Enabling Autonomous and Adaptive Social Robots in Education: A Vision for the Application of Generative AI

Eva Verhelst, Ruben Janssens, Tony Belpaeme

IDLab-AIRO, Ghent University – imec

Abstract The limited autonomy of social robots currently prevents many ambitions in educational robotics from being realised. This leads to scripted dialogues, content that fails to adapt to individual students and conversations remaining largely text-based. Recent advances in generative artificial intelligence (AI) might alleviate these issues, allowing for educational robots whose dialog can be flexibly generated based on the lessons to be taught, the student's needs and personality, and the environment. This chapter presents a vision of how generative AI can power truly autonomous and adaptive social robots in education, discussing limitations of past educational robotics research, recent technical advances in AI, as well as concrete examples of applications of AI in educational human-robot interaction, and a reflection on limitations of current AI. By bridging technical and pedagogical perspectives, it shows what the next step in the evolution of human-robot interaction in educational contexts might look like.

Keywords Educational Social Robots, Generative Artificial Intelligence, Educational technology, Adaptive Learning, Conversational Artificial Intelligence

1. Introduction

Commercial generative artificial intelligence (AI) tools such as ChatGPT, Copilot and Gemini have had a noticeable impact on education, sparking both enthusiasm and apprehension among educators and students. Students use it to write their assignments – forcing teachers to reevaluate their evaluation methods – while teachers use it as a tool for automatic grading (Adeshola & Adepoju, 2023; Dempere et al., 2023). Here, we will step away from students and teachers using the AI tools directly and focus on how generative AI can turn social robots into more successful teachers, teaching assistants, and tutors.

Social robots for education have been developed for a more than three decades now (Belpaeme et al., 2018). However, as we discuss in Section 2, they still often lack real autonomy and generalisability, limiting their applicability in real-world classrooms. Meanwhile, the last few years have seen the emergence of many powerful *generative AI* models, which we define here as AI models that generate new content (Feuerriegel et al., 2024). This content can be natural language text, which is generated by *(large) language models (LLMs)*, often based on a piece of input text which is called the *prompt*. They are not only limited to text, but *multi-modal* modals can also use other modalities such as images as input (Gan et al., 2022). Other models also generate new images or sound, often of remarkable human-like quality (Podell et al., 2023).

We argue that these generative AI models will provide new capabilities to social robots, enabling them to surpass the limitations that have kept social robots from entering real-world classrooms. We focus in particular on their autonomy and on the possibilities for *adapting* the interaction between the robot and the learner to the learner's skill level and *personalising* the interaction to e.g. their interests. As AI researchers that develop autonomous social robots for educational scenarios, with this chapter we aim to present our vision of how generative AI will enable these new capabilities to educational social robots. We provide both background information about the concrete generative AI models we deem important, as well as several concrete applications showing how they can be integrated in social robots. With this, we hope to inspire anyone who researches or develops educational social robots, either from a technical or a pedagogical perspective.

This chapter is structured as followed. We first discuss the constraints that currently limit educational social robots in Section 2. Then, we present the various types of generative AI models that we believe will be powerful for educational social robots in Section 3, along with a short history of these recent advances in generative AI. Section 4 then presents how they can concretely be applied in educational social robots, taking a few specific applications that were recently developed as examples. In Section 5, we discuss the remaining and new technical and societal limitations of these technologies, and finally summarize our findings in the conclusion in Section 6.

2. Current Limitations of Educational Social Robots

People have been dreaming of robot teachers for decades – already in his 1951 short story "The Fun They Had", Isaac Asimov speculated about children each having their own individual mechanical teacher. Today, social robots are real, and commercially available robots like Nao, Pepper, Furhat, and others, can be used in real-world settings. In the meantime, class groups are becoming more diverse, with children (and parents) expecting more individualised support and tutoring, and many countries struggling with teacher shortages. Then why are social robots not yet ubiquitous in our schools? We examine what limits social robots today.

2.1 Natural language interactions

Perhaps the most prevalent mode of communication humans use, especially in teaching and learning interactions, is natural language. A teacher explains concepts to their students in natural language, students ask and answer questions, write answers on tests, give presentations, write essays, and of course, in language education, people learn by practicing a target language.

However, this immediately lays bare one of the central challenges of social robotics: holding natural conversations. Until recently, most robots' conversational abilities were powered by rule-based chatbots. This means that the robot programmer had to envision all possible utterances that the user could say and program appropriate responses for each of them. Templates provided some flexibility, meaning the user can substitute some words in the programmed template sentences, and more advanced *intent recognition models* could automatically categorize which option the user intended to say, within a predefined list of options. Data-driven approaches allow for more flexibility, but require data collection which is very expensive in HRI (Reimann et al., 2024).

This approach clearly limits the possibilities and autonomy of the social robot. Suppose the user answers the robot's question or instruction in a different way than was envisioned by the programmer, or worse, when the user asks the robot an unexpected question, the robot will often not understand the user, leading to the interaction failing and possibly the user losing trust in the robot (Flook et al., 2019)

As a consequence of this, many research studies that investigate the effectiveness of social robots in educational scenarios use tightly scripted interactions, or even rely on remote-controlled interactions. The latter is often called the "Wizard-of-Oz" approach, and relies on a human behind the scenes choosing what the robot says and does. While these scripted or "wizarded" studies provide insights into the potential benefits and effects of social robots in education, these systems are not yet ready for autonomous deployment in the real world (Belpaeme et al., 2018).

Another major limiting factor for human-robot conversations is the performance of automatic speech recognition (ASR) systems. While ASR has made great strides in recent years, *really good* performance is only achieved for typical populations, i.e. adult native speakers of a language for which many training examples are available, such as English, and preferably not speaking a dialect of this language. Especially for young children, speech recognition has long not performed well enough to work in autonomous interactions (Kennedy et al., 2017). Robot designers have had to resort to other modalities of interaction to cope with the lack of robust ASR, such as touchscreens where a child can select an option, or perhaps type a response (Belpaeme et al., 2018).

Taking into account both the necessity to rely on (partly) scripted conversations, and the lack of robust ASR for children, it becomes clear that social robot tutors could so far only be deployed in very narrow situations, only focussing on specific topics for which time-intensive work was needed from the robot designer to program the interaction, and not supporting free-form input from the user. This has also limited which didactic methods could be used in the teaching and learning interactions, as methods that rely heavily on conversations, with a lot of learner input, could not yet be reliably implemented. Specifically for (second) language education, this has resulted in research focussing mostly on vocabulary teaching, as real conversational practice was not yet feasible (Randall, 2019).

2.2 Multi-modal inputs

Of course, humans do not only use language to communicate. We also convey information using other modalities, such as through what we can see or non-verbal sounds. People use social cues such as facial expressions, eye gaze, gestures, body pose, and prosody, to express emotions, and to infer information about others' mental states (Mehrabian, 1972). Crucial in education, teachers can use this information to assess how engaged and motivated a student is, but also whether they understand the material that is being discussed or if they are confused (Schutz et al., 2006).

However, as we also saw in the previous section, tasks that are mostly intuitive and natural for people, such as conversations in natural language, are exceedingly difficult for AI – a principle often referred to as Moravec's paradox (Moravec, 1988). The same goes for interpreting such social signals. People express emotions and social signals in very subtle ways, often using movements that are only shortly visible (Krumhuber et al., 2013), and much less exaggerated than the stereotypical emotional expressions that are often found in datasets used to train AI models. These expressions are also highly individual (D'Mello et al., 2018), and furthermore difficult or even impossible to interpret without the context of the interaction (Mesquita & Boiger, 2014).

While social signal processing, in educational applications especially focused on detecting engagement, has been an active research area for a number of years, systems for detecting engagement are often limited in their success to very specific contexts, and not yet ready for complex, real-world scenarios (Cumbal et al., 2020; Gunes & Churamani, 2023).

Besides understanding social cues, educational robots should also be aware of their physical environment. As they are physically embodied, they share this space with the user. Their embodiment, which is often humanoid and thus contains eyes, also creates the expectation in the user that the robot can see. However, the conversation models discussed in the previous section are often not, or only loosely, connected with the visual input of the robot: chatbots are not designed with eyes in mind (Reimann et al., 2024; Janssens, 2024).

Using visual information about the environment, however, holds many opportunities for educational robotics. The robot could refer to objects, people, or actions that are happening, which could aid in vocabulary acquisition (Bara & Kaminski, 2019; Wolfert et al., 2024), or gestures could be used by the user to provide more information in the interaction. If the robot does not have access to this visual information, it might also not understand some of the user utterances, if they refer to elements of the environment.

The visual information can also be used to make the robot more sociable: it can personalise the conversation based on what the robot sees of the user. Going much further, the robot could personalise the entire interaction. Humans also adapt the way they speak, and especially how they explain things, to the person they are speaking to, and we also base this on the way someone looks. Simple personalisation of interactions, such as using the user's name, has already been shown to increase engagement in human-robot interactions (Henkemans et al., 2013), so adapting the interaction more extensively to the specific user could lead to even better outcomes.

However, currently, social robots are limited in how they can use visual information in interactions. Just as the conversations need to be largely scripted, systems currently need to be specifically designed to recognise a given set of objects at certain moments during the interaction (Wolfert et al., 2024).

2.3 Adaptation and personalisation

Perhaps the most promising aspect of using social robots in education is the potential for a completely personalised tutor for each learner – as Asimov already dreamt of in 1951. However, given the limitations already discussed, this dream is still far-removed. As interactions needed to be extensively pre-programmed, all options for modifying the interaction needed to be foreseen by the designer. Concretely, this means that interactions cannot be easily personalised to the learner's interest, by for example letting the learner choose a topic on which to converse during language learning. Also, providing new didactic content is often challenging, as the programming of the robot is often not sufficiently accessible to teachers (Chevalier et al., 2016). Adapting the content of the learning interaction to the learner's performance, current skill level, or prior knowledge is also an important aspect of personalised robot tutors, but requires more flexibility than is available with the current conversational models.

Besides providing different options the robot can choose to use in its interaction with the learner, a major challenge is also knowing which of them to choose. The robot can, for example, choose interventions to improve the learner's engagement, or to adapt to the learner's skill level and aim to improve learning gain. However, for both cases, a model for the learner's engagement or knowledge is needed (also called *student modelling* (Chrysafiadi & Virvou, 2013)), as well as a method to predict which impact each option has, allowing the robot to select the optimal action.

Both building these models and designing the action selection methods are challenging. One of the aspects that makes this challenging is the difficulty of detecting and interpreting non-verbal signals. However, progress has been made in building action selection models, e.g. using interactive reinforcement learning (Belpaeme et al., 2018).

3. Recent Advances in generative AI

As discussed in the previous section, one of the main limiting factors for social robots in education is their ability to process and express themselves in natural language: most of our educational interactions take place in the form of natural language. We claim that for social robots to fully integrate into education, they must master this form of communication. This would have remained a pipe dream, if not for generative AI. Generative AI offers the potential to converse with tutoring systems using natural language. This section describes the technology that made Generative AI, such as Claude or ChatGPT, possible, and paints a picture on how it created a revolution not only in language modelling, but also in speech recognition, image generation and multi-modal language models. We also look at how the field of natural language processing (NLP) has gotten to this point, and what technologies made this possible, as well as the advancements those technologies introduced for speech recognition, image generation, and multi-modal language models.

3.1 Language Modelling

Much of the data created by people is unstructured: natural language, meant for other humans – think about books, social media and blog posts, news articles and much more. This means that it is not organised in a way that computers can directly use it. The field of NLP tries to bridge that gap, by processing human language so that it is useable for computers. It has typically focussed on solving tasks that involve natural language, and one of them is intent recognition. It is useful for, among other things, recognising what a user is trying to communicate. Related is sentiment analysis, where the emotional intent behind a piece of text is analysed. An example application of this is a company trying to estimate the public opinion of their product based on what is said about it on the internet. To gain structured information from a text, named entity recognition can be used. This task is about trying to extract named entities, such as people, places and organisations, from text. This can then be used for further classification. Another NLP task that is useful in (educational) robotics is ASR: transcribing text from speech in an audio file. Then, the transcription can be fed to further processing steps that expect written text.

The NLP task that we will focus on here is language modelling, as the recent advancements in it are what we believe will have a great impact on educational social robotics. Modelling languages boils down to predicting the probability of a word given previous words. Many attempts have been made, staring with statistical models to, more recently, neural networks. If, given a few words, we know which words are more or less likely to follow it, we can use this information to choose a next word. Then, we add it to the words we already had, look at the probabilities of the next word, and again choose the following one. This way, language modelling is the key to generating text, which is what we needed to ease the natural language limitation in educational robotics. For a model to learn how to do this, it needs data – luckily natural language examples are abundant on the internet. Multiple approaches are used for training models to do this task. You can train a model by giving it the start of the sentence and having it predict the next word, as this word is of course known. BERT, the language model developed by Google in 2018, was trained using a *masked language objective*: some words in the training text were randomly deleted (*masked*), and the model had to estimate what word it should have been (Devlin et al., 2018). This task allowed the model to learn probabilities of words in context. Then, even if the model was not directly trained for predicting the next word given a part of a sentence, it still learns this task as it is closely enough related to the task it was trained on. This concept is called transfer learning: training a model on one task, and using it for a related, but different task. Transfer learning makes a model much more useful, as it can be applied to more than just the specific task that it was trained for. Often, language models are trained on a modelling objective as described above, making them a *general-purpose* model, after which they are re-trained using smaller amounts of data for a more specific downstream task, such as intent recognition, sentiment analysis, or even summarising text and answering questions about it. This process is called *fine-tuning*, and it is a form of transfer learning.

Now that we understand the goal of NLP and the importance of language modelling, we can look at the revolution in generative AI that we believe will have an impact on educational social robots. The part of the revolution that was most visible to the general public was OpenAI's release of ChatGPT at the end of 2022 (Kasneci et al., 2023). ChatGPT's sudden success was made possible by a technology that was proposed a few years before, in a widely cited publication: *Attention is all you need*, by a team of researchers at Google (Vaswani et al., 2017). This publication introduced the Transformer architecture. What sets this architecture apart from previous language modelling techniques, is that it allows text to be processed in a much more efficient way. The previous state-of-the-art – recurrent neural networks – tried to capture the sequential nature of textual data by processing it piece by piece while keeping a hidden state – a summary of all that it had processed already. While this technique showed great potential, its inherent limiting factor is that text must be processed sequentially. When processing large amounts of text – as is necessary for modern, data-hungry techniques – this is just too slow. The Transformer architecture avoided the need for processing sequential data sequentially by introducing the concept of *attention*, as referenced in the name of the publication that started this revolution.

To understand attention, we will first take a look at why processing language is a difficult task. Within a sentence, words influence each other. Most of this influence is short distance: the article is decided by the word that follows. But, in some cases, this influence reaches a word that is further in a sentence, with many words in between. For example, the conjugation of a verb is decided by the subject, which can be multiple clauses away. Recurrent neural networks process such a sentence word by word, while keeping a summary of what was seen before. To predict the next word, this summary must contain the number of the subject – which might have been the first word of the sentence – so it can be remembered until the verb appears – which might be the last. The challenge of this long-term memory that is necessary for language understanding, together with the inefficiency of processing text piece by piece, is what the attention mechanism solves.

The idea behind attention is that, when processing a part of the data, the model first determines what relevant information can be found in other parts of the data. Following the example of the subject's influence on the verb's conjugation, the attention mechanism would decide that the subject is important for predicting the verb, as well as the object of the verb, while other words in between might not be. In attention, this problem is solved as follows. A sum of all words, weighted by their importance to the current word, is used as input during processing. As this weighted sum can be calculated for each word in relation to all other words simultaneously – allowing for parallelisation, the limitation of processing the data piece by piece, as in recurrent neural networks, is overcome. This way, the Transformer architecture can easily model long term interactions in sequential data, while processing more of this data at the same time.

The introduction of the Transformer architecture allowed for faster, more parallel processing of data, fitting perfectly in an era of end-to-end, data-driven technologies: large models, trained directly for their end-task, using large amounts of data.

The new possibility to train these models on such large amounts of data, allowed for advancements in their performance that were unreachable with previous technologies. These models are often referred to as Large Language Models (LLMs).

At their core, LLMs do one thing: given some text, they predict the next word – just as your smartphone's keyboard does. Therefore, you can ask it a question, and it generates an answer word by word. The resulting text is often very impressive and human-like, which makes it seem like there is more going on than just predicting a next word. The difference between a LLM and your smartphone's keyboard is that LLMs are trained on such large amounts of data, and on larger pieces of text at once, that they can find a pattern in the structure of language. It is important to note that this pattern recognition, which leads to very human-like text, is not the same as reasoning or even understanding. LLMs are just very good at generating text that seems like what an intelligent answer would be, but the reasoning steps that happen behind the scenes when a human is talking are not implemented in LLMs.

We have established that LLMs take text as input and generate text that seems likely to follow the given text. The text that is given as input is often called the *prompt*. Using this prompt intelligently, can greatly improve how well the generated text matches what you are trying to do. This is a new form of transfer learning: the model is trained to model the probabilities of words, but using a prompt, we have it fulfil a different task – like before, such tasks can include intent recognition, sentiment analysis, summarizing texts, answering questions, or generating example sentences about a specific concept. Explaining the task or asking a question is called *zero-shot* prompting. Providing some examples of the task at hand is called *few-shot* prompting and often already provides a noticeable improvement (Brown, et al., 2020). Prompting an LLM to not immediately generate an answer, but to provide a *chain of thought* – some intermediate reasoning steps – improves the model's reasoning skills, or it at least appears to do so (Wei et al., 2022). Actually, while the model generates these intermediate steps, it is still choosing what is most likely in a statistical sense. As the intermediate steps are usually more obvious and simple tasks, there is a higher chance that the model encountered something similar during training. Then, these intermediate steps are more likely to be correct. As the model then takes these generated steps as input when generating the final answer, its probability of correctness improves. Asking the model to generate intermediate steps is forcing it to reason out loud – which is the only kind of reasoning such a model can do.

Prompting LLMs makes them suited for a variety of tasks, as they can generate language that is generally correct, human-like and - with intelligent prompting – adheres to the task quite well. Their usefulness for educational social robots is easy to imagine given this impressive ability, and knowing what limited social robots from fully succeeding in educational applications until now.

3.2 Automatic Speech Recognition

A similar story can be found in the performance of ASR: recognizing the speech in audio fragments and transcribing it. ASR has been researched extensively, and useful results have existed for years. In optimal settings, error rates of around 5% are reported, which is similar to or even better than human annotation (Xiong et al., 2018). Though this sounds very promising, robust speech recognition that can handle suboptimal environments, such as a noisy room, has been hard to attain (Karpagavalli & Chandra, 2016). Early models were heavily engineered, with handcrafted features that relied heavily on expert knowledge. This is often not robust and does not necessarily transfer between languages. Then, similar to the evolution in language modelling, there was a shift to end-to-end models. These were trained on raw audio and transcriptions, skipping the feature engineering step. End-to-end models tend to be too large for use on mobile devices, so cloud-based solutions appeared, such as the ones provided by Google and Microsoft. Finally, attention-based – the core improvement of the Transformer – ASR were designed, by universities and companies such as Google (Chan et al., 2015; Bahdanau et al., 2016). Around 2022, OpenAI presented Whisper, an open-source ASR system. It uses a Transformer model as published in the original paper and is trained on an impressive 680,000 hours of data (Radford et al., 2023). This again shows what is made possible by the Transformer architecture: an end-to-end (speech-to-text) application trained on large amounts of data. Next to the computational ability to train a model on so much data, another challenge is to actually gather that data: expert-transcribed audio data is not available in abundance. The data that was collected for training Whisper is *weakly supervised* – not only transcriptions by expert are used, but the vast majority of the data is also made up of messier transcriptions that are not necessarily high quality. This allows for much more training data, which makes a noticeable difference. Whisper has an impressive performance close to that of humans for English (Radford et al., 2023), and as the training data contains several languages, it can also transcribe other languages – with the performance of course dependent on the available resources of the language.

In interaction with humanoid robots, humans expect to be able to use speech – so speech recognition is a necessary tool. The advent of the Transformer and attention-based neural networks shows how robust speech recognition is possible, even in suboptimal situations. Especially in education, this is important. Classrooms are seldom not noisy, people learning languages do not always articulate perfectly and speak with accents, and child speech is generally much harder to understand than adult speech. Therefore, robustness is of utmost importance in this application, and Transformers show great potential in achieving it (Janssens et al., 2024a).

3.3 Image Generation

AI generated images are appearing everywhere the last few years, causing worry among artists (Ghosh & Fossas, 2022) and enthusiasm amongst the less artistically gifted. Not too long ago, generating an image based on a description of it was still science fiction. In the early days of image generation, many images of a certain subject were fed to a model, allowing the model to generate a similar image. This was often done with Generative Adversarial Networks (GANs). In a GAN, one neural network – the generator – generates images following the distribution of the training set, while another neural network – the discriminator – tries to distinguish fake, generated images from the training set. Then, these two models are trained simultaneously: the generator becomes better at generating images, while the discriminator becomes better at discriminating between real and generated images (Goodfellow et al., 2014). While these GANs showed some promising results, having to provide a large enough set of example images to generate a new image, is still very far away from the text-to-image generation we see nowadays. An attempt at a text-to-image model based on GANs was made, with encouraging results on specific subjects and at most *visually plausible* results for more general subjects (Reed et al., 2016).

Again, the real revolution that brought AI generated images into the public eye, was started by applying the Transformer architecture to the problem. In 2021, OpenAI released DALL-E, a Transformer-based text-to-image model, which is an extension to GPT-3, their LLM. It takes text and possibly images as input and returns images as output. A year later, DALL-E 2 was released, as their newest model that generates more accurate and realistic images. DALL-E 2 is a diffusion model: it starts from an image that is just noise, and iteratively removes noise based on the text it was given, to increase the likeness of the image to the content of the text. Stable Diffusion is a similar text-to-image model that is open-source (Podell et al., 2023). Many of these image generation models are also used to edit existing images, such as changing the style of the image, changing the posture of the subject and adding things as the user wants (Kawar et al., 2023).

As high-quality image generation is now possible, advances are made in how to personalise these images. Techniques such as *textual inversion* are used to generate images with specific objects, animals or people in them, based on just a few pictures of it (Gal et al., 2022). This works by finding which 'words', in the vocabulary of the image generation model are closest to the object that we want to represent. Of course, the model – as all models do – doesn't use natural language words internally, it uses a numerical representation of them. Therefore, it can make up new words – new numbers – as needed, which can then be exploited for the personalisation of the image it generates.

Finally, the latest step in the evolution of image generation is a natural extension of it: video generation. As a video can be considered a sequence of images, or even an image with the dimension of time added to it, the aforementioned progress in image generation also translates to video generation, or text-to-video models. Again, OpenAI impressed the world in early 2024 by demonstrating Sora (Brooks et al., 2024): a text-to-video model that is not yet released to the public at the time of writing. It uses similar technologies as the DALL-E models, with a combination of diffusion models and transformers. As this technology is still quite new, it is exciting to wonder what the future of text-to-image and text-to-video will bring, but it is not hard to image that there will be numerous possibilities to apply this to educational settings.

3.4 Multi-Modal Language Models

While language models as discussed above have shown impressive performance in understanding natural language and generating new text or even images based on their input, one piece of the puzzle is still missing: understanding visual input.

Initial language models were not made with multi-modality in mind. They only process information from one modality – text – and generate more text in return, or use that text for other tasks such as predicting a sentiment score or classifying it. However, as we discussed in our first section, social robots are situated in the physical world. They need to be aware of their environment and need to be able to integrate what they see into their language. As social robots are often embodied in a humanoid shape, which usually includes eyes, users expect that the robot is able to see. They might ask the robot questions about an element of the environment, or make references to the environment in their speech. Hence, the robot requires visual information to understand the user.

This problem is often called *situated* language interaction: making the robot produce and understand language that refers to the physical and social world around it (Goodwin, 2000). This is also related to the concept of *grounding*: connecting words that refer to elements of the world, such as “red” or “table”, or even abstract concepts such as “jealousy” to semantics and properties (Janssens, 2024). Grounding is seen as one of the important challenges that the field of human-robot interaction (HRI) poses for AI (Lemaignan et al., 2017). It is also a concrete implementation of the wider symbol grounding problem (Harnad, 1990).

Efforts have been long underway to combine vision with natural language. One of the first and most-studied tasks in this area is that of *image captioning*, or writing a description of an image in natural language. Researchers were able to develop well-performing models for this task because of the release of large-scale annotated dataset such as MS COCO (Lin, et al., 2014) and Visual Genome (Krishna et al., 2017). These models often consisted of a convolutional neural network (CNN) that processes the image and a recurrent neural network (RNN) that generates the text. Later, Transformer-based models were introduced to replace the text-generating and also the image-processing models.

This work was then extended to more conversational tasks, such as visual question answering, where the AI model answers natural language questions about an image (Gan et al., 2022), and visual dialogue, where the user and the model can have a longer conversation about the image (Das et al., 2017).

These models were originally built by training them for one specific task – however, this required a large dataset for that task. However, just like for text-only models, pre-training on large datasets that were scraped from the internet, such as images together with their alt-text or already provided captions, allowed the models to become more powerful and be fine-tuned to specific tasks (Gan et al., 2022). Some work-arounds also exist to integrating this visual context into language models when visual data is not sufficiently available, such as using a model that was specifically trained for image captioning, and then providing those captions to a text-only model (Janssens et al., 2024b).

Today, vision-language models are slowly becoming as powerful as the language-only models. As the models and the datasets they are trained on are becoming larger and larger, they are becoming better and better at many *zero-shot* tasks, i.e. without being fine-tuned. Prominent examples include GPT-4o (OpenAI, 2022), BLIP-2 (Li et al., 2023) and LLaVa (Liu et al., 2024), the latter two of which are open-source. These models can write descriptions of images, answer questions or hold dialogues about them, but also perform more specific tasks, without needing to be fine-tuned.

4. Application of generative AI in Educational Robotics

In the previous section, we looked at the technological advances that allowed for what can be considered a revolution in AI in the last years. We believe that this wave of advancements can have a strong impact on the success of educational social robots, so this section will describe our vision on how to leverage it optimally. In our vision, we put the social robot in a tutoring role: it does not take the teacher's place, but it can support the teacher during classes, exercises, extra tutoring sessions and homework. Within the framework of this robot tutor, we see a few current limitations that generative AI can help alleviate. First, we look at content generation through generative AI: using tools like LLMs and text-to-image models to write or illustrate the educational content that is delivered by the robot. Secondly, we discuss adaptation, where the flexibility of generative AI allows for adjusting the difficulty of the content to the student based on their personal progress, and thirdly, personalisation, where the content's theme can be fit to the personal interests of the student. Finally, the potential of integrating multi-modal inputs such as vision and speech into a robotic teacher is explored.

Because of personal interests and experience, our focus will be slightly biased towards – but not limited to – language education.

4.1 Content Generation

A large part of education is providing students with the content that contains what students should learn: explanations, both visual and textual illustrations, examples … In language education this is especially important, as much of language learning comes down to exposure to the language (Ellis & Wulff, 2020).

If a social robot is used for educational purposes, there are three options: all content is provided to the robot beforehand by teachers and educators – in other words, what the robot does is scripted, there is a human in the loop that adjusts the robots behaviour to allow for flexibility in the lessons – in other words, the robot is (partially) teleoperated, or the robot must be able to adjust its lesson on the fly (Senft et al., 2019). The first option is inherently limited: it is not possible to predict all interactions between student and robot, we cannot foresee with which part of the lesson students will struggle or what questions will be asked. This is not to say that there is no use in scripted lessons – sometimes delivering fixed content is the main point of a lesson, but to fully exploit the advantages of educational social robots, some flexibility in the robot's behaviour is necessary. The second option – a (partially) teleoperated robot tutor – is useful and necessary in a research context, but not scalable and most of all, it does not actually help alleviate teacher shortages: if we use robot tutors as classroom support, but each robot is controlled by a teacher, then we could have just added extra teachers, and the robots would not have been necessary. Therefore, we believe that the future of educational social robots is in flexible, adjustable content and that this is one of the main current limitations where generative AI can have a real impact.

When a robot tutor is asked questions, asked for an explanation or to give examples, it should give a natural language response. LLMs are good at exactly this: generating text based on other text. In language education, the correctness of the language of the answer is the most important, but during, for example, a history class, the information provided by the robot must also match the school curriculum while being factually correct. LLMs can therefore be extended using Retrieval Augmented Generation (RAG), where – next to the implicit knowledge that an LLM has by knowing the probability of words – an external memory is added to the model (Lewis et al., 2020). In educational context, this could be the course's textbook. The LLM can then search the textbook for the relevant information, add this to its inputs, and generate a factually correct answer based on its language knowledge combined with the relevant information. The combination of LLMs and techniques like RAG can allow social robots to support teachers without the added work of preparing scripted lessons.

When delivering content to a student, it can help to also illustrate this visually. In language learning, the meaning of words and concepts can be conveyed to the student visually. That way the need to translate everything to the student's native language is avoided, as the benefit of translation in language learning is a contested topic (Dagilienė, 2012). An additional benefit is that a language learning system that does not depend on translation, is agnostic to the native language of the student so it can be more readily reused. The most intuitive application for visually representing words is vocabulary learning: images of object are shown, either during exercises or when the robot talks about these words. We have shown this to be effective in a proof-of-concept study, where students play a game to learn vocabulary. They are shown a set of images and are told a description of one of these images – either by a robot or played from a tablet. All the images and descriptions were AI generated, based on the estimated language level of the student. The students had to choose which image fits best to the description. The game was played in a language that the students had very limited knowledge of, and even though no translations to their native language were used, they showed a statistically significant learning gain. This experiment shows the power of generative AI in adaptive content generation in educational contexts (Verhelst et al., 2024).

In other parts of language learning, similar approaches can be developed. As much of language learning depends on exposure to the language, adding visual support of what the robot is talking about can help the student to better understand what is said. Imagine a robot telling a story interactively, using students' input for the progression of the story – which is of course generated using an LLM. This will help engage the students, as they are actively part of the lesson, and they can influence the story to be more interesting to them. The students are still learning the language, so they might not understand each word or language construction the robot uses, but generative AI is used to visualize the story as it is told. This can keep the students' attention, while filling in the gaps of what the students don't yet understand independently: a perfect combination for language learning.

Generating visual content can of course also be used in education other than language learning. Everywhere, from geography to mathematics classes, illustrations are used in textbooks. Ensuring that we can generate illustrations that consistently adhere to the curriculum is not self-evident, but using techniques like editing existing images using generative AI can circumvent these issues: starting from correct images and editing only what is needed to fit the context.

4.2 Difficulty adaptation

In an educational context, it is important to deliver learning material of the correct difficulty level. Following Vygotsky's theory on the zone of proximal development, a student learns most from content that is slightly too difficult for the student to solve independently, but just right with the help of a more advanced peer or an adult (Chaiklin, 2003) - or, of course, a robot. Following this theory, educational content should not be the same for every student. An educational social robot should have some way to estimate the student's level, either by modelling the student's knowledge, or collecting feedback during the lesson, in the form of incorrect answers and expressions of confusion, both spoken and facial. If interactions with the social robot are largely scripted, it is difficult to dynamically adapt the difficulty of educational content to the student.

As generative AI allows for exercises to be generated during the interaction, it is possible to use what is known about the student's level to influence which exercises are generated. This can mean that we either adjust the language that is being used, or which content is being delivered to the student. During a mathematics class, we could monitor whether a student regularly makes mistakes on a certain topic. If such a topic has been identified, the lesson can be dynamically adjusted to focus more on this topic. It can be explained again or in more detail, and – using LLMs to generate the text, while the content is taken from the textbook using RAG – the explanation will be on the same topic but phrased differently. More exercises on the same topic can be chosen, until the student consistently shows good results. This way, the same concept can be taught to all students, starting from the same textbook and a loosely scripted lesson, that is then adapted to each student based on their answers to exercises, their feedback, their non-verbal communication and their questions.

To allow knowledge to be transferred to students, their understanding of this content should be maximalised. Therefore, all language that is used in class should be in reach of the student's language level. Even if the content that is to be delivered is scripted, generative AI allows for flexibility in the difficulty of the language in which it is delivered.

For language learning, most of the content is of course language itself, so this is extra relevant. As mentioned before, in language learning, exposure to the language is of utmost importance (Ellis & Wulff, 2020) and ensuring that the language level is matched to that of the student has been shown to increase learning, and might increase the students' participation during class (Randall, 2019). Adaptation of the lesson can be applied to exercises on e.g., grammar, similar to what was mentioned above, but it can also be applied to conversational learning. If a robot tutor is available during a second language course, it can play the role of a native speaker in the language that the student is learning, providing examples of typical ways of speaking and correct pronunciation, while serving as a conversation partner. The content of the conversation does not have to be educational, as chatting with a native speaker about anything is increased exposure, thus useful. As many schools struggle with a shortage of teachers, and these teachers often do not have the time to practice conversational skills with their students, this might be a valuable asset for schools to bridge the gap from theoretical language learning to actual communicative skills.

Influencing the difficulty level of language that is generated by a model can be done by prompting the model to generate language of this level, by prompting the model to match the difficulty level of the student's inputs or by using a language model that is specifically trained to generate language of a certain difficulty level. The last option is the least flexible, as training a model must be done beforehand, but is likely the most robust way to ensure the desired behaviour (Lester et al., 2021).

Previously, adjusting the difficulty level of the content in educational technology could only be done by choosing which exercise to give the student, or by changing the exercises based on some predefined parameters that change the level of the exercise. Using generative AI for generating educational content gives us much more flexibility: the difficulty of content can be dynamically adapted to the student. As this can be done based on any input by the student – not only answers to exercises, this can happen at a much more fine-grained level than what can be predefined. This allows educational tutoring systems to adapt to every student individually, to create the best possible learning environment.

4.3 Personalisation

We have seen that generative AI can be used to generate educational content and that it can be used to adapt the content to the student's educational level. Personalisation also involves adapting the educational content to the student, but with a focus on improving the student's engagement, by ensuring a connection between the content and the student's personal life, making it more relatable (Belpaeme et al., 2018). Students who enjoy sports more than they do mathematics classes might be more motivated to learn mathematical operations if they can do so by calculating the score of their team given a description of the match. So, while the content stays the same, personalisation can be used to package the contents in an engaging way.

LLMs can generate exercises that involve students' hobbies, and image generation models can be used to illustrate them. If these illustrations involve people, we can change their appearance to be more familiar to the student, as people tend to learn better from those who are more like themselves, and it can make students feel more included and involved in the lesson (Binderkrantz & Bisgaard, 2024).

Additionally, new techniques like textual inversion, allow for inserting certain people, animals or objects into generated images. Catering the content to students' interests might help to engage them, which is beneficial for learning. We can use these techniques to insert elements of the student's life into educational content. In educational materials, illustrations are often used to make the lesson more visually appealing (Houghton & Willows, 1987). If these illustrations are altered using generative AI techniques, to include the student's favourite tv character, family or even the student themselves, it can be more engaging, while starting from a textbook image ensures that the content is still correct.

For language learning, the conversational robot tutor, with which you practice language of your own level, can talk about anything you like, even remembering your interests and asking about your hobbies. If the main goal of the conversation is practicing the language, the options for personalisation are endless. If we are trying to convey specific concepts, personalisation can be done as described similarly to other subject, conveying the same content but packaging it according to the student's interests. Visual support can be adapted to the student's life: imagine the storytelling robot from before, where students could change the story's narrative interactively. Personalisation can allow it to include their pet dog, and generative AI is now able to actually show this pet in whatever adventure the student chooses for it.

4.4 Multi-modal interactions

The revolutions in AI allow educational robots to do more than generate textual or visual content with endless possibilities for on-the-fly adaptation – they also allow the robot to perceive the world around it through multiple modalities, such as what they see and hear. This enables even more immersive and effective learning interactions.

Let us return to the example of language learning. Earlier, we discussed how visual content could be generated to support vocabulary learning. Now, this visual content could also be replaced by the physical world around the robot and the learner themselves. In Wolfert et al. (2024), we showed that students learn vocabulary better when they are provided a visual referent of the concept they are learning, either on a screen or with a physical object. This approach could be much expanded. Everything the robot sees about its environment could be provided to the language model that is driving the conversation. The robot could therefore make references to objects or people in the surroundings, or even to movement, and the robot could ask or answer questions about the environment. These strategies can enrich natural conversations between the robot and the learner and could help the learner by grounding the language they are learning in the physical world, and by offering opportunities to practice concepts that require a physical referent in the environment. Visual content could also help to personalise the interaction with the user, potentially boosting motivation and social bonding, such as in (Janssens et al., 2024b), where the robot addresses the user by asking a question about a feature of the user that the robot sees, such as their apparel.

Not only visual input can enrich the learning interaction: also auditive information can be important. We discussed earlier that speech recognition systems have improved considerably in the last decade, including for smaller populations such as children. Now, finally, real conversational interactions between robots and children or non-native speakers might be possible. One could even dream of systems giving feedback on pronunciation. As an example, (Kennedy et al., 2024) have managed to extract pronunciation information from users saying their names, to enable the robot to correctly pronounce those names. (Amioka et al., 2023) also explored pronunciation learning with social robots, by generating realistic lip movements for the robot when pronouncing the words to be taught – however, learners who do not yet have any knowledge of the language do not seem to benefit from these realistic lip movements.

All modalities will have to be integrated to make the robot aware of its social environment. Social cognitive aspects of the learning interaction, such as the learner voluntarily or involuntarily communicating that they are interested, engaged, motivated, uncertain, struggling … should be detected by the robot and provided to the AI models generating the interaction, in order to adapt the interaction to the learner, improving both the learning and social effectiveness of the interaction.

Finally, we recognise that we have so far omitted the modalities of touch and proxemics (i.e. how close the human and robot are standing from each other). Both play important roles in human-human social interactions, and have been shown to also modulate human-robot interactions (Ren et al., 2023; Mumm & Mutlu, 2011). This could be leveraged to further improve educational robots, but has not been widely studied yet.

5. Technical and societal limitations of generative AI

Recent advances in generative AI allow us to dream big: it is now possible to have conversations with robots about anything, speech recognition seems to work for typical populations and is rapidly improving for atypical populations, and we can generate images and even videos that contain anything that can be thought of. Still, some challenges are left before generative AI can freely be applied to educational robotics.

5.1 Real time conversations

For a conversation to feel natural, the timing must be right. In human-human conversation, a lot happens to decide who speaks when: people know when their turn to speak arrives, based on eye contact, tone of voice and context. When changing turns – one person stops talking and another starts – people leave only 200ms of silence on average, and often our speech even overlaps. For conversational systems, a response time of 700-1000ms is often deemed acceptable, even though this is already much longer than humans usually take (Skantze, 2021). That means that a robotic tutor must be able to provide an answer to students in less than one second. In this short time, many things must happen: the speech recognition must provide a transcript of the question, in some cases, e.g., when using RAG, a search must happen, then, a language model must generate an answer based on the question and perhaps the search results and finally, speech must be synthesized to play from the robot's speakers. There are ASR systems that manage to generate a transcription in sub-second transcription time, but this is only one step of the way (Janssens et al., 2024a). Answer generation by an LLM is reported to sometimes take up to three seconds, bringing us far over the maximum of one second, even without speech synthesis (Irfan et al., 2023). Therefore, before natural human-robot conversations can take place, the efficiency of the involved systems must be dramatically improved.

5.2 Controlling generative AI

As discussed before, controlling the output of generative AI is generally done with prompt engineering: describing the task at hand to the model, using zero-shot or few-shot learning. Although this is a powerful way of applying a general model to a task it was not specifically trained for, tasks must be clearly and unambiguously described, using at most a few examples. This technique does not suit all tasks and there are limits to the control the user has over generated output when prompting a model. Prompting can have very variable results, with changes as small as adding one word already leading to vastly different results (Liu et al., 2023). Another often occurring problem is so-called hallucination: generated output diverging from the desired output, earlier output or conflicting with real-world facts (Zhang et al., 2023; Alkaissi & McFarlane, 2023). This is detrimental to the reliability of the outputs, especially when used in real-world applications. Additionally, strict control over the tone and language level of generated output is difficult, while ensuring appropriate text – both in difficulty level and in content – is of utmost importance in educational contexts. Additionally, if a social robot is expected to deliver some parts of a school's curriculum, we must be able to guarantee that it actually does this, and that it does not tell the students anything incorrect or confusing. Finally, as LLMs are trained on human data, if an LLM is told that what it said was incorrect, it tends to agree, no matter the truth. So, if a conversational system that adheres to the curriculum and that tells the truth is created, students might think it is wrong, question it or correct it, and get fed false information anyway. In educational settings, this can have strong consequences, with students learning incorrect things and not knowing who to trust. Therefore, having strict control over generative AI's output is needed for reliable educational social robots.

5.3 Sustainability

While generative AI and all technologies based on large neural networks are very powerful, its sustainability leaves much to be desired. When a neural network processes information, it does calculations with this data. Each of these calculations takes a small amount of energy. The revolution in these technologies, as discussed in earlier sections, allowed much more calculations to be done in shorter times. Processing data – doing calculations – could now be done in parallel. Of course, if each calculation takes some energy, parallelizing the calculations so we can do many of them at once makes it more energy consuming per time unit, while requiring less time. But the revolution in AI showed that these models had the capacity to find patterns and mimic reasoning when given enough data. So, these models were trained on more data, giving us more impressive models: larger, with more input and better output – and a larger energy consumption. The number of computations that our AI models require has been growing exponentially for a while now, together with their energy demands (Strubell et al., 2020). Of course, there is something to say for an increase in energy demand if it comes with increased performance, but the demand that comes with exponential growth is not something that our planet can afford. There are a few things that we can do to minimise this impact while still reaping the benefits of these impressive technologies (Lacoste et al., 2019). The location where a model is trained determines the energy source, how sustainable it is, and the carbon footprint of the model. Choosing this wisely is impactful and should be communicated transparently when making a model available. When using a model, its impact can be influenced by being mindful of the energy sources used in deployment. Additionally, choosing not to train your own model, but finetuning an existing one or even just prompting it without additional training minimizes the computational resources used. Finally, the hardware on which the AI models are trained and used, as well as the hardware of the robot, have a climate impact. More efficient hardware, which requires less compute time or performs your computations more effectively, will minimize the impact of this hardware. So, while using this powerful technology opens a lot of doors, especially in educational robotics, its consequences should be considered every step of the way, so that climate impact can be minimized where possible.

5.4 Privacy

When using LLMs and other generative AI in an educational context – often with minors, privacy is an important concern (Mhlanga, 2023). Most of the currently best performing LLMs, such as OpenAI’s GPT family and Google’s Gemini, are commercially available through an API. This raises the problem that using it entails sending potentially sensitive data of vulnerable people to a company without knowing exactly how this data is saved and whether the company uses it, while issues of these models' adherence to privacy standards have been reported (Cartwright et al., 2024). It is not unthinkable that companies that provide LLM services, would use this user data to further optimize their models. While this is understandable from the perspective of optimizing the models, children's data in schools should not be freely used like that. Therefore, to ensure privacy on the data given to educational social robots, local LLMs should be used. Many open-source, local models exist, but running them can be complicated for people that do not have a technical background, such as often those in educational sectors. Therefore, the development of social robots for education that use generative AI should include providing easily accessible ways to run local LLMs, to ensure the privacy of students.

5.5 Ethical problems

If generative AI is used in educational robots, we can make them more human-like and social than ever. We believe this can bring great benefits to students' learning, but it also carries risks. The more socially human-like a conversational system or robot is, the higher the chance that students form attachments to it (Huber et al., 2016). This can be dangerous for their development of social skills, as it may start replacing social contact with peers. Especially for younger students, attachment may pose a problem if the robot then breaks, is removed or replaced or the student loses access to it for other reasons. It is important that it is clear for students at all times that the educational robot is a tool for learning, not a person. On the other hand, if students know that the robot is not human, which means there are no consequences to misbehaving towards it, this may lead to the normalisation of negative behaviour towards the robot, that later transfers to peers or teachers as well (Nomura et al., 2015).

In current conversational systems using generative AI, there is already a risk of spreading false information (Chen & Shu, 2023). Put such a system into a role with authority on their education – such as a tutor, a teacher's assistant or worse, a teacher – and students will not doubt its factuality. This is amplified by the appearance of reasoning in LLMs, which can take away doubt if it does arise. Additionally, LLMs tend to agree with whatever (false) information is suggested to them, which carries the risk of becoming an echo chamber (Nehring et al., 2024). Especially for young, vulnerable people, this can quickly become damaging, and should be avoided at all costs in educational environments.

As governments and other regulatory bodies are exploring how to regulate AI in order to minimize the risks the technology poses to citizens and society, education is seen as an application domain of special importance. As an example, in the European Union's AI Act, education is identified as a high-risk application domain, meaning AI systems deployed in education should comply to quality, safety, and transparency requirements, and could be subjected to human oversight. In particular, using emotion recognition in educational institutions is classified as an *unacceptable risk* and will be prohibited. It is yet unclear what emotion recognition entails exactly, but care should be taken to ensure that such regulation does not limit opportunities for making socially interactive systems more adaptive, while ensuring that systems that would track students' emotions at scale remain banned. Beyond education as a specific application area, general-purpose AI models, such as some LLMs discussed above, will undergo additional scrutiny, and have to comply to transparency requirements (Artificial Intelligence Act, 2024).

6. Conclusion

Today, social robots are not yet present in our classrooms. New, rapidly changing technology might change this in the near future. While educational robotics was limited by scripted interactions, failing speech recognition and only rigid options for adaptation to students, a revolution in generative AI technology has occurred, that might just change everything.

In this chapter, we described our vision of the future that will follow from this revolution. We started by investigating the aforementioned limitations of educational robotics and their causes. What followed, was an introduction to the rapid changes in generative AI technologies that caused what we call a technological revolution. These came together in the main section of this chapter, where we described in detail our future vision of educational robotics: how will this technological revolution solve the current problems, and what is the potential impact of this on our classrooms.

In this vision, the content that is provided in lessons will not be the same for every student, as there will be room for content generation on the fly. This can range from interactive explanations by a robot tutor, to illustrations of the stories that students are making up then and there. Generative AI will allow tutoring systems to adapt their difficulty level to each student – in a much more personal, fine-grained, and generalizable way than what is done today. Robots will manage to engage students even more by finding out what interests them specifically and use this to package content, so they want to know more about it. And as social robots are embodied, they share the physical world with us, which we can use to our advantage – from recognising when a student has lost interest, to involving the world around them in conversation.

Of course, this technology is still in rapid development. It allows us to dream big, but as is always the case with new technology, it also introduces new issues. Some of these issues are technological challenges waiting to be solved. As technology evolves, generative models' inference speed will decrease, allowing shorter generation times. It is reasonable to assume that the progression of AI technologies will allow for real time, natural flowing conversation with educational robots. The constraining of generative AI to be trustworthy is issue that asks for future research, with promising techniques arising rapidly. The sustainability of AI models is a problem that cannot be left unsolved, and future research should take this into account every step of the way, by minimizing impact where possible and ensuring transparency of the results. The privacy of children in schools should be preserved, by either demanding transparency from commercial partners, or avoiding them by running AI models locally as much as possible and keeping full control over the involved data. Finally, the position of educational robots and, more generally, generative AI in our school should be chosen with care and communicated clearly to everyone involved, including those who might still be too young to grasp the possible dangers.

It is clear that there is ample room for future research in the field of generative AI-driven educational robotics. On the one hand, the current limitations require new advancements, and existing technologies require care in their application to a field as critical and vulnerable as education. On the other hand, the revolution in generative AI opened many doors that have not yet been taken. The dream described in this chapter has not yet been realized, although it might already be possible. Only time will tell if the future of social robots in education will change as we described, and if Asimov's dream of mechanical teachers will come true.

Bibliography

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, pp. 1-14.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*.

Amioka, S., Janssens, R., Wolfert, P., Ren, Q., Pinto Bernal, M. J., & Belpaeme, T. (2023). Limitations of Audiovisual Speech on Robots for Second Language Pronunciation Learning. *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 359–367). Stockholm, Sweden: Association for Computing Machinery.

Artificial Intelligence Act. (2024). European Parliament and Council.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4945-4949). IEEE.

Bara, F., & Kaminski, G. (2019). Holding a real object during encoding helps the learning of foreign vocabulary. *Acta Psychologica*, *196*, 26-32.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: a review. *Science Robotics*.

Binderkrantz, A. S., & Bisgaard, M. (2024). A gender affinity effect: the role of gender in teaching evaluations at a Danish university. *Higher Education*, pp. 591-610.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., . . . Ramesh, A. (2024). Video generation models as world simulators. Retrieved from https://openai.com/research/video-generation-models-as-world-simulators

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems.*

Cartwright, O., Dunbar, H., & Radcliffe, T. (2024). Evaluating privacy compliance in commercial large language models-chatgpt, claude, and gemini.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's educational theory in cultural context*, pp. 39-64.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend and spell. *arXiv preprint arXiv:1508.01211.*

Chen, C., & Shu, K. (2023). Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*.

Chevalier, M., Riedo, F., & Mondada, F. (2016). Pedagogical uses of thymio II: How do teachers perceive educational robots in formal education?. *IEEE Robotics & Automation Magazine*, *23*(2), 16-23.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715-4729.

Cumbal, R., Lopes, J., & Engwall, O. (2020). Detection of listener uncertainty in robot-led second language conversation practice. *Proceedings of the 2020 International Conference on Multimodal Interaction*, (pp. 625-629).

Dagilienė, I. (2012). Translation as a learning method in English language teaching. *Studies about languages*, pp. 124-129.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., . . . Batra, D. (2017). Visual dialog. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 326-335).

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

D'Mello, S., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, *10*(2), 174-183.

Ellis, N. C., & Wulff, S. (2020). Usage-based approaches to L2 acquisition. In G. D. Bill Vanpatten, *Theories in second language acquisition, an introduction* (pp. 63-82). New York: Routledge.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 111-126.

Flook, R., Shrinah, A., Wijnen, L., Eder, K., Melhuish, C., & Lemaignan, S. (2019). On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy? *Interaction studies*, 455-486.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. a. (2022). Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends in Computer Graphics and Vision*, 163-352.

Ghosh, A., & Fossas, G. (2022). Can there be art without an artist? *arXiv preprint arXiv:2209.07667*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of pragmatics*, 1489-1522.

Gunes, H., & Churamani, N. (2023). Affective Computing for Human-Robot Interaction Research: Four Critical Lessons for the Hitchhiker. *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1565-1572). IEEE.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 335-346.

Henkemans, O. A., Bierman, B. P., Janssen, J., Neerincx, M. A., Looije, R., van der Bosch, H., & van der Giessen, J. A. (2013). Using a robot to personalise health education for children with diabetes type 1: A pilot study. *Patient education and counseling*, 174-181.

Houghton, H. A., & Willows, D. M. (1987). *The psychology of illustration* (Vol. 1). Springer.

Huber, A., Weiss, A., & Rauhala, M. (2016). The ethical risk of attachment how to identify, investigate and predict potential ethical risks in the development of social companion robots. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 367-374). IEEE.

Irfan, B., Kuoppamäki, S.-M., & Skantze, G. (2023). Between reality and delusion: challenges of applying large language models to companion robots for open-domain dialogues with older adults.

Janssens, R. (2024). Multi-modal Language Models for Human-Robot Interaction. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 109-111). Boulder, CO, USA: Association for Computing Machinery.

Janssens, R., Verhelst, E., Abbo, G. A., Ren, Q., Bernal, M. J. P., & Belpaeme, T. (2024a). Child speech recognition in human-robot interaction: Problem solved?. In *International Conference on Social Robotics* (pp. 476-486). Singapore: Springer Nature Singapore.

Janssens, R., Wolfert, P., Demeester, T., & Belpaeme, T. (2024b). Integrating Visual Context into Language Models for Situated Social Conversation Starters. *IEEE Transactions on Affective Computing*, 1-14.

Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, pp. 393-404.

Kasneci, E., Se{\ss}ler, K., K{\"u}chemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . H{\"u}llermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., . . . Irani, M. (2023). Imagic: Text-based real image editing with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 6007-6017).

Kennedy, J., Kumar, N., & Paetzel-Prüsmann, M. (2024). Name Pronunciation Extraction and Reuse in Human-Robot Conversations. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 593–597). Boulder, CO, USA: Association for Computing Machinery.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., . . . Belpaeme, T. (2017). Child speech recognition in human-robot interaction: evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 82-90). ACM/IEEE.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., . . . Shamma, D. A. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 32-73.

Krumhuber, E. G., Kappas, A., & Manstead, A. S. (2013). Effects of dynamic as pects of facial expressions: A review. *Emotion Review*, *5*(1), 41-46.

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., & Alami, R. (2017). Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence*, 45-69.

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Rockt{\"a}schel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, pp. 9459-9474.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning* (pp. 19730-19742). PMLR.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Computer Vision--ECCV 2014: 13th European Conference, Zurich, Switzerland,* *September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Zurich, Switzerland: Springer.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in Neural Information Processing Systems.*

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023). GPT understands, too. *AI Open*.

Mehrabian, A. (1972). *Nonverbal Communication.* Transaction Publishers.

Mesquita, B., & Boiger, M. (2014). Emotions in context: A sociodynamic model of emotions. *Emotion Review*, *6*(4), 298-302.

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. In *FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals* (pp. 387-409). Springer.

Mumm, J., & Mutlu, B. (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction. *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 331–338). Lausanne, Switzerland: Association for Computing Machinery.

Nehring, J., Gabryszak, A., Jürgens, P., Burchardt, A., Schaffer, S., Spielkamp, M., & Stark, B. (2024). Large Language Models Are Echo Chambers. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, (pp. 10117-10123).

Nomura, T., Uratani, T., Kanda, T., Matsumoto, K., Kidokoro, H., Suehiro, Y., & Yamada, S. (2015). Why do children abuse robots? *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts*, (pp. 63-64).

OpenAI. (2022). *GPT-4 Technical Report.*

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., M{\"u}ller, J., . . . Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International conference on machine learning.* PMLR.

Randall, N. (2019). A survey of robot-assisted language learning (RALL). *ACM Transactions on Human-Robot Interaction (THRI)*, pp. 1-36.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. *International conference on machine learning* (pp. 1060-1069). PMLR.

Reimann, M. M., Kunneman, F. A., Oertel, C., & Hindriks, K. V. (2024). A survey on dialogue management in human-robot interaction. *ACM Transactions on Human-Robot Interaction*, 13(2), 1-22.

Ren, Q., Hou, Y., Botteldooren, D., & Belpaeme, T. (2023). Behavioural Models of Risk-Taking in Human–Robot Tactile Interactions. *Sensors*, 4786.

Schutz, P. A., Hong, J. Y., Cross, D. I., & Osbon, J. N. (2006). Reflections on investigating emotion in educational activity settings. *Educational psychology review*, *18*, 343-360.

Senft, E., Lemaignan, S., Baxter, P. E., Bartlett, M., & Belpaeme, T. (2019). Teaching robots social autonomy from in situ human guidance. *Science Robotics*, *4*(35), eaat1186.

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*.

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI conference on artificial intelligence*, (pp. 13693-13696).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*.

Verhelst, E., Janssens, R., Demeester, T., & Belpaeme, T. (2024). Adaptive Second Language Tutoring Using Generative AI and a Social Robot. *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 1080-1084).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems.*

Wolfert, P., De Gersem, L., Janssens, R., & Belpaeme, T. (2024). Multi-modal Language Learning: Explorations on learning Japanese Vocabulary. *1129-1133* (pp. Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction). Boulder, CO, USA: Association for Computing Machinery.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., . . . Chen, Y. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.