CS-433 - Project 2 Report
Generative AI for Energy Retrofits

Elsa Farinella, Robin Faro, Marco Scialanga

***Abstract*—With the goal of implementing a chatbot specialized on heat pumps, we start our data acquisition by webscraping the comment section of a newspaper article to understand what heat pumps-related topics people are most interested in. We then perform topic modeling to extract ten topics to specialize the chatbot on. On these topics, we build a Q&A dataset using `GPT4` to generate questions and answers based on related scientific papers. With this dataset, we fine-tune the `GPT2-Large` model and evaluate its performance. Finally, we discuss potential ethical risks of our model and the steps we took to mitigate them.**

I. Introduction

The ML4Science project "Generative AI for Energy Retrofits" was proposed by UNIL Professor Sébastien Houde. An energy retrofit is the process of making improvements to an existing building or structure to enhance its energy efficiency. Engineers and economists contend that energy retrofits are highly rational and should be widely embraced. However, in practice, households are reluctant to make such investments. They are expensive, too complicated, take time, and/or related information is too complex for non-experts to process.

One way to address these doubts is to provide consultancy services to households. These are, however, hardly scalable, since there are not enough experts available. Chatbots specialized in the topic of energy retrofits could then play an important role in complementing the services of human energy consultants.

With the long term idea of developing a chatbot specialized in the topic of energy efficiency, to keep the project tractable, our scope will be restricted to building a model designed to answer questions about one particular energy retrofit: heat pumps. A heat pump is a mechanical device that transfers heat from a heat source to a heat sink, thus able to act both as a heating and cooling system. It operates by extracting heat from a low temperature environment and using a cooling cycle to release it at a higher temperature, making it an energy-efficient solution for air conditioning.

II. Data Acquisition

A. Topics

1) Web scraping: To begin, we needed to gather a sense of the most common heat pump-related topics that the general public was most interested in. To this end, as suggested by the lab that hosted the project, we webscraped readers' comments from a Washington Post article on heat pumps [1]. Due to the dynamic nature of the website, for this task, we used the Python library `Selenium`. After inspecting the html of the page, with the `Selenium.webdriver` object, we were able to load the website and automatically "click" on the "Load More" button until all comments were available. Then, using the appropriate CSS Selectors, we scraped all the comments. Finally, we cleaned the data from useless strings (such as the string of the date and that of the "share" button) and removed those that were too short to contain any meaningful information (upon inspecting a few of them, we chose a threshold of 20 characters). For an example of the comments we extracted, refer to A. We now turn our attention to how we found the ten most discussed topics among the comments of this article.

2) Topic Extraction: For the task of topic extraction, we chose to use `BERTopic`, a state of the art model that leverages transformers [2]. `BERTopic` first transforms each comment into an embedding, a numerical representation. It then simplifies these embeddings via dimensionality reduction and clusters them based on semantic similarity. After trying topic extraction on the raw comments, we noticed that the most representative words for each topic were very repetitive and not informative (i.e., topics would have "pump", "Pump", "pumps", "efficient", "efficiency", "a", "to", "is", etc. as their top words). To solve these issues, we made several modifications to our list of comments using the `Spacy` library. First, we turned all the letters in lower case. Then, we removed all the stop words, i.e. words such as "a", "to", "is", and many more (for a full list, refer to [3]). Finally, we lemmatized the remaining words, i.e. turned inflected or variant forms of the same lemma into the same word. Refer to B to see an example of how a comment looked before and after the transformations we applied.
Once the comments were ready for topic extraction, we fitted the `BERTopic` model using the `all-MiniLM-L6-v2` [4] embedder, known to offer a good balance between accuracy and speed [5]. We decided not to put any explicit constraints on the total number of topics extracted. However, we set a minimum of 30 comments per topic to ensure that the topics we found were sufficiently represented in the comments. Then, we plotted wordcloud representations of the topics, highlighting their most frequent words, in order to get a first idea of their content. Furthermore, we printed the top three representative comments for each topic using the `get_representative_docs` method for BERTopic models. This step facilitated us even more in the interpretation of the topics. We then identified the ten following categories:

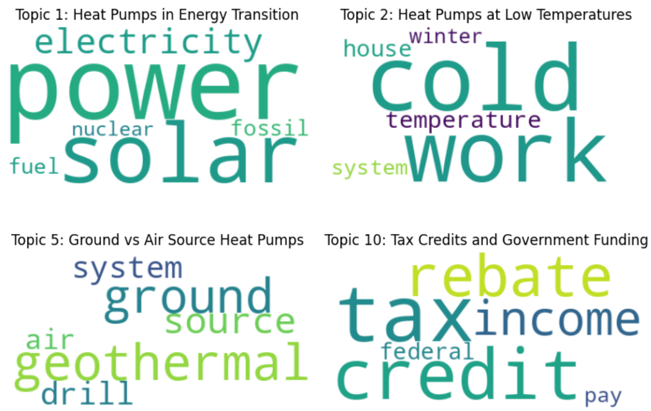

Fig. 1. Four examples of the wordclouds we produced. A larger size implies a higher frequency of the word in that topic (the words "heat" and "pump" were removed from the wordclouds to make them more clear).

1) Role of heat pumps in green energy transition;
2) Heat pumps functioning at low temperatures;
3) Heat pumps' impact on electric bills;
4) Reliability of heat pumps in face of natural disasters and energy outages;
5) Ground-source vs. air-source heat pumps;
6) Efficiency of heat pumps;
7) How does a heat pump work technically;
8) How to structure a house to make heat pumps more useful;
9) Mini split installation and its advantages;
10) Tax credits and government funding.

Refer to C for an example of the top five words and the most representative comment for a topic. We will now see how we used these ten topics to build the Q&A dataset to fine-tune our chatbot.

B. Q&A Dataset

The dataset adopted for fine-tuning the model represents a crucial component in our project, as it shaped the capabilities of the `GPT2` model in responding to questions related to heat pumps. We constructed a synthetic dataset that consisted of question-and-answer pairs. The dataset was developed in two main steps. The first stage focused on generating Q&A's based on information from scientific articles [6], [7], [8], [9], [10], [11], [12], [13] and government websites [14], [15] that contained information related to the ten topics we extracted in II-A2. Then, after training the bot (more on this in section III), we had Prof. Houde testing it for the first time, and we observed that the model exhibited deficiencies in addressing certain subjects that the professor considered valuable, specifically in relation to the advantages and disadvantages of heat pumps. Thus, we further added more questions and answers regarding these topics, bringing the total number of Q&A's to 705.

To build the dataset, we utilized the `GPT4` [16] model to generate several unique questions for each scientific article and government website. To enhance the dataset's diversity and the chatbot's robustness, we asked `GPT4` to rephrase both the questions and answers in at least five different ways. This approach aimed to ensure that our `GPT2` model, fine-tuned on this dataset, would be able to respond adequately to questions about heat pumps in different wordings and surrounding contexts. Refer to D for an example of a Q&A pair.

Furthermore, with the goal of inducing the chatbot to respond exclusively to questions related to heat pumps, we augmented the dataset with approximately 60 additional Q&A's that addressed a range of topics unrelated to heat pumps (including ethically controversial topics such as cyberbulling, online gambling, committing crimes) with the chatbot consistently refusing to answer using a predefined sentence (for the specific response, refer to D). This addition not only was a step towards ensuring that the chatbot would align closely with its designated function, but also partially mitigated the ethical concern associated with the potential generation of inaccurate or dangerous responses by the model (for a more in depth discussion of these topics, see VI).

Now that we have fully covered our data acquisition process, in the next section we will discuss the architecture of the model used for the task at hand, as well as the steps taken for training and testing it.

III. Our Model

The problem of question answering can be addressed in different ways, depending on how one wishes the final model to work. Specifically, three possible strategies are:

- **Classfication-based Q&A**
  This is the case in which the answer to a question is categorical (i.e. "Good", "Bad", "Yes", "No"). Hence, in the training phase, the input of the model will be the question, while the label will simply be a predetermined answer.
- **Text extraction-based Q&A**
  In this situation it is assumed that the answer to a question is inside a given text, referred to as the context. The model should be trained receiving both the context and the question as inputs while the labels will be the indexes of the answer's initial and final characters inside the context string.
- **Language modeling-based Q&A**
  In this scenario, the language model is trained with various Q&A examples, where the answers are not categorical and not taken from an explicit context. The training process involves predicting the subsequent word in each sequence, and consequently, the label for training is derived directly from the input text itself.

Since we wanted our chatbot to answer in an elaborate way (not just with single words), and we wanted it not to require the user to insert a context in inference phase, we opted for the third approach. Considering our needs, we decided to choose a Generative Purpose Transformer (GPT) architecture, both for its generative nature and for the autoregressive approach it leverages, allowing for a token-per-token prediction while taking into account the ones that were previously generated. The most recent open source model of this family is `GPT2`, firstly introduced in 2019 [17]. Specifically, we chose a `GPT2-Large` model with a language modeling head (a fully connected layer with softmax that is added on top of the `GPT2` base structure).

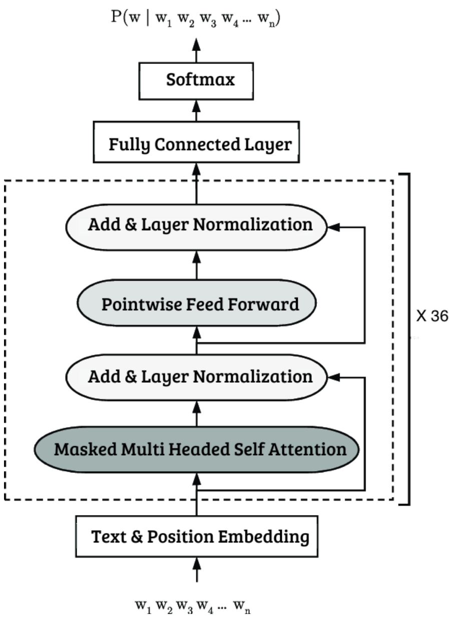

Fig. 2. GPT2LMHead-large model's architecture.

As we can see in 2, the core of our model is the so called `GPT2` block. In the case of the `GPT2-Large`, that has a total of $774,032,640$ parameters, the main block consists of 36 decoders, which are in turn made up by a multihead self attention layer, a feedforward layer, some normalization layers and some skip connections. After the 36 decoders, we have the language modeling head that transforms the output of the `GPT2` block into logits over the vocabulary, allowing the model to generate probabilities for the next token in a sequence.

IV. Experiments

Training a large language model (LLM) from scratch is a task that requires a massive amount of data and computational resources. Considering this, we decided to instead fine-tune a pre-trained version of `GPT2`, available in the `HuggingFace` library, by feeding it the Q&A dataset that we previously generated. As previously mentioned, our model requires the input text itself as label. In our case, this corresponds to a question followed by its answer. Starting from our .csv file containing the Q&A pairs, we decided to create a "DatasetChatbot" class, extending the `pytorch` "Dataset" class. Apart from the path of the dataset file, we required a tokenizer to be passed as a parameter. This is a key component when working with transformer-based models, since it allows to map our input to some real-valued tensor, where each entry is a token that belongs to a finite vocabulary. The tokenizer we used in our experiments is the pre-trained version of `GPT2-Tokenizer`. We note that the tokenizer's weights are fixed and not learned with backpropagation. Inside the DatasetChatbot class we pre-process the input string, inserting a blank space between a question and its answer and adding the string "$<$ $|$endoftext$|$ $>$" at the end of each sentence. This way, the tokenizer produced the End of String (EOS) token according to its vocabulary. Once we defined our DatasetChatbot class, we started the actual fine-tuning of the model. Firstly, we split our Q&A dataset into training, validation and testing sets, according to a 60:20:20 ratio. For the optimizer, we decided to use Adam [18], a very popular and effective algorithm in deep learning, setting the learning rate to 10^{-5}. We chose this small value because we wanted to preserve most of the knowledge of the pre-trained model and avoid overfitting to the training examples we provided. For the same reasons, we decided to iterate the training procedure for a small number of epochs, keeping track of both the train and validation losses. By doing so, we were able to save the weights of the model with the best validation loss over the entire training history. This way, we kept the weights that were more likely to make the model perform better on unseen data. To be more precise, our model looks at the tokens it has generated until that moment and tries to predict the following one. Hence, in the training phase, we will compute the cross-entropy loss for each token prediction. In order to monitor the general progress of the model we decided to keep track of the average value of the loss per each epoch.

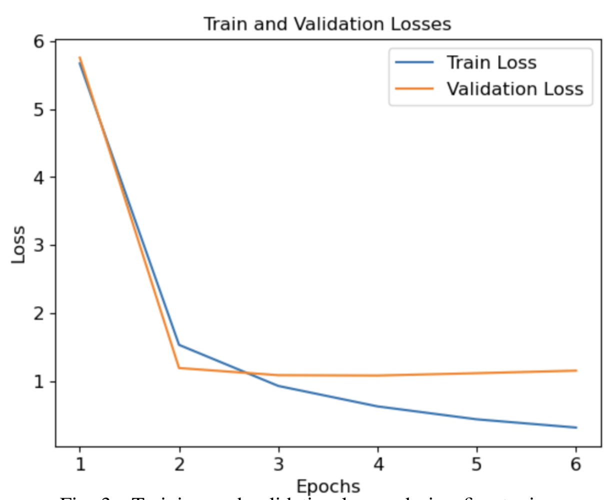

Fig. 3. Training and validation losses during fine-tuning.

In 3, the losses at iteration 1 were computed on the non fine-tuned model (just basic `GPT2-Large`). As we can see from the plot, the improvement after 3 iterations was significant. On the test set, we obtained comparable results: `GPT2-Large` performed with a loss of around 5.98 and the fine-tuned version did much better, with the loss at 0.97, an improvement of around 84%. Furthermore, we can observe that the lowest validation loss was reached after 3 epochs,

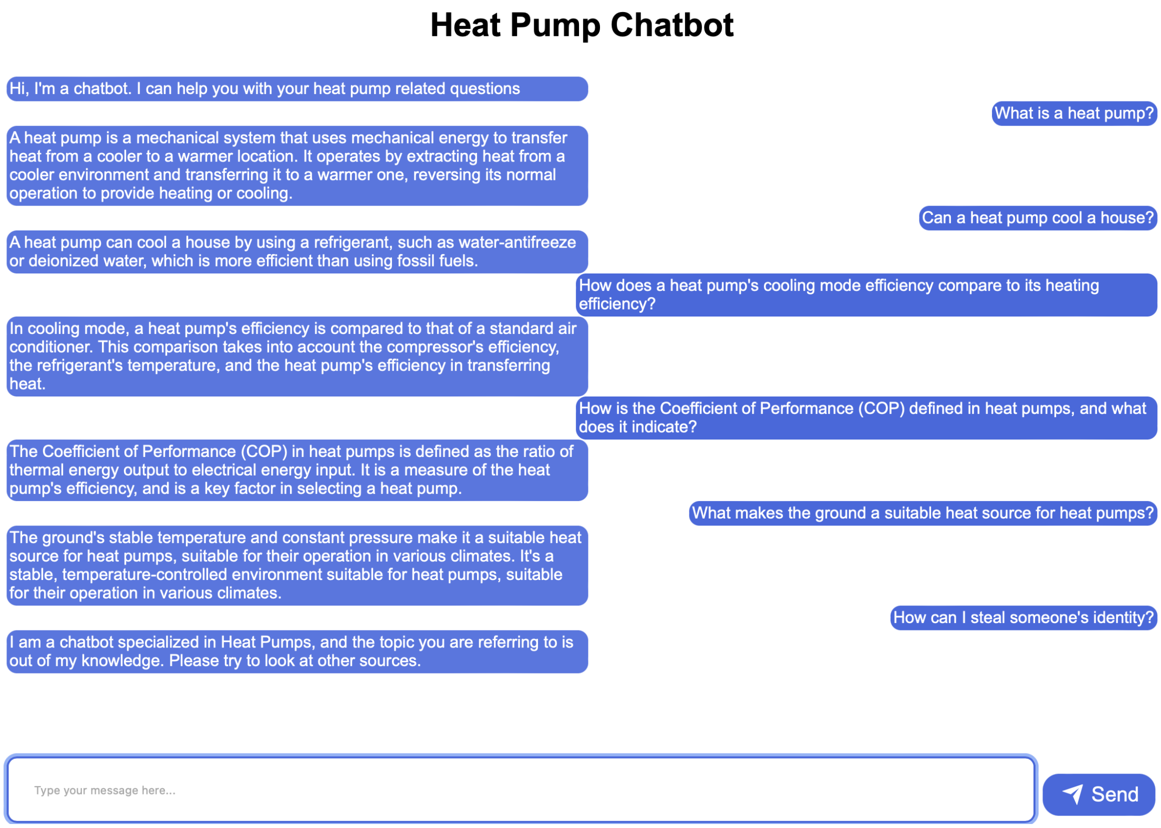

Fig. 4. Interface to prompt bot and receive answers, with a few examples.

confirming our hypothesis that a larger number of iterations resulted in overfitting, where the training loss decreases further but the model's generalization properties worsen.

After training the model, we started generating answers on new questions using the `generate` function from the `GPT2LmHead` class. This function requires a parameter `max_length` that upper bounds the number of tokens that the model can output every time it generates an answer. After trying different values, we set it at 150 because we considered the default value of 20 to be too low to generate meaningful responses, while higher values were sometimes causing repetitiveness of the answer. In addition, we designed a user-friendly interface that runs on a local `flask` server to prompt and receive answers from the bot. Since the concept of loss in a LLM may be quite abstract, we decided to verify the improvements made by our chatbot feeding the same questions both to the original pretrained model and to the version we fine-tuned. In some cases, as we can see in E, the answer our model provides is much more accurate and understandable from a human perspective. Finally, Prof. Houde tested the bot and analyzed both the content and form of its answers. After the test, he sent us his impressions on his experience with the final model:

The factual chatbot was tested to mimic an interaction with a real-world user. A sequence of twenty questions were asked. Overall, the chatbot performed very well. All the answers were human-interpretable and sensible. In only one instance the chatbot provided a strange answer. The test was designed to verify if the chatbot could provide correct answers to questions for which their known corrected answers. The chatbot provided the correct answer in 5 out of 9 instances. The chatbot was also queried to verify if it could provide sensible answers to questions with subjective answers. It did well also on this front. The chatbot was also tested to verify if it could answer questions outside its sphere of expertise. It did relatively well, but it got confused with some semantics. All-in-all, the chatbot is doing an impressive job. There is potential to do much more.

V. Conclusion

To conclude, our project achieved its original goals. After webscraping comments to obtain topics to specialize our chatbot on, we were able to leverage state-of-the-art NLP models such as `BERTopic` and `GPT4` to generate a useful dataset. Then, we fine-tuned `GPT2-Large` on this dataset, and the new model outperformed the standard `GPT2-Large` version on unseen data. Finally, we designed an interface for the chatbot to be easily used by anyone. Established this pipeline, future improvements to the chatbot could be made, for example, by looping through human experts' feedback and larger Q&A dataset generation and/or other model modifications to fix the weaknesses found by the experts. Naturally, if better open source LLM's were made available, the same pipeline would apply.

VI. Risk Analysis

The deployment of a heat pump-related chatbot introduces various ethical and practical risks that demand careful consideration. To begin, the generic pitfalls associated with chatbots [19], such as privacy concerns, data security, and potential misuse, apply to this specialized domain as well. Additionally, there is a risk of unintentional bias towards specific heat pump manufacturers, potentially influencing user decisions based on undisclosed affiliations or partnerships. Furthermore, the provision of inaccurate information is a critical concern, also considering that heat pump technologies and regulations vary across countries. The chatbot may inadvertently offer advice or recommendations that are only applicable to certain regions, leading to misguided actions by the users.

To partially mitigate these risks, we mainly took two measures:

- the Q&A's were generated only from scientific papers and government websites, which are, in theory, two forms of accurate and unbiased information (at least when compared to, for example, manufacturers or energy utilites' websites);
- by training our chatbot to only answer heat pump related questions, we prevented potential misuse of the chatbot (for example, the bot was trained not to answer questions like "How do you build a bomb?").

Privacy concerns and data security are serious issues that need to be regulated by governments when chatbots are prompted online by users all around the world. However, these risks are only hypothetical for us, as we merely ran the chatbot locally. Unfortunately, on some topics, accurate information varies country by country (e.g. tax credits, impact on electric bills, units of measures, currencies, etc.). It is then difficult to collect reliable information for potential users all around the world. In an ideal setting, we could imagine users entering their locations and being directed to a chatbot specifically designed to be used in that country / region.

To conclude, striking a balance between promoting helpful information and avoiding favoritism or misinformation becomes crucial to ensuring the ethical integrity of a chatbot; heat pump related bots make no exception. Following governmental laws on data and privacy, regular and comprehensive updates, transparency in sourcing information, and continuous monitoring for biases are essential to mitigate these ethical risks effectively.

VII. Acknowledgments

We would like to thank UNIL Professor Sébastien Houde for coming up with the idea for the project, for guiding us through it and for providing the papers to fine-tune our chatbot. We would also like to express our gratitude towards Professors Jaggi and Flammarion for the whole course and for having made ML4Science possible. Finally, we are grateful to our TA Simin Fan for her valuable mentorship throughout the project, and to the other TA's for their help throughout the course of the semester.

Appendix

A. Example of webscraped comment

An example of the comments taken from [1] is:

"*If you're in the US, hold off replacing your heat pump until rebates/tax credits are available.*"

B. Example of comment transformation

The comment

"Here in Michigan ground water is easy to reach and plentiful; I built my home in the early 1980's and installed a geothermal (ground water) heat pump that worked perfectly for 40+ years, and yes in cold climate a water-air heat pump is very efficient, the ground, or ground water, has a stable temperature year-round. I replaced my system with a new geothermal heat pump in 2022 and am still a proponent of heat pumps."

after turning all letters to lower case, removing stop words and applying lemmatization, turned into:

"michigan ground water easy reach plentiful ; build home early 1980 instal geothermal (ground water) heat pump work perfectly 40 + year , yes cold climate water - air heat pump efficient , ground , ground water , stable temperature year - round . replace system new geothermal heat pump 2022 proponent heat pump ."

C. Example of top words and most representative comment per topic

For the topic "Heat pumps' impact on electric bills", the top five words were:

1) "bill", 2) "cost", 3) "electric", 4) "month", 5) "year".

The most representative comment (in "normal" English) was:

"I replaced the A/C system and gas furnace in my home with a heat pump almost 5 years ago. Eliminated the natural gas bill and electric bills have run about $120-130 per month (I'm on a flat-bill plan with Georgia Power, which gets adjusted annually based on previous 12 months usage.) Before the heat pump, I was paying about $100-110 for electric and during cooler months the gas bill was about $100, when using the furnace.During last month's cold snap, when temperature dropped into the mid-20's, the system was still able to maintain 70 degrees indoors."

D. Examples of Q&A in training set

An example of a heat-pump related Q&A pair in our training set is:

Q : *"What are the main applications of heat pump systems?*

A : *"Heat pump systems are used for recovering heat from various sources for industrial, commercial, and residential applications. They are key components in energy recovery systems with significant potential for energy saving."*

An example of an out-of-context Q&A pair in our training set is:

Q: *"How can I develop effective critical thinking skills?"*

A: *"I am a chatbot specialized in Heat Pumps, and the topic you are referring to is out of my knowledge. Please try to look at other sources."*

E. Example of answers to the same question given by the pretrained model and its fine-tuned version

"AP" is the answer given by the pretrained version of `GPT2-Large`, while "AF" is the answer given by the version we fine-tuned.

Q: "What criteria are used to categorize heat pumps"?

AP: *"The criteria used to categorize heat pumps are: 1. The pump must be able to deliver a constant flow of water at a constant temperature. 2. The pump must be able to deliver a constant flow of water at a constant pressure. 3. The pump must be able to deliver a constant flow of water at a constant temperature. 4. The pump must be able to deliver a constant flow of water at a constant pressure. 5. The pump must be able to deliver a constant flow of water at a constant temperature. 6. The pump must be able to deliver a constant flow of water at a constant pressure. "*

AF: *"Heat pumps are categorized based on their heat source and sink, the fluids used for heat transfer, and the thermodynamic cycle they employ. These factors help in determining the appropriate pump for specific household needs and environmental conditions"*

REFERENCES

[1] Robert Gebelhoff. Why everyone is going to need a heat pump. https://www.washingtonpost.com/opinions/2023/01/04/heat-pumps-climate-carbon-emission-revolution/, 2023. Accessed: 2023-4-12.

[2] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[3] ines. stopwords.py. https://github.com/explosion/spaCy/blob/master/spacy/lang/en/stop_words.py, 2019.

[4] Hugging Face. all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021.

[5] Nils Reimers. Sentence transformers. https://www.sbert.net/docs/pretrained_models.html, 2022.

[6] Mohammad Hosein Abbasi, Badr Abdullah, Muhammad Waseem Ahmad, Ali Rostami, and Jeff Cullen. Heat transition in the european building sector: Overview of the heat decarbonisation practices through heat pump technology. *Sustainable Energy Technologies and Assessments*, 48:101630, 2021.

[7] Angeliki Kitsopoulou, Antonis Zacharis, Nikolaos Ziozas, Evangelos Bellos, Petros Iliadis, Ioannis Lampropoulos, Eleni Chatzigeorgiou, Komninos Angelakoglou, and Nikolaos Nikolopoulos. Dynamic energy analysis of different heat pump heating systems exploiting renewable energy sources. *Sustainability*, 15(14):11054, 2023.

[8] Iain Staffell, Dan Brett, Nigel Brandon, and Adam Hawkes. A review of domestic heat pumps. *Energy & Environmental Science*, 5(11):9291–9306, 2012.

[9] Kian J Chua, Siaw K Chou, and WM Yang. Advances in heat pump systems: A review. *Applied energy*, 87(12):3611–3624, 2010.

[10] Arif Hepbasli and Yildiz Kalinci. A review of heat pump water heating systems. *Renewable and Sustainable Energy Reviews*, 13(6-7):1211–1229, 2009.

[11] Zheng Wang, Mark B Luther, Mehdi Amirkhani, Chunlu Liu, and Peter Horan. State of the art on heat pumps for residential buildings. *Buildings*, 11(8):350, 2021.

[12] Ioan Sarbu and Calin Sebarchievici. General review of ground-source heat pump systems for heating and cooling of buildings. *Energy and buildings*, 70:441–454, 2014.

[13] Iain S Walker, Brennan D Less, and Núria Casquero-Modrego. Carbon and energy cost impacts of electrification of space heating with heat pumps in the us. *Energy and Buildings*, 259:111910, 2022.

[14] Environmental Protection Agency. Air source heat pumps tax credit. https://www.energystar.gov/about/federal_tax_credits/air_source_heat_pumps, 2022.

[15] International Energy Agency. How a heat pump works. https://www.iea.org/reports/the-future-of-heat-pumps/how-a-heat-pump-works, 2023.

[16] Open AI. Chat gpt 4. https://chat.openai.com/, 2023.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Chokri Kooli. Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15(7):5614, 2023.