A teacher's guide to ChatGPT and remote assessments

FernUni.ch
UniDistance.ch

01

What you need to know about ChatGPT and other Large Language Models (LLMs)

According to ChatGPT's own explanation:

> ChatGPT is a large language model that uses a deep neural network to generate humanlike text. The model is trained on a dataset of text, such as books, articles, and websites, and uses this training data to learn patterns and structure of language. During the generation process, the model takes in a prompt, or a piece of text as input, and uses the patterns it learned during training to generate a response that is similar to the input. The model can generate a wide variety of text, including answering questions, writing stories, and carrying out conversation.
>
> (generated by chat.openai.com on 22.01.2023)

A text generated by ChatGPT is based on probabilities (what is the word that is most likely to come next given the dataset it was trained on). It does not use logic, which explains why discussions sometimes appear nonsensical. The dataset it was trained on goes up to 2022, it currently does not have access to the Internet, and it is unclear whether it had access to research journals (it has given us contradicting responses to the question).

The output was supervised by humans using Reinforcement Learning from Human Feedback (RLHF). Humans wrote models of possible dialogues that were fed to the programme, and then newly AI-generated texts were ranked by human trainers. This taught the chatbot to answer in ways preferred by the humans chosen to train the model, as well as teaching it not to answer offensive or dangerous questions (although it is still quite easy to get around these safeguards).

We have not been able to find any data on how often responses were correct or not. However, many academic users seem to share a similar experience: the first answer seems surprisingly good, but as teachers try out more advanced questions in their areas of expertise, AI-generated answers remain superficial or are no longer correct. The issue for non-specialists is that, according to Openai's own blog:

> ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.
>
> (https://openai.com/blog/chatgpt/)

The issue of the validity of ChatGPT output is central. One factor is the quantity and quality of data used to train the model and the choice of prompts. When using the Internet, we are now used to looking for quality sources. In the case of AI chatbots, the sources are unknown and information cannot be linked back to a specific source. Equally problematic, LLMs sometimes produce hallucinations, that is, texts that are stated with certainty but are nonsensical or not based on the information from the training text set. This also raises epistemological issues about truth, and the different understandings of facts and the world. ChatGPT repeats what has been most shared, combines elements that may not belong together, and the training dataset is dominated by anglosaxon culture.

It is essential that students are made aware of the fact that despite sounding highly academic, the output may be wrong. ChatGPT has no understanding of what it is writing, of the world in which it works, nor is it ethical. ChatGPT has made up entire bibliographies that looked plausible and described experiments experts in the field could not trace in the literature.

For an expert in the field, ChatGPT can help summarize a text, compare two concepts, write code, and much more, but only if the human knows enough to write good prompts, make a judgement about the quality of the output and never assumes that something is true if the only source is ChatGPT (Will, 2023). We need to work together with students to learn how to use it to benefit their learning and understand its shortcomings.

What does this mean for assessments?

- AI can write well-written texts in all different styles, better than most students, within seconds.
- It is currently impossible to detect reliably whether a text was written by AI or a human. Although companies are trying to develop new tools, it will be an unwinnable arms race, especially as it is always possible to ask AI to write so that it will not be detected. Moreover, most students will not simply copypast, but use the output as one source amongst others. How to reference the use of AI is not yet established.
- The use of AI may help improve the level of student work, saving time on certain tasks (e.g.: structure, polishing the language, correct citation formats) to leave more time for indepth work in other areas (more careful critical analysis, for example).

02

Ideas for assessments you may want to try out in your modules

The way we assess students is strongly determined by tradition in the discipline and personal experience. However, are they the best way to foster the type of learning we wish for and students' curiosity and motivation (Forsyth, 2022)? ChatGPT is forcing us to reconsider our assessments. We should take the opportunity to make sure they are aligned with our purposes and values.

The suggestions below are relevant now, as we start learning to work with ChatGPT. Our approaches may change over the next few months when new insights and experiences become available. UniDistance Suisse will update you on a regular basis.

2.1   General considerations when redesigning an assessment

You may want to consider the following aspects:

1. What is the **purpose** of your module, the programme, and higher education more broadly? Review your intended learning outcomes. What should students be able to do by the end of your module? Why do they need to know these things and gain these skills? Should these be adapted for the world in which our students will be working 10 years from now? Are your assessments aligned with these goals?
2. What are important **values** for you as a teacher? Academic integrity, trust (needed for students to take risks, ask and answer questions, and accept feedback), fairness (all students have the opportunity to demonstrate their knowledge and skills), engagement and motivation, validity (is my assessment measuring what I want), surveillance (how much surveillance is acceptable in our society and university?). With the choices we make, some students will gain, and others lose. We need to pay attention and openly discuss these ethical questions (Cowen, 2022).
3. What is the **context**? Field of study, level, institution, type of students, professional expectations, etc.

General tips:
• Set **clear expectations** for the use of LLM tools in your syllabus: what can students use in their work, why, the need to reference their use or not, and own responsibility for the final content of the work (see Dr Christian Spannagel's example *here* or a model by Ryan Baker, *here*, discussed by Geroge Veletsionos *here*).
• Make exams **meaningful and motivating** to the students: choice (topic, specific case, but also in the ways of showing their knowledge), test higher cognitive skills, authentic tasks.
• You may want to ask for **shorter essays** so you can focus more closely on the argument, critical analysis, link to a specific context, check the references, etc.

- Offer a **combination of different types of exams.**
- Learning requires **trust** (to ask questions, take risks, accept feedback, for example). As educators, we need to focus on building this relationship through conversations in the classroom. Starting from a position of suspicion will make this difficult.
- Make sure the **time constraints** are flexible and encourage students to use the flexibility if they do not feel prepared. Cheating increases if students feel they are trapped. Offer extensions when needed (better than cheating).
- Some students tend to cheat when they feel they cannot do the task or they do not understand it. Make sure to take time to **explain the assignment and discuss it** with the students. See Amigud and Lancaster (2019) on reasons for cheating.

2.2   Ideas for assessments in a world with ChatGPT

2.2.1   Use ChatGPT in your assessments:
Rational:
Probably the most meaningful way to learn with our students about the possibilities and limitations of ChatGPT as an academic tool is to use it to answer assignment questions.

How:
Start by generating a ChatGPT output and then critiquing it in class (D'Agostino, 2023). Work with the students to see how well they identify shortcomings (lack of logic, biases, false statements, superficial arguments, made up references, etc.). Tip: save a ChatGPT output you would like to work with, each one is different, you might not want any surprises on the day of your class.

Option 1
Ask the students to use ChatGPT to answer a question you have asked or that they have come up with. It should be a challenging question without a clear answer. Then ask the students to do one or all of the following:

- Critically evaluate the response
- Add references
- Improve the outcome using 'track changes' to highlight their work. *Here* is a model of what such an assignment could look like.

Option 2
Prompt Competition (taken from *Ryan Watkins*)

- Identify a major question or challenge in your field or discipline that ChatGPT could write about. Preferably a question with no clear single right answer.
- Have students collaborate (in pairs or small teams) on developing 5 to 10 criteria for assessing ChatGPT responses to the major question. For example, ChatGPT's output references more than one theoretical perspective.
- Ask students to individually write a prompt for ChatGPT to answer the major question.
- Have students use their criteria to judge other students' responses (in pairs or small teams) and rate the ChatGPT prompts/responses from best to worst.

2.2.2   Dual assignments (taken from *Ryan Watkins)*

Let students choose between working with ChatGPT or without. Students who use AI should include their prompts and ChatpGPT's output, and use tracking to highlight their work. Students who decide not to use ChatGPT should sign a statement that AI was not used.

Grading will need to be adapted to each context. A whole class discussion about the quality of the assignments would most certainly be valuable.

2.2.3   Contextualise the assignment

Explicitly ask students to link their work to theories or cases used in class, or a new case which they need to research (a local company for example, the latest national data, or a recent research paper). If they use AI, they will need to know which theories or information to include in their questions and evaluate the answers given by the chatbot.

2.2.4   Assessment with multiple checkpoints

Build the assessment over the semester with assignments to be handed in regularly. Teachers can give feedforward toward the next step. This builds a closer relationship, which has been shown to reduce cheating, and the teacher gets to know the students and their work.

A research project could include defining the research topic and question, an annotated bibliography, a first outline or draft and final copy.

A weekly blog with individual feedback in the form of comments. The students can choose the topic of the blog, in relation to the material covered during the week, and relate it to their own experience or a current event, for example.

2.2.5   Process over product

Thinking about how one goes about a task helps develop metacognitive understanding and reflectivity. Ask the students to document and explain the process behind their work: what questions they asked themselves? What choices did they make and why?

2.2.6   Oral examinations

These may be carried out during the semester or at the end of the semester. Alternative formats (may also be combined): debates, presentations, role play, moot court, etc.

2.2.7   New media to share knowledge

Ask students to use different media to share their knowledge: video, podcast, multimedia presentation, infographic, mind map, website, poster presentation, etc. These offer alternative ways of showing one's knowledge and valuing students with different skill sets. Given the time it takes to master these tools, it may be good to suggest group work, and a choice of media. The marking criteria must be clear (e.g., marking grid).

2.2.8   What about assessing basic knowledge?

Rational:

Students need to memorize some facts to be able to carry out more complex tasks as they advance in their studies. However, there are clear issues with cheating when remotely assessing anything that can be 'Googled' or asked from ChatGPT. It is important to take this into account when writing MCQs.

Regular practice at recalling information (spaced retrieval practice) and mixing the topics (interleaving) throughout the semester has been proven to be more effective for memory retention than end-of-semester cramming (Brown et al., 2014).

How:

- Discuss with your students the importance of the respective basic knowledge for their studies in the long run.
- Make the assessments lowstakes and repetitive, throughout the semester. Weekly multiple-choice questions, mixing new and old concepts, with a minimum grade that needs to be achieved each week. The tests are required but only count for a small amount of the overall grade.
- Have an oral exam or ask 3–5 of these memorized facts during an oral exam later in the module or study program (tell the students at the beginning what they need to memorize and may be asked at any point during their studies).
- Have a time-limited MCQ test with randomized questions and order of answers. This is an option for learning outcomes that require memorization and automatic applications, not reflection or complex applications.

If you have tried something else that works in your class, please do share it with us henrietta.carbonel@unidistance.ch so we can add it to our suggestions. This is still very much work in progress.

03

References

Amigud, A., & Lancaster, T. (2019). 246 reasons to cheat: An analysis of students' reasons for seek-ing to outsource academic work. Computers & Education, 134, 98–107. https://doi.org/10.1016/j.compedu.2019.01.017

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). Make it stick: The science of successful learning. The Belknap Press of Harvard University Press.

Cowen, T. (2022, December 16). Who gains and loses from the new AI? Marginal Revolution. https://marginalrevolution.com/marginalrevolution/2022/12/who-gains-and-loses-from-the-new-ai.html

D'Agostino, S. (2023, January 12). ChatGPT advice academics can use now. Inside Higher Ed. https://www.insidehighered.com/news/2023/01/12/academic-experts-offer-advice-chatgpt

Fawns, T. (2022). An entangled pedagogy: Looking beyond the pedagogy—technology dichotomy. Postdigital Science and Education, 4(3), 711–728. https://doi.org/10.1007/s42438-022-00302-7

Forsyth, R. (2022). Confident assessment in higher education (1st ed.). Sage Publications.

Will, M. (2023, Month DD). With ChatGPT, teachers can plan lessons, write emails, and more. What's the catch? https://www.edweek.org/technology/with-chatgpt-teachers-can-plan-les-sons-write-emails-and-more-whats-the-catch/2023/01

FernUni.ch
UniDistance.ch