutendo le prestazioni dell'LLM nell'estrazione degli spostamenti chimici da articoli singoli e multipli, e conclude con riflessioni sulle capacità e i limiti degli LLM nell'analisi dei dati XPS, nonché potenziali future direzioni di ricerca in questo campo.

Contents

Introduction

This thesis explores the integration of Large Language Models (LLMs) into the analysis of X-ray Photoelectron Spectroscopy (XPS) data, assessing their capabilities and limitations in querying knowledge bases and extracting chemical shift information from research papers. The goal is to evaluate the potential of LLMs to assist material engineers in interpreting complex XPS data. This work has been conducted in collaboration with Applied Materials Inc., an American corporation that supplies equipment, services, and software for the manufacture of semiconductor chips for electronics.

Material engineering for semiconductor devices requires precise measurements of uneven and ultra-thin (often less than a nanometer thick) layers of material through the use of various surface analysis techniques, including XPS, a widely used method for determining the elemental composition and chemical states present on a material's surface. The complexity of these measurements and the vast diversity of processes, materials, and experiments makes the volume and intricacy of past knowledge largely unavailable to material engineers, who must rely on their expertise and the available literature to interpret the data.

Artificial Intelligence (AI) and LLMs have experienced remarkable advancements in recent years, revolutionizing the way we interact with technology. These sophisticated models, trained on vast amounts of text data, have shown exceptional capabilities in Natural Language Processing (NLP), advancing the automation of various natural language tasks such as text summarization [1], question answering [2], and language translation [3].

Remarkably, the largest LLMs, have demonstrated a strong ability to generalize across multiple tasks [4, 5], leveraging their extensive training to achieve high performance on assignments they were not explicitly optimized for. Moreover, these massive models have exhibited emergent capabilities [6] that were not present in smaller models and seemingly arose from scaling up the model's size.

These rapid developments have sparked significant interest in integrating LLMs into several fields, including software development [7, 8], chemistry [9, 10], medicine [11, 12] and financial analysis [13], to take advantage of their language understanding and generation abilities. Similarly, in material science, an LLM-powered assistant able to select relevant information from multiple existing sources and synthesize it to extract considerations applicable to novel materials, could significantly accelerate the development cycle of semiconductor devices.

While the potential applications are vast, it is crucial to thoroughly evaluate the capabilities and limitations of LLMs within specific domains before deploying them: these models can exhibit inconsistencies, hallucinations and biases [14, 15] stemming from the data they were trained on or gaps in their knowledge. The experiments conducted aim to evaluate the model's performance in understanding complex scientific discussions, responding to specific informational requests, and processing and presenting data extracted from academic articles. By investigating the effectiveness of LLMs in this specific domain, this study provides valuable insights into their potential applications in material science and contributes to the broader understanding of their capabilities and limitations.

This thesis is organized in 4 chapters: Chapter 1 provides an overview of the evolution of Large Language Models, discussing key architectures and developments in the field. Chapter 2 introduces X-ray Photoelectron Spectroscopy, explaining its principles, the generation of photoelectrons, chemical shifts, and the challenges in XPS spectrum analysis. Chapter 3 details the experimental setup, including the selection of research papers, data extraction methods, and the challenges encountered during the study. Finally, Chapter 4 presents the results of our experiments, discussing the performance of the LLM in extracting chemical shift data from individual and multiple papers, and concludes with insights into the capabilities and limitations of LLMs in XPS data analysis, as well as potential future directions for research in this area.

Chapter 1

Evolution of LLMs

The field of NLP has seen a significant evolution in the last decade, the development of techniques such as word2vec [16, 17] and GloVe [18] introduced new architectures to compute word embeddings, which allowed for the representation of words as vectors in a high-dimensional space, capturing the semantic relationships between them. The adoption of embeddings led to the development of pre-trained language models, early examples of which include ELMo [19], based on a bidirectional Long Short-Term Memory (LSTM) architecture [20]. However, it was the introduction of the transformer architecture [21] that laid the foundations for the development of modern LLMs.

1.1 Transformers

Transformers are a type of neural network architecture which relies on self-attention, a mechanism that allows the model to weigh the importance of different parts of the input sequence in relation to each other when making predictions, thus enabling the model to capture long-range dependencies in the data. Using an encoder-decoder architecture, transformers create embeddings of the input, capturing the semantic and positional information of the tokens in the sequence, these embeddings are then weighted by the self-attention layers, based on the other tokens in the input sequence, to pro-

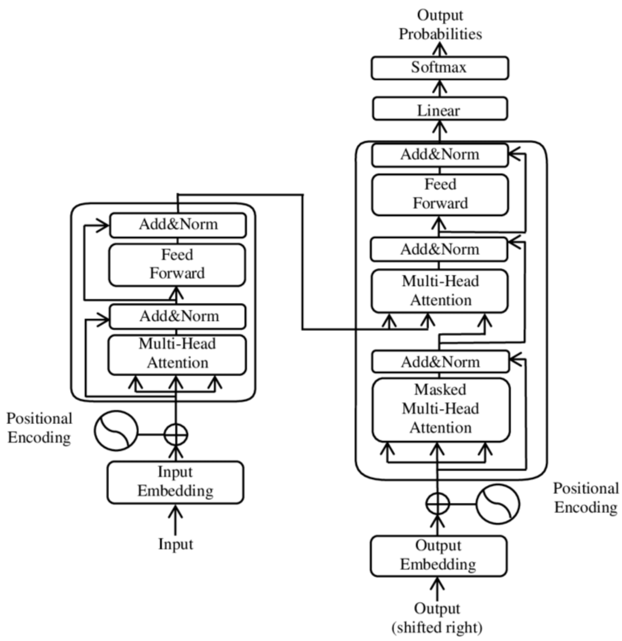


Figure 1.1: The Transformer architecture. By Yuening Jia - DOI:10.1088/1742-6596/1314/1/012186, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=121340680

duce the encoded representation of the input. The decoder then uses this representation, along with that of previous outputs, to generate the output sequence. Transformers were shown to outperform previous state-of-the-art models for NLP tasks.

1.1.1 Self-Attention

The attention mechanism was first introduced in models using recurrent neural networks to select important information from the input sequence, self-attention, the core mechanism of transformers, extends this concept allowing the network to discern internal correlations between the elements of a sequence and dynamically adjust their importance. The attention mechanism used in transformers is computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1.1}$$

where:

- Q, K, and V are the query, key, and value matrices, respectively, obtained by multiplying the input sequence's embeddings by the weight matrices learned by the model.
- d_k is the dimension of the query and key vectors.

1.1.2 Multi-Headed Attention

Transformers further extend the self-attention mechanism by applying it multiple times in parallel through a process called multi-headed attention, by which the outputs of the different attention heads are concatenated and linearly transformed to produce the final output. Thus, the model can capture different aspects of the input sequence, allowing it to obtain a more comprehensive understanding of the data.

1.1.3 Positional Embedding

As seen in previous architectures, transformers rely on embeddings to represent the input sequence. However, since the model does not have any recursion or convolutions, it lacks the ability to capture the order of the tokens within the sequence. Hence, positional information must be added to the input through a technique known as positional embedding, which involves creating fixed-length vectors that encode the position of each token in the sequence, these vectors are then added to the input embeddings; in doing so the information obtained from the positional embeddings is combined with the semantic information of the token embeddings, allowing the model to understand the sequential order of tokens in the input.

1.1.4 Encoder

The encoder of a transformer is composed of a stack of identical layers, each of which consists of two sub-layers: a multi-headed self-attention mechanism and a feed-forward neural network. The self-attention module captures the relationships between the different tokens in the input sequence, while the feed-forward neural network processes the output of the previous module to produce the final representation of the input. The output of the encoder is then passed to the decoder, which generates the output sequence.

1.1.5 Decoder

The decoder of a transformer is similar to the encoder, with the addition of a third sub-layer, which performs multi-headed attention over the encoder's output, this enables the decoder to integrate information from the encoder with its own outputs from the previous time step, thus generating the output sequence autoregressively. Additionally, by masking the self-attention module, the decoder is only allowed to perform self-attention on the tokens that precede the current position in the sequence, ensuring that only the information available at the time of prediction is used to generate the output.

1.2 LLM Architectures

Transformers have played a foundational role in the development of LLMs, primarily due to their ability to scale efficiently with increasing data and computational resources, enabling models to scale to billions of parameters and effectively capture the complexity of natural language data. The performance of LLMs has thus seen significant progress with the development of ever-larger models. While transformers are fundamentally an encoder-decoder model, researchers have experimented with different architectures derived from the original transformer, thus Large Language Models can be classified, depending on the underlying architecture, in three main categories:

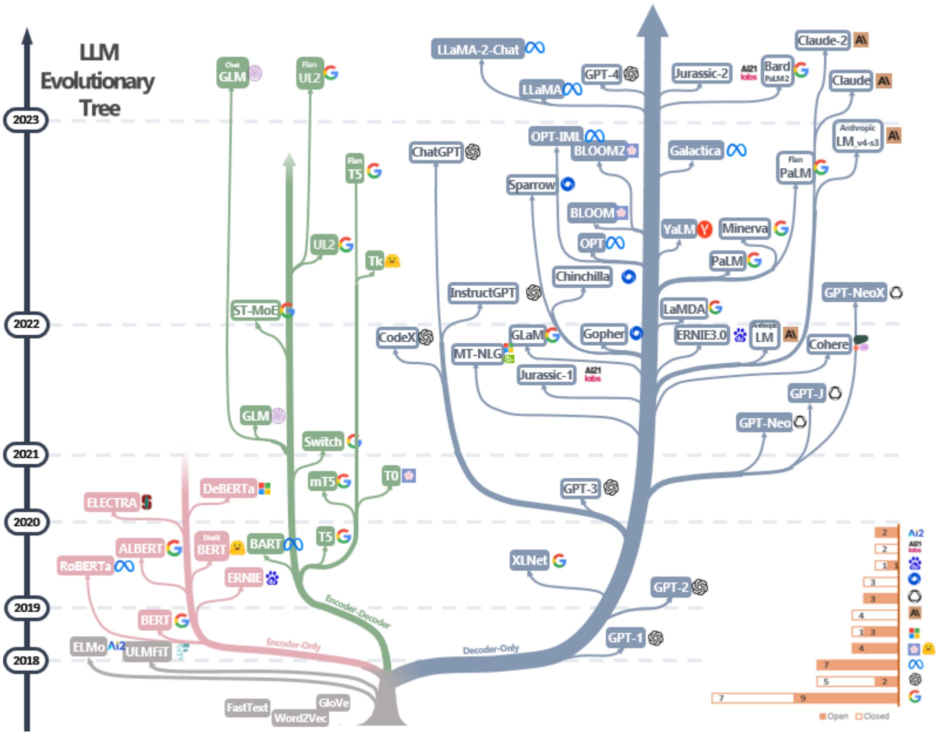


Figure 1.2: LLM evolutionary tree [22].

- Encoder-only models
- Encoder-decoder models
- Decoder-only models

1.2.1 Encoder-Only Models

Encoder-only models focus on encoding the input text into a meaningful representation, but do not include a decoding component; they are largely trained using a technique called Masked Language Modeling (MLM), where a percentage of the input is masked, and the model is trained to predict the original tokens. One of the first examples of transformer-based models is BERT (Bidirectional Encoder Representations from Transformers) [23], an encoder-only model, which brought significant improvements over previous state-of-the-art techniques. Following BERT's success, several models, such as ALBERT [24], have built upon this architecture improving its performance. Despite these advancements, the application of encoder-only configurations in LLMs has seen a decline, since the lack of a decoder module limits the model's ability to generate text.

1.2.2 Encoder-Decoder Models

LLMs built on the encoder-decoder architecture are based entirely on the traditional transformer model, with the encoder applying self-attention to the input and the decoder generating the output in an autoregressive manner. The most notable examples of this architecture are T5 [25] and BART [26].

1.2.3 Decoder-Only Models

The decoder-only architecture is currently the most popular for building LLMs, in this configuration the model is trained to sequentially generate text attending to previous tokens, without the need for an encoder module. The most famous example of this architecture is the GPT (Generative Pre-trained Transformer) family of models, including the GPT-3.5 subfamily and the more recent GPT-4 [27], respectively powering the free and plus versions of OpenAI's ChatGPT. The decoder-only architecture has been widely adopted in other models, such as Anthropic's Claude and Meta's LLaMA [28], showing impressive performance across a wide range of benchmarks.

For the purposes of this work, the model used is GPT-4, accessed through the ChatGPT plus subscription. The decision to employ GPT-4 was based on its state-of-the-art performance in understanding and generating human-like text and its training on a diverse and extensive corpus, including a wide range of scientific materials. Furthermore, GPT-4 has shown better capabilities in various reasoning tasks, compared to other models [29], making it an interesting candidate for this study.

Chapter 2

X-ray Photoelectron Spectroscopy

X-ray Photoelectron Spectroscopy (XPS) is a surface sensitive analytical technique, based on the photoelectric effect, in which X-rays are used to irradiate a material, causing the emission of photoelectrons from the surface atoms and allowing for the measurement of their kinetic energies. The main characteristic of XPS is its ability to provide information about the elemental composition, chemical state, and electronic state of the elements present on the surface of a material.

2.1 Generation of photoelectrons

When a material is exposed to an electromagnetic radiation, such as X-rays, electrons within its atoms absorb the incident radiation. This absorption can cause the electrons to gain sufficient energy to overcome the binding forces holding them within their respective orbitals; this phenomenon, known as the photoelectric effect, results in the ejection of electrons from the material's surface. The kinetic energy of the emitted electrons, also called photoelectrons, is directly related to the energy of the X-ray photons used to eject them, and the binding energy of the electrons within the material;

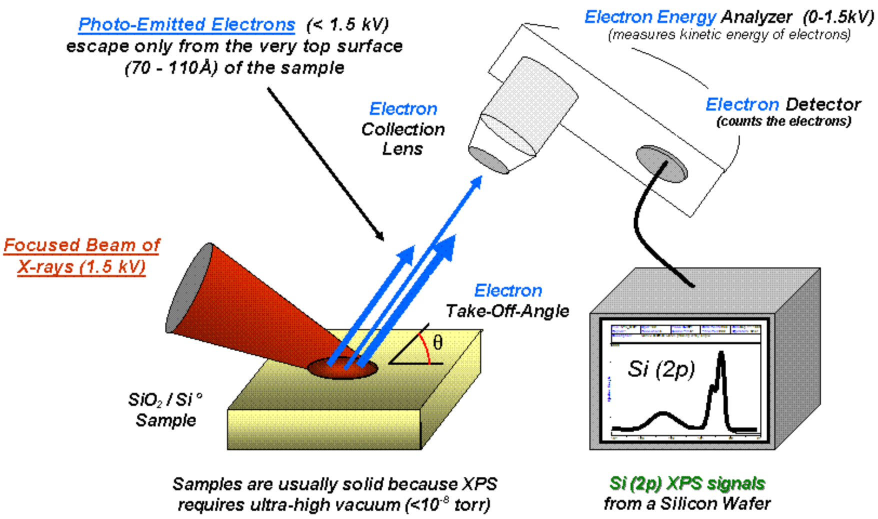


Figure 2.1: Schematic representation of the XPS system.

such relationship is expressed by the following equation [30]:

$$E_B = hv - E_k - \varphi \tag{2.1}$$

where:

- E_B is the binding energy of the electron within the material.
- hv is the energy of the X-ray photon used to eject the electron.
- E_K is the kinetic energy of the emitted photoelectron.
- φ is a constant value known as the spectrometer work function.

This equation forms the basis for the XPS technique: it enables a detailed analysis of the material's surface as provided by the XPS spectrum, which records the intensities of photoelectrons as a function of their binding or kinetic energy, revealing the elemental composition and chemical states on the surface of the sample trough the characteristic peaks in the spectrum. A typical XPS spectrum is a plot of the number of electrons detected at

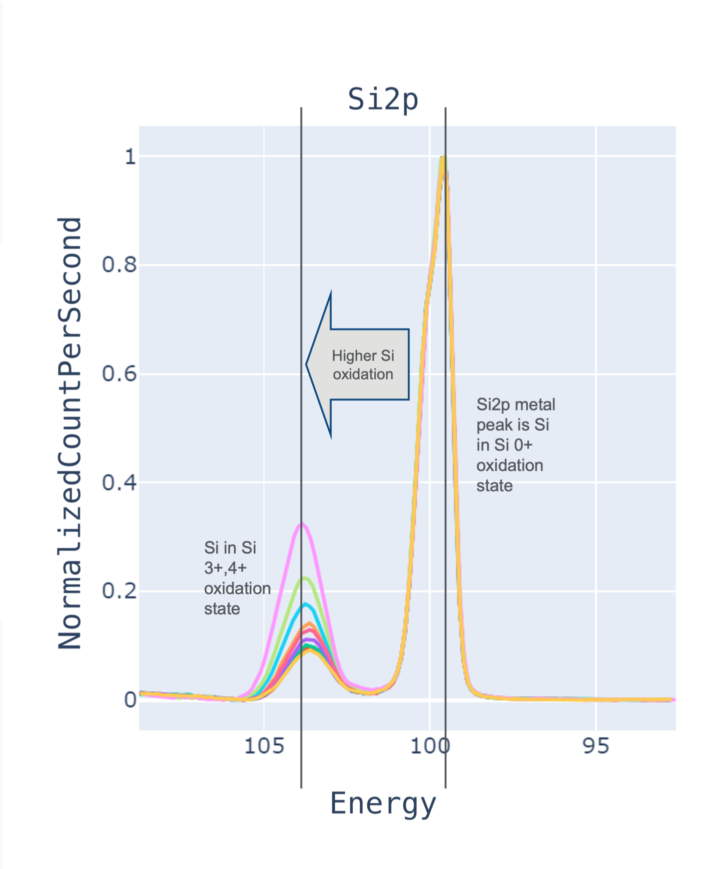


Figure 2.2: XPS spectrum of the $2p$ electrons of an Si sample.

a specific binding energy, usually focusing on regions surrounding selected electronic transitions, thus only capturing electrons ejected from specific atomic orbitals; distinct peaks in the spectrum correlate to characteristic electron configurations of different elements. This spectral signature enables the identification and quantification of elements present on the surface of the material. Figure 2.2 shows an example of an XPS spectrum with multiple peaks at different binding energies, suggesting the presence of silicon atoms at different oxidation states.

2.2 Chemical Shifts

The binding energy of a core electron is primarily determined by the electrostatic interaction between it and the nucleus, however it may also be influenced by the chemical environment of the atom from which it originates [31]; indeed, the effective nuclear charge experienced by an electron, and thus its binding energy, can be altered by the electrostatic shielding provided by other electrons in the atom, including the valence electrons, as well as the removal or addition of electronic charge through processes such as oxidation or reduction. This means that, depending on the chemical bonds formed by an atom, the nuclear charge experienced by its core electrons may vary. Hence, when the bonds between atoms in a material change, the binding energy of the innermost electrons can also be altered, these variations in binding energy, caused by changes in chemical state, are known as chemical shifts. Chemical shifts are generally computed as the difference between the binding energy of an electron in a given chemical state and that of the same type of electron in a reference state, typically a pure element or a standard compound.

2.3 XPS Spectrum Analysis

Interpreting data from an XPS spectrum can be a challenging task that demands a deep understanding of both the underlying physics and chemistry. The process of identifying the elements present in the material and determining their chemical states requires a thorough analysis of the spectral data, including the position and shape of the peaks, as well as the chemical shifts that may be present. This data is often compared with known references from scientific literature or software tools that provide databases of XPS spectra for different elements and chemical states. The interpretation of XPS spectra is further complicated by the presence of overlapping peaks, which can make it difficult to distinguish between different elements or chemical states. To address this issue, researchers often combine XPS data with other measurements, to provide a more comprehensive understanding of the material's surface. This process is often time-consuming, requiring researchers to consult a vast amount of scientific literature to correctly interpret the data. Automating the analysis of XPS spectra through the use of LLMs could significantly reduce the time and effort required to extract pertinent findings and synthesize diverse data sets, thereby enhancing the precision of spectral analysis and contributing to deeper insights and more reliable conclusions.

Chapter 3

Experimental Setup

The goal of this study was to assess how effectively Large Language Models could comprehend and extract relevant data from research papers within the domain of X-ray Photoelectron Spectroscopy. Specifically, the experiments performed aimed to evaluate the ability of LLMs to identify and extract information on chemical shifts observed in the analyzed samples. Understanding how well current state-of-the-art technologies perform in this specific domain could provide valuable insights into their strengths and limitations, and thus guide the development of more specialized tools and models optimized for XPS data analysis.

Such tools could significantly enhance the efficiency and accuracy of spectral analysis, ultimately advancing research and applications in material science. Furthermore, evaluating the performance of LLMs in this context contributes to the broader understanding of their capabilities and potential, informing future innovations and improvements in artificial intelligence technologies. This section describes the methodology used to conduct the experiments and evaluate the performance of the LLMs in extracting chemical shift data from XPS research papers.

3.1 Experimental Design

The experiments were designed to evaluate the capabilities and limitations of an LLM-powered chatbot in extracting chemical shift data from XPS research papers. The primary focus was on the model's ability to comprehend intricate scientific discussions, respond to specific informational requests, and its proficiency in processing and presenting data extracted from academic articles. To this end the LLM was engaged in a series of conversations, each initiated by providing it with one or more research papers in PDF format, and tasked with extracting specific pieces of information from the given documents.

3.1.1 Selection of Research Papers

The articles used in the experiments were selected, with the help of domain experts, from the NIST Database for the Simulation of Electron Spectra for Surface Analysis (SESSA) [32], containing the chemical shifts for various samples as reported in the literature. The selection focused on papers that presented binding energies related to the electrons in the $2p$ orbital of Silicon atoms, as this element is commonly analyzed using XPS and has well-documented chemical shifts. The chosen papers covered a range of experimental conditions, sample compositions, and chemical states, to ensure a diverse and representative dataset for the experiments.

3.1.2 Data Extraction

The study was conducted using ChatGPT powered by the GPT-4 model, which allows users to upload files for the chatbot to analyze. Focusing on the chemical shifts of Silicon $2p$ electrons, the model was tasked with identifying the relevant data, including the binding energies of the electrons in different chemical states of Silicon, and presenting this information in a structured and coherent manner.

These conversations were conducted both with individual papers and multiple papers simultaneously, to test the extent to which the model's contextual understanding could handle and integrate information from various sources. The extracted data was then compared with the actual values reported in the research papers, as well as in the SESSA database, to evaluate the model's performance.

3.2 Challenges and Adjustments

Throughout the course of this project, two primary challenges, involving ChatGPT's ability to parse PDF files and the extraction and interpretation of chemical shifts from articles, were encountered. Identifying and addressing these challenges enabled the project to progress effectively while providing valuable insights into the capabilities and limitations of LLMs in scientific data analysis. The following sections detail these challenges and the adjustments made to address them.

3.2.1 PDF Parsing

Challenge:

- **Inability to Extract Data:** Initial attempts to upload PDFs directly to ChatGPT led to inconsistent results. PDF files often contain complex formatting, images, and non-standard text encodings. When attempting to parse some of these documents directly, the chatbot would either hallucinate, providing incorrect results, or fail to extract any data, stating it wasn't able to find the requested information. Figure 3.1 shows such an example: after being queried about the chemical shifts in the document, in its response ChatGPT states it could not find the data in the provided article, despite the information being reported in the paper.

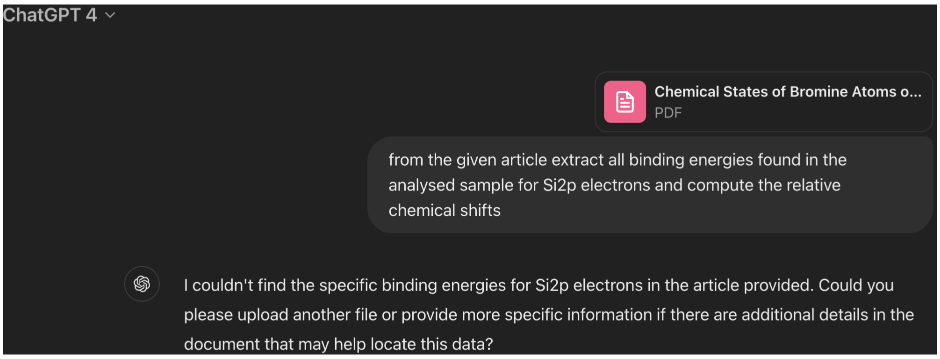


Figure 3.1: Example of ChatGPT failing to extract data from a document.

Adjustment:

- **Conversion Using OCR:** To address this challenge, Optical Character Recognition (OCR) technology was employed. The process involved converting the original PDFs into new, OCR-processed PDF documents. These new PDFs contained the same visual content as the originals, but with an added layer of machine-readable text. This conversion transformed the content of the PDF files, including table data, into selectable and searchable text within the new PDF, while maintaining the original layout and appearance. As a result, the text became accessible and readable by ChatGPT, enabling it to analyze and extract the necessary information effectively from the articles. This adjustment significantly improved the model's ability to parse and interpret the data from the research papers.

Future Improvements:

- **Errors in OCR Conversion:** While OCR allowed text extraction from all of the selected documents, such techniques might introduce mistakes into the parsed text, such as misrecognized characters or formatting errors. Future work could focus on implementing additional verification and correction steps to minimize such mistakes, enhancing the accuracy of the data extracted from the documents.

- **Data Preprocessing:** Not all of the information contained in the articles was relevant to the task at hand. Further preprocessing of the text data could help filter out irrelevant information and focus the chatbot's attention on the key details needed for the extraction of important data.

3.2.2 Interpreting Chemical Shifts from Articles

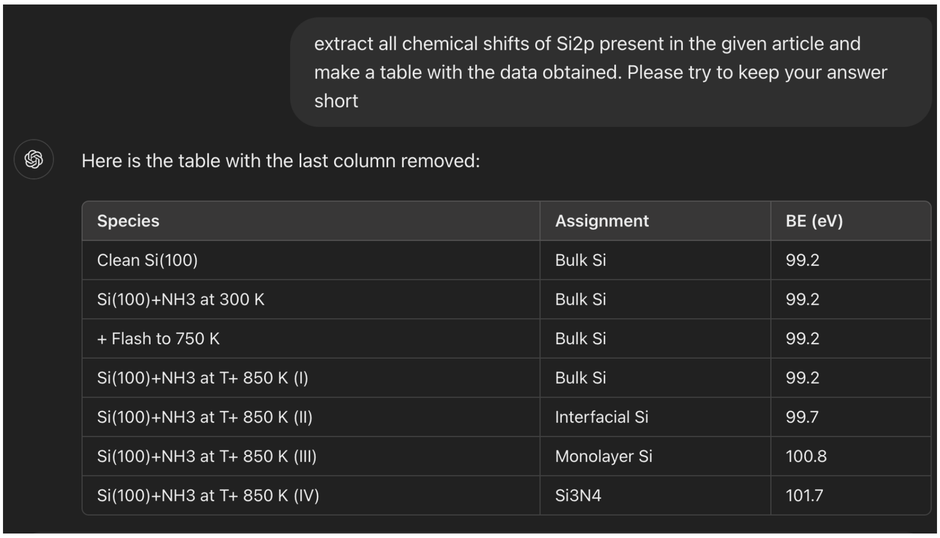
Challenge:

- **Extracting Relevant Chemical Data:** The initial prompts asked ChatGPT to directly extract the chemical shifts from the documents. However, several articles did not directly mention the observed shifts, instead, they provided only the binding energy for the electrons in the sample, along with the binding energy found for the element in the reference state. This posed a challenge as the chatbot was unable to autonomously infer that it could compute the chemical shifts from the provided binding energies. Furthermore, some articles did not report the binding energy of the reference state, making it impossible for the chatbot to compute the chemical shifts. As a result, it would either note that the chemical shifts were not present or report only the binding energies extracted from the articles, as shown in Figure 3.2.

Adjustment:

- **Optimized Prompts for Computation:** The adopted approach involved adjusting the prompts to suggest computing the chemical shifts from the extracted binding energies. Implementing this adjustment required manual intervention to guide the chatbot through the calculations, prompting it to first extract the binding energies and then compute the corresponding chemical shifts.

  The revised prompts were structured as follows:


extract all chemical shifts of Si2p present in the given article and make a table with the data obtained. Please try to keep your answer short

Here is the table with the last column removed:

Species	Assignment	BE (eV)
Clean Si(100)	Bulk Si	99.2
Si(100)+NH3 at 300 K	Bulk Si	99.2
+ Flash to 750 K	Bulk Si	99.2
Si(100)+NH3 at T+ 850 K (I)	Bulk Si	99.2
Si(100)+NH3 at T+ 850 K (II)	Interfacial Si	99.7
Si(100)+NH3 at T+ 850 K (III)	Monolayer Si	100.8
Si(100)+NH3 at T+ 850 K (IV)	Si3N4	101.7

Figure 3.2: ChatGPT extracting binding energies instead of chemical shifts

> *"from the given article extract all of the reported Binding Energies for Si_{2p} electrons, then from the extracted values compute the corresponding chemical shifts and compile a table containing the chemical formula of the sample and the corresponding chemical shift."*

Additionally, in cases where the reference binding energy was omitted by the article, the prompts were adjusted to explicitly specify its value, which had to be manually retrieved from the SESSA database.

Interestingly, it was unnecessary to explicitly provide the formulae for these calculations, as the chatbot was able to infer the required operations, supposedly from its training data. With the adjusted prompts, ChatGPT was able to accurately compute chemical shifts from the given binding energies. This aligns with previous research demonstrating that well-constructed prompts improve LLM performance on various tasks [33].

Future Improvements:

- **Refining Prompts:** While the revised prompts were effective in guiding ChatGPT to compute the chemical shifts, further refinement could enhance the model's ability to autonomously identify and interpret chemical data from the articles. By incorporating more detailed instructions and contextual hints, the chatbot could be guided more effectively in recognizing relevant data.

By addressing these challenges and implementing the adjustments described above, the experiments were able to proceed effectively, providing insights into the capabilities and limitations of LLMs in extracting chemical shift data from XPS research papers. The adjustments made during the course of the project have not only improved the performance of the chatbot in processing and interpreting scientific data but have also highlighted areas for further refinement and optimization in future studies.

Chapter 4

Results and Conclusions

The experiments conducted in this study yielded valuable insights into the capabilities and limitations of state-of-the-art Large Language Models in understanding and extracting chemical shift data from research papers in the field of X-ray Photoelectron Spectroscopy. By employing the GPT-4 model, accessed through the ChatGPT interface, the study assessed the system's performance in identifying relevant information, computing chemical shifts, and presenting the extracted data in a coherent manner.

The results demonstrated the LLM's proficiency in comprehending complex scientific discussions and accurately extracting binding energies reported in the research papers. Furthermore, with appropriate prompting, the model exhibited the ability to compute chemical shifts from the extracted binding energies, leveraging its extensive knowledge to perform the necessary calculations autonomously.

However, the experiments also revealed certain limitations in the LLM's ability to independently identify and interpret chemical shift data from the articles: on several occasions, the model failed to recognize the relevance of the provided binding energies for computing chemical shifts, and, while optimized prompts helped in some cases, there were instances where the LLM was unable to perform the desired computations even with explicit guidance.

4.1 Performance with Individual Papers

To assess the LLM's ability to comprehend and extract chemical shift information from XPS research papers, a total of 30 articles were selected for analysis. These papers covered a diverse range of sample compositions, experimental conditions, and chemical states of Silicon, providing a challenging test for the model's understanding and data extraction capabilities. The performance of the LLM in accurately identifying and computing the chemical shifts of Silicon $2p$ electrons was evaluated by comparing the extracted data with the reference values reported in the respective papers or available in the SESSA database.

In the majority of cases, namely 25 out of 30 articles, the LLM demonstrated an impressive ability to understand the content of the articles, accurately extracting the chemical shift values for all the reported samples or chemical states, with a high degree of accuracy. However, in some instances the chatbot struggled to retrieve all of the binding energies and required further guidance to correctly extract the data, this issue was prevalent in those articles where the data was not presented in a structured manner or was reported in different tables across the paper.

4.2 Performance with Multiple Papers

While the previous section evaluated the LLM's ability to extract chemical shift data from individual research papers, a significant challenge lies in integrating and reconciling information from multiple sources. In the context of XPS data analysis, researchers often need to consult and synthesize findings from various studies to gain a comprehensive understanding of the chemical shifts associated with different materials and experimental conditions. To assess the LLM's proficiency in this regard, experiments were conducted by providing the model with two or more research papers simultaneously and tasking it with extracting and consolidating the chemical shift data reported across these sources.

The articles were selected among those that had been successfully processed singularly and grouped based on the elements found in the studied samples. The tests carried out with two articles showed a decline in the model's performance, only 11 out of 30 pairs of articles resulted in the chatbot successfully extracting the requested data from both files, with most chats extracting data from only one of the papers or using only one reference to compute the shifts, especially when one of the articles contained data from multiple samples.

Further experiments with three articles revealed a similar trend, with the chatbot often failing to integrate the data from all sources and sometimes hallucinating values that were not present in the documents. These results can likely be attributed to the limited context size of the model, which may hinder its ability to retain and utilize information from multiple sources effectively, as well as the complexity of the task, which requires a higher level of reasoning and integration of data across different papers.

4.3 Examples

To illustrate the LLM's performance in analyzing research papers, some examples of interactions with the model are presented below. These examples showcase the chatbot's ability to extract chemical shift data from the articles, as well as the challenges encountered during the experiments.

Successful Extraction

In this example, ChatGPT was able to accurately extract the requested data using the standard prompt presented in the previous chapter:

Article: Thermally grown Si_3N_4 thin films on $Si(100)$: Surface and interfacial composition

ChatGPT: *This table summarizes the different chemical environments of silicon in the samples discussed in the article, showing how the binding energy shifts relative to pure silicon (reference Si_{2p} energy of 99.2 eV). The* *chemical shift is calculated as the difference between the binding energy of the species and the reference energy for elemental silicon.*

Chemical Formula	**Binding Energy (eV)**	**Reference Si_{2p} Energy (eV)**	**Chemical Shift (eV)**
Si	99.2	99.2	0.0
Si (outer layer)	98.7	99.2	-0.5
$Si + NH_3$ (300K)	99.7	99.2	+0.5
$Si + N(Si)_2$	100.4	99.2	+1.2
Interfacial Si	100.8	99.2	+1.6
Si_3N_4	101.7	99.2	+2.5

Partial Extraction

During this interaction, the chatbot was able to extract all of the binding energies reported in the article, but failed to correctly compute the chemical shifts without additional guidance:

Article: An X-ray photoemission spectroscopy investigation of oxides grown on Au_xSi_{1-x} layers

Prompt: *from the given article Extract all of the reported Binding Energies for Si_{2p} electrons, then from the extracted values compute the corresponding chemical shifts and compile a table containing the chemical formula of the sample and the corresponding chemical shift.*

ChatGPT: *The chemical shift is calculated as the difference between the oxidized and nonoxidized Si_{2p} binding energies. For the $Au_{0.81}Si_{0.19}$ sample, only the oxidized Si_{2p} value is reported, and thus no chemical shift can be computed. Here is the compiled table showing the chemical formula of each sample and the corresponding chemical shift:*

Sample Composition	**Oxidized** Si_{2p} **(eV)**	**Nonoxidized** Si_{2p} **(eV)**	**Chemical Shift (eV)**
$Au_{0.81}Si_{0.19}$	103.3	-	-
$Au_{0.59}Si_{0.41}$	103.1	98.9	4.2
$Au_{0.45}Si_{0.55}$	103.4	99.1	4.3
$Au_{0.34}Si_{0.66}$	103.2	98.8	4.4
$Si(100)HT$	103.5	99.1	4.4

The initial calculation of chemical shifts was flawed: the model erroneously subtracted non-oxidized binding energies from oxidized ones, rather than using the reference $Si(100)$ energy, which wasn't extracted. To address this issue, the chatbot was given two prompts: one to extract the binding energies and another to compute the chemical shifts using the reference $Si(100)$ energy. This two-step approach allowed the model to accurately determine the chemical shifts:

Prompt: *from the given article, extract all of the reported Binding Energies for Si2p electrons and compile them in a table, make sure to include all binding energies even if different ones are given for the same sample*

ChatGPT: *Based on the document, I've compiled a table listing all the reported binding energies for* Si_{2p}*electrons from the XPS analysis of oxidized AuxSi1-x layers. Here's the table:*

Sample Composition	**Nonoxidized** Si_{2p} **(eV)**	**Oxidized** Si_{2p} **(eV)**
$Au_{0.81}Si_{0.19}$	-	103.3
$Au_{0.59}Si_{0.41}$	98.9	103.1
$Au_{0.45}Si_{0.55}$	99.1	103.4
$Au_{0.34}Si_{0.66}$	98.8	103.2
$Si(100)$	99.5	103.6
$Si(100)HT$	99.1	103.5

Prompt: *now compute the chemical shifts for every sample using as* *reference both of the binding energies given for $Si(100)$ and create a table out of the results.*

ChatGPT: *To compute the chemical shifts, we will use the binding energies of both nonoxidized and oxidized Si_{2p} electrons from the "$Si(100)$" sample as reference values. The reference values for nonoxidized and oxidized Si_{2p} from the $Si(100)$ sample are 99.5 eV and 103.6 eV, respectively.*

Sample Composition	**Nonoxidized Si_{2p} Shift (eV)**	**Oxidized Si_{2p} Shift (eV)**
$Au_{0.81}Si_{0.19}$	-	-0.3
$Au_{0.59}Si_{0.41}$	0.6	0.5
$Au_{0.45}Si_{0.55}$	0.4	0.2
$Au_{0.34}Si_{0.66}$	0.7	0.4
$Si(100)$	0	0
$Si(100)HT$	0.4	0.1

Failed Extraction

In this example, the model was provided with two articles to analyze, but proved unable to extract data from both sources, only managing to process the information from one of the papers and failing to integrate the data from the second article:

Articles:

- Fabrication of minority-carrier-limited $n - Si$/insulator/metal diodes
- Nondestructive Depth Profile Analysis by Changing Escape Depth of Photoelectrons

Prompt: *from the given articles find and extract all binding energies given for Si_2p electrons and compile them in a table alongside the chemical formula of the analyzed sample. Make sure to find every binding energy in both articles.*

ChatGPT: *Here are the binding energies for Si_{2p} electrons compiled from the documents you provided, alongside the chemical formula of the analyzed samples:*

Chemical Formula (Sample)	**Binding Energy (eV)**	**Source**
Si/SiO_2	103.3 eV (SiO2), 99.1 eV (Si substrate)	Nondestructive Depth Profile Analysis by Changing Escape Depth of Photoelectrons

4.4 Conclusions

The research presented in this thesis has investigated the integration of Large Language Models into the analysis of X-ray Photoelectron Spectroscopy data. The findings confirm that LLMs possess impressive capabilities in the analysis of research articles within the field of XPS, demonstrating their proficiency in comprehending complex scientific discussions, accurately extracting binding energies reported in the research papers, and computing chemical shifts from the extracted data when provided with appropriate prompting. However, our results also highlighted the current limitations of LLMs, primarily their dependency on the quality and extent of the training data provided.

Without sufficient domain-specific training, the models may produce less accurate or relevant interpretations, which could lead to misleading conclusions. To address these challenges and enhance the accuracy of results, it is essential to adopt a more refined approach: other works have seen success in improving the performance of LLMs by integrating them with external, expert-designed tools and pre-existing specialized chatbots [34, 35]; nonetheless the involvement of human experts still remains crucial to ensure the accuracy and reliability of the model's outputs. Moreover, the development of more robust evaluation metrics and benchmarks is imperative for assessing the true capabilities and limitations of LLMs in specific domains.

Significant efforts have been dedicated to constructing comprehensive datasets to evaluate the performance of models on various tasks in fields such as mathematics [36, 37] and logical reasoning [38, 39]. Similarly, in the realm of XPS data analysis, the creation of standardized datasets and evaluation protocols could play a crucial role in objectively measuring the progress of LLMs facilitating the comparison and ranking of different models while also providing valuable insights into the strengths and weaknesses of the approaches employed, guiding future research and development efforts. In conclusion, the limitations and challenges encountered in this study underscore the importance of a cautious and informed approach to the integration of LLMs into scientific research and data analysis.

While these models hold great promise in enhancing the efficiency and accuracy of data processing and interpretation, their deployment must be accompanied by rigorous evaluation and validation to ensure the reliability and validity of the results obtained. By addressing these challenges, LLMs can be improved and optimized for specific domains, contributing to the advancement of scientific research and the development of innovative tools and technologies.

Bibliography

[1] A. S. A. Rasheed, M. M. Masud, and M. Abduljabbar, "A comparative study of transformer based pretrained AI models for content summarization," in *15th International Conference on Innovations in Information Technology, IIT 2023, Al Ain, United Arab Emirates, November 14-15, 2023*. IEEE, 2023, pp. 79–84. [Online]. Available: https://doi.org/10.1109/IIT59782.2023.10366411

[2] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi, "Can chatgpt replace traditional KBQA models? an in-depth analysis of the question answering performance of the GPT LLM family," in *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I*, ser. Lecture Notes in Computer Science, T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, and J. Li, Eds., vol. 14265. Springer, 2023, pp. 348–367. [Online]. Available: https://doi.org/10.1007/978-3-031-47240-4_19

[3] S. C. Siu, "Chatgpt and gpt-4 for professional translators: Exploring the potential of large language models in translation," *Available at SSRN 4448091*, 2023.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[6] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, vol. 2022, 2022. [Online]. Available: https://openreview.net/forum?id=yzkSU5zdwD

[7] J. Chen and J. Zacharias, "Design principles for collaborative generative AI systems in software development," in *Design Science Research for a Resilient Future - 19th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2024, Trollhättan, Sweden, June 3-5, 2024, Proceedings*, ser. Lecture Notes in Computer Science, M. Mandviwalla, M. Söllner, and T. Tuunanen, Eds., vol. 14621. Springer, 2024, pp. 341–354. [Online]. Available: https://doi.org/10.1007/978-3-031-61175-9_23

[8] A. Ziegler, E. Kalliamvakou, X. A. Li, A. Rice, D. Rifkin, S. Simister, G. Sittampalam, and E. Aftandilian, "Productivity assessment of neural code completion," in *MAPS@PLDI 2022: 6th ACM SIGPLAN International Symposium on Machine Programming, San Diego, CA,*

USA, 13 June 2022, S. Chaudhuri and C. Sutton, Eds. ACM, 2022, pp. 21–29. [Online]. Available: https://doi.org/10.1145/3520312.3534864

[9] A. D. White, G. M. Hocky, H. A. Gandhi, M. Ansari, S. Cox, G. P. Wellawatte, S. Sasmal, Z. Yang, K. Liu, Y. Singh, and W. J. Peña Ccoa, "Assessment of chemistry knowledge in large language models that generate code," *Digital Discovery*, vol. 2, pp. 368–376, 2023. [Online]. Available: http://dx.doi.org/10.1039/D2DD00087C

[10] M. P. Polak and D. Morgan, "Extracting accurate materials data from research papers with conversational language models and prompt engineering," *Nature Communications*, vol. 15, no. 1, p. 1569, 2024.

[11] Y. Mathur, S. Rangreji, R. Kapoor, M. Palavalli, A. Bertsch, and M. R. Gormley, "Summqa at mediqa-chat 2023: In-context learning with GPT-4 for medical summarization," in *Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023*, T. Naumann, A. B. Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds. Association for Computational Linguistics, 2023, pp. 490–502. [Online]. Available: https://doi.org/10.18653/v1/2023.clinicalnlp-1.51

[12] Q. Lyu, J. Tan, M. E. Zapadka, J. Ponnatapura, C. Niu, K. J. Myers, G. Wang, and C. T. Whitlow, "Translating radiology reports into plain language using chatgpt and GPT-4 with prompt learning: results, limitations, and potential," *Vis. Comput. Ind. Biomed. Art*, vol. 6, no. 1, p. 9, 2023. [Online]. Available: https://doi.org/10.1186/s42492-023-00136-5

[13] X. Li, S. Chan, X. Zhu, Y. Pei, Z. Ma, X. Liu, and S. Shah, "Are chatgpt and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December*

6-10, 2023, M. Wang and I. Zitouni, Eds. Association for Computational Linguistics, 2023, pp. 408–422. [Online]. Available: https://doi.org/10.18653/v1/2023.emnlp-industry.39

[14] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 248:1–248:38, 2023. [Online]. Available: https://doi.org/10.1145/3571730

[15] A. T. Kalai and S. S. Vempala, "Calibrated language models must hallucinate," in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, B. Mohar, I. Shinkar, and R. O'Donnell, Eds. ACM, 2024, pp. 160–171. [Online]. Available: https://doi.org/10.1145/3618260.3649777

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., 2013, pp. 3111–3119. [Online]. Available: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1532–1543. [Online]. Available: https://doi.org/10.3115/v1/d14-1162

[19] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 2227–2237. [Online]. Available: https://doi.org/10.18653/v1/n18-1202

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[22] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *CoRR*, vol. abs/2304.13712, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.13712

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[24] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.703

[27] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[28] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[29] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang, "Evaluating the logical reasoning ability of chatgpt and gpt-4," *ArXiv*, vol. abs/2304.03439, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258041354

[30] F. A. Stevie and C. L. Donley, "Introduction to x-ray photoelectron spectroscopy," *Journal of Vacuum Science & Technology A*, vol. 38, no. 6, p. 063204, 09 2020. [Online]. Available: https://doi.org/10.1116/6.0000412

[31] S. Hagström, C. Nordling, and K. Siegbahn, "Electron spectroscopic determination of the chemical valence state," *Zeitschrift für Physik*, vol. 178, no. 5, pp. 439–444, 1964. [Online]. Available: https://doi.org/10.1007/BF01379473

[32] C. Powell, "Nist database for the simulation of electron spectra for surface analysis (sessa), version 2.1.1," 2018-09-18 2018.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[34] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “Augmenting large language models with chemistry tools,” *Nature Machine Intelligence*, pp. 1–11, 2024.

[35] G. Gamage, S. Kahawala, N. Mills, D. D. Silva, M. Manic, D. Alahakoon, and A. Jennings, “Augmenting industrial chatbots in energy systems using chatgpt generative AI,” in *32nd IEEE International Symposium on Industrial Electronics, ISIE 2023, Helsinki, Finland, June 19-21, 2023*. IEEE, 2023, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ISIE51358.2023.10228101

[36] S. Frieder, L. Pinchetti, A. Chevalier, R. Griffiths, T. Salvatori, T. Lukasiewicz, P. Petersen, and J. Berner, “Mathematical capabilities of chatgpt,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html

[37] K. Yang, A. Swope, A. Gu, R. Chalamala, P. Song, S. Yu, S. Godil, R. J. Prenger, and A. Anandkumar, “Leandojo: Theorem proving with retrieval-augmented language models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[38] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, “Reasoning with language model prompting: A survey,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 5368–5393. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.294

[39] B. Wang, X. Yue, and H. Sun, "Can chatgpt defend its belief in truth? evaluating LLM reasoning via debate," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 11 865–11 881. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-emnlp.795