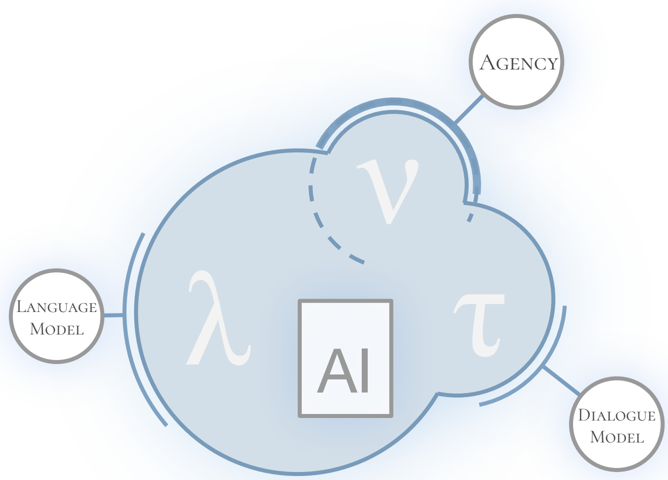
Perspectives on Generative AI

Extended Slide Notes
2023-07



Boris Steipe

boris.steipe@utoronto.ca · https://sentientsyllabus.substack.com

In these slide-notes, I discuss aspects of Generative AI *technology* that affect its use in practice and explore the resulting *abilities* of AI systems. I showcase examples of the surprising emergence of understanding, and explore the *consequences* for learning and teaching in a higher-education context.

The field is complex, not just complicated: all facets influence each other, and, mapping out the landscape in broad strokes still reflects that complexity.

My original presentation was part of a technical session at the Council of Graduate Schools (CGS) Summer Workshop, July 12. 2023 in Denver, Colorado.

Restructuring and interpreting the original presentation slides allowed me to flesh out some ideas that we discussed at the workshop: a principles-based approach to academic integrity, AI policies, and priorities for institutional support. Thus the notes cover more ground.

I am grateful to the CGS organizers: Suzanne Ortega, Heidi Shank, and Marlena Wolfgramm for the opportunity. Thank you to Cari Moorhead, University of New Hampshire, for the introduction in Denver. And a big thank to my co-presenter at the workshop: Betsy Barre, Wake Forest University.

The engagement, profound interest, and lively contributions of our workshop participants were inspiring.

Toronto, August 08. 2023

Boris Steipe

Generative AI – Concept Map

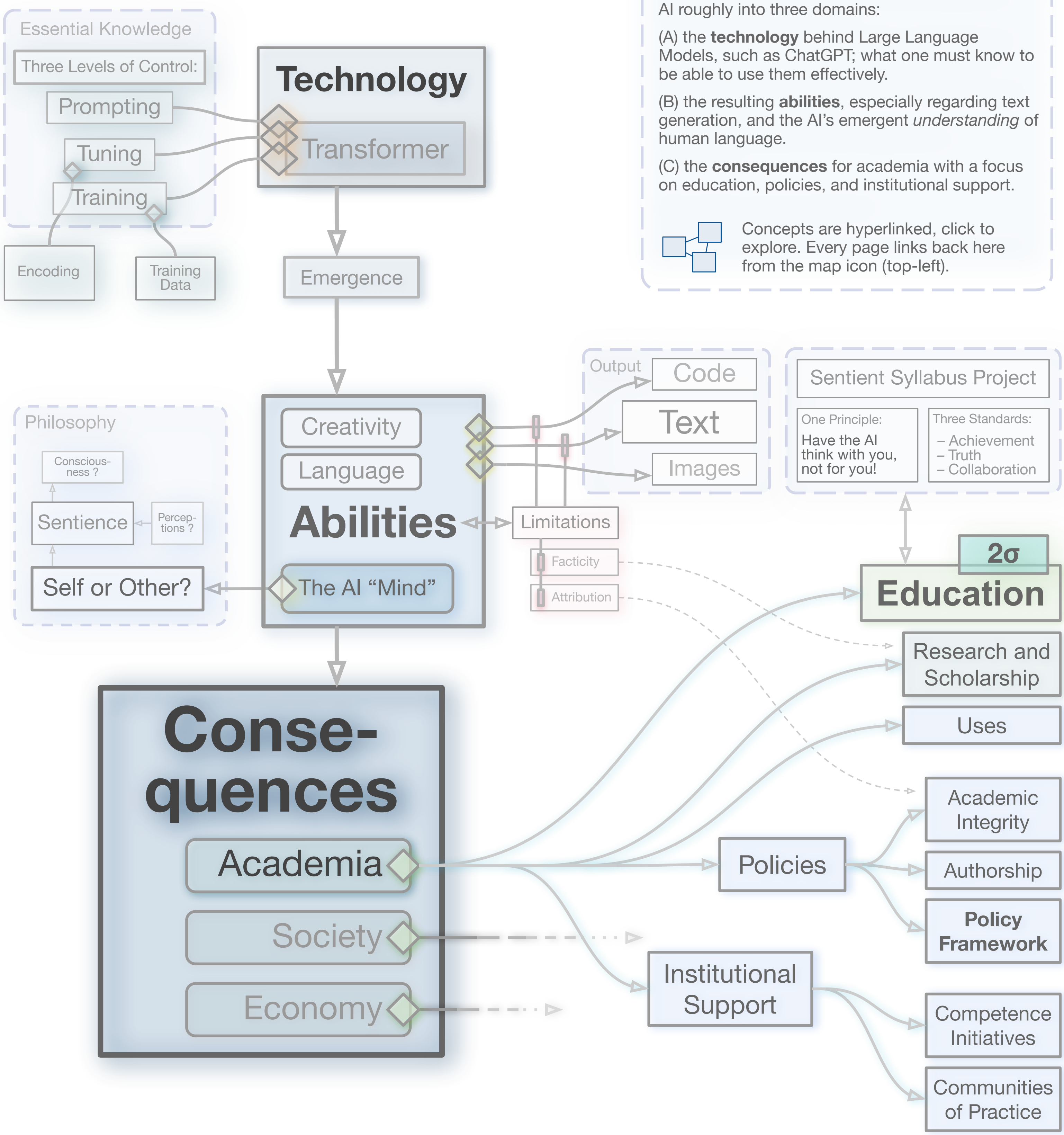
In this document, we divide the field of Generative AI roughly into three domains:

(A) the **technology** behind Large Language Models, such as ChatGPT; what one must know to be able to use them effectively.

(B) the resulting **abilities**, especially regarding text generation, and the AI's emergent *understanding* of human language.

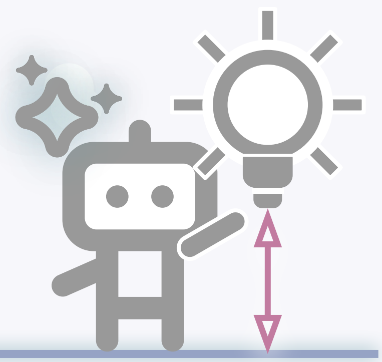
(C) the **consequences** for academia with a focus on education, policies, and institutional support.

Concepts are hyperlinked, click to explore. Every page links back here from the map icon (top-left).



Graduate Schools: Selected Themes

The Graduate Perspective



At the graduate level, human work must be able to surpass the AI, whatever the AI's abilities are.

Since AI abilities are continuously improving, *surpassing* the AI entails to make full use of the AI, and to "stand on its shoulders"...

Generative AI

What is it?

- Technology Foundations
- Available Systems

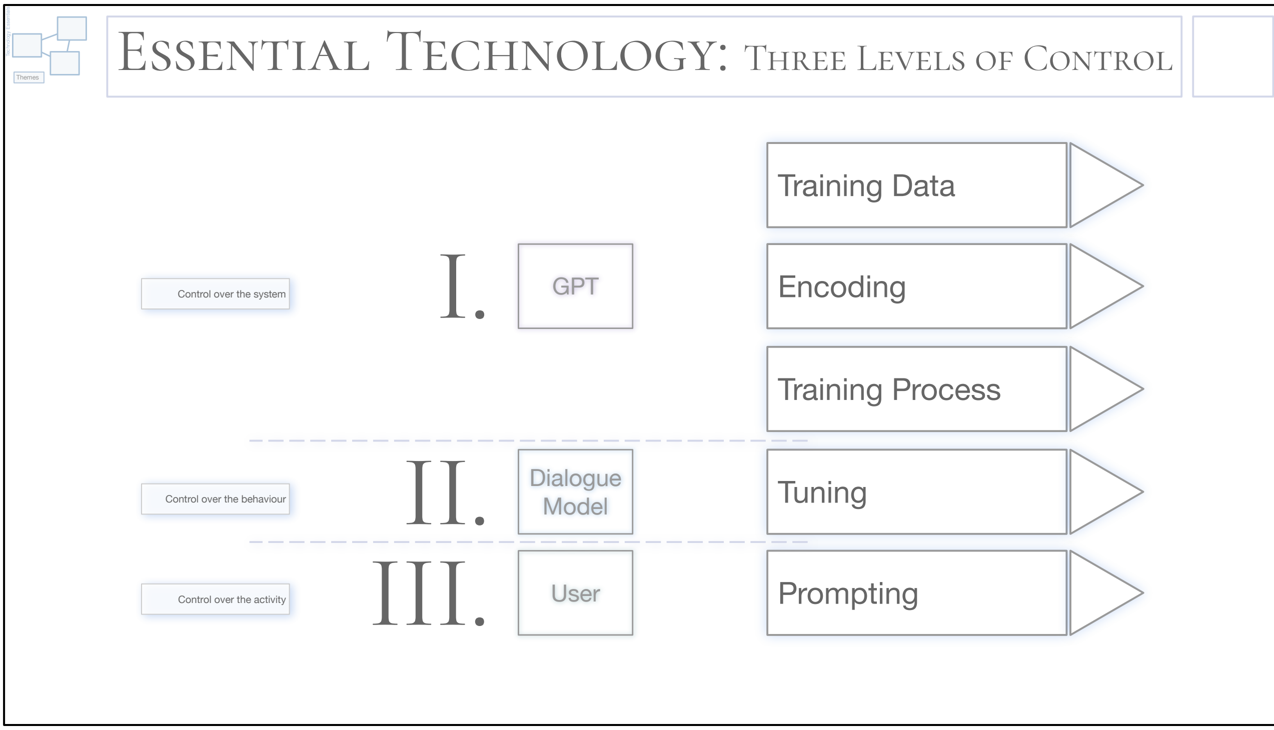
What can it do?

- Abilities
- Education
- Uses

AI and the Academy...

- AI Policies
- Academic Integrity
- AI Support
- AI Competence

Communities of Practice



What we need to know about the technology of Generative AI is determined by how we interact with the systems and what we expect them to do.

I. We are dealing with Large Language Models (LLM), specifically GPT systems – Generative Pre-trained Transformers. They *generate* output from input, by *transforming* an input prompt into an output string, and they are pre-trained for their tasks. Some aspects of the training are important in practice: what the *training data* is, how it gets into the computer, and what happens in the *training process*. This will helps us develop an intuition about such systems' abilities and limitations.

II. We need to understand the process of *tuning* – the extra training that builds useful dialogue abilities on top of the AI's general language abilities. This controls the parameters under which the language abilities of the GPT system are doing something useful.

III. We need to understand the proper use of user prompts. You can think of this *prompting* as the programming of the system: crafting the specific request that elicits a response from the AI. Prompts determines the type and quality of output that we obtain.

Training Data

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Brown et al. (2020) “Language Models are Few-Shot Learners”. *arXiv.* https://arxiv.org/abs/2005.14165

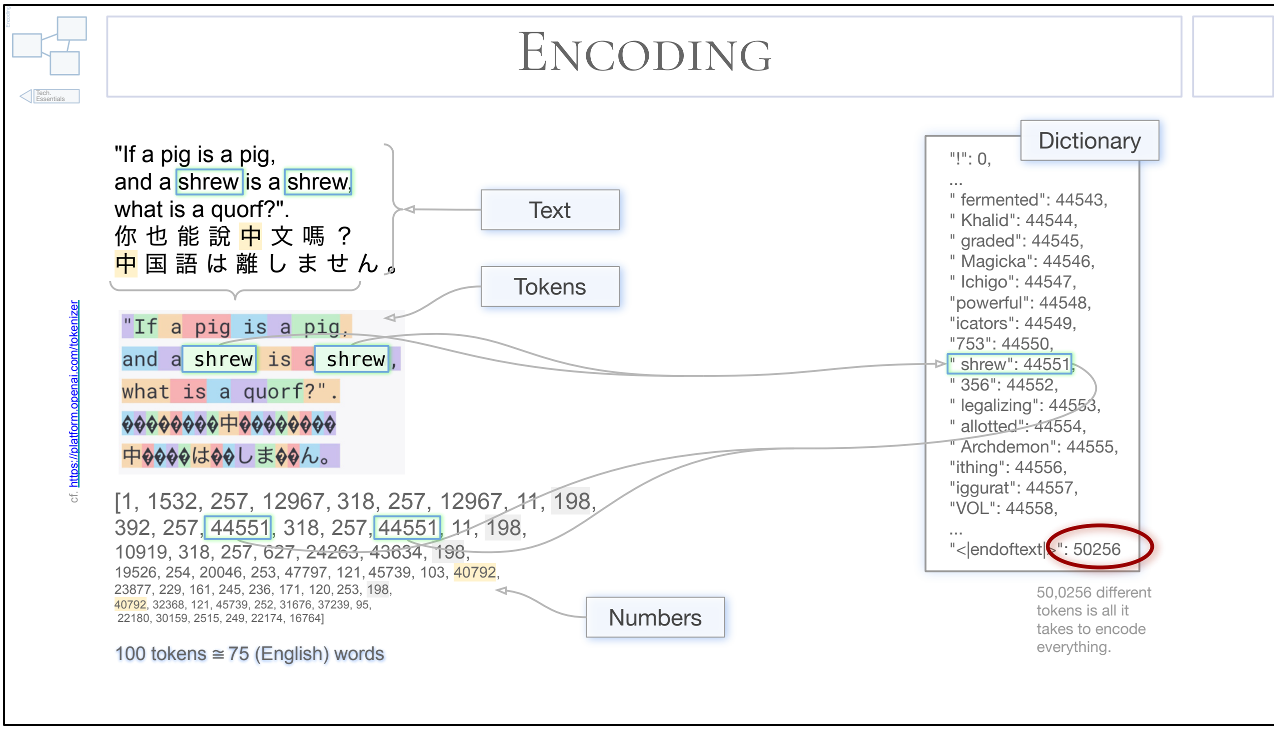
This is the foundational technical publication that describes OpenAI’s GPT-3 Language Model.

GPT-3

GPT-4: ?

The training data contains a significant fraction of human thought. Interestingly, performance of OpenAI’s ChatGPT models, Anthropic’s Claude, and Google’s Bard is roughly comparable – even though the details of training data are not identical. This is reassuring: the training process converges on relevant and universal principles. It underscores the fact that language models do not copy and store what they have been trained on, but learn an interpretation.

The numbers are taken from the technical publication of the GPT-3 model (Brown 2020). Details of GPT-4, OpenAI’s most recent, much more capable language model, have not been published.



Thinking of a GPT system as operating on words is a good approximation. In reality however, as with any computer, the system operates on numbers, and the numbers are related to words, or rather word-fragments – "tokens" – that are stored in a remarkably small dictionary. The dictionary contains 50,256 entries, which are used in a clever encoding scheme that can represent not only all the words of English and many other languages, but also a large variety of writing systems including the CJK scripts, Hindi, Inuktitut, and even Sumerian cuneiform. The sample text contains an English sentence, a line of Chinese and of Japanese with kanji and kana. I have highlighted the glyph "中" (middle) which is used in both Chinese (*zhòng*) and Japanese (*naka*) and mapped to its own codepoint in the dictionary: 40,792. Less frequently used characters are assembled from typically three components.

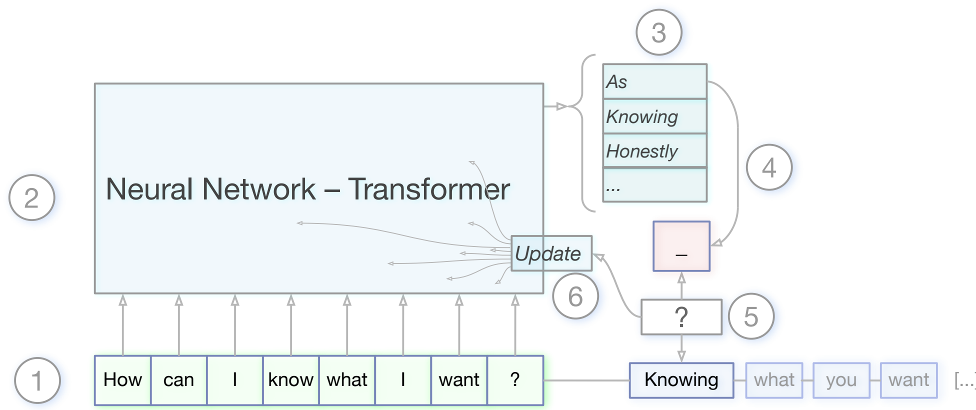
Text is the source material. Tokens are a structured interpretation of the text. Numbers is what the algorithm “sees”.

The excerpt from the dictionary shows entries around the string “ shrew”. Note that the entries do not have any particular ordering, most words include a leading blank space, or they are re-useable word fragments (e.g. “icators”).

The encoding system is efficient. It takes only about 100 tokens to represent 75 English words, including spaces, capitalization, and punctuation.

Note that the AI system learns language from the relationships between the tokens, not words. However, for simplicity the following slides will pretend that GPT systems work on words, merely as an approximation.

Training the GPT

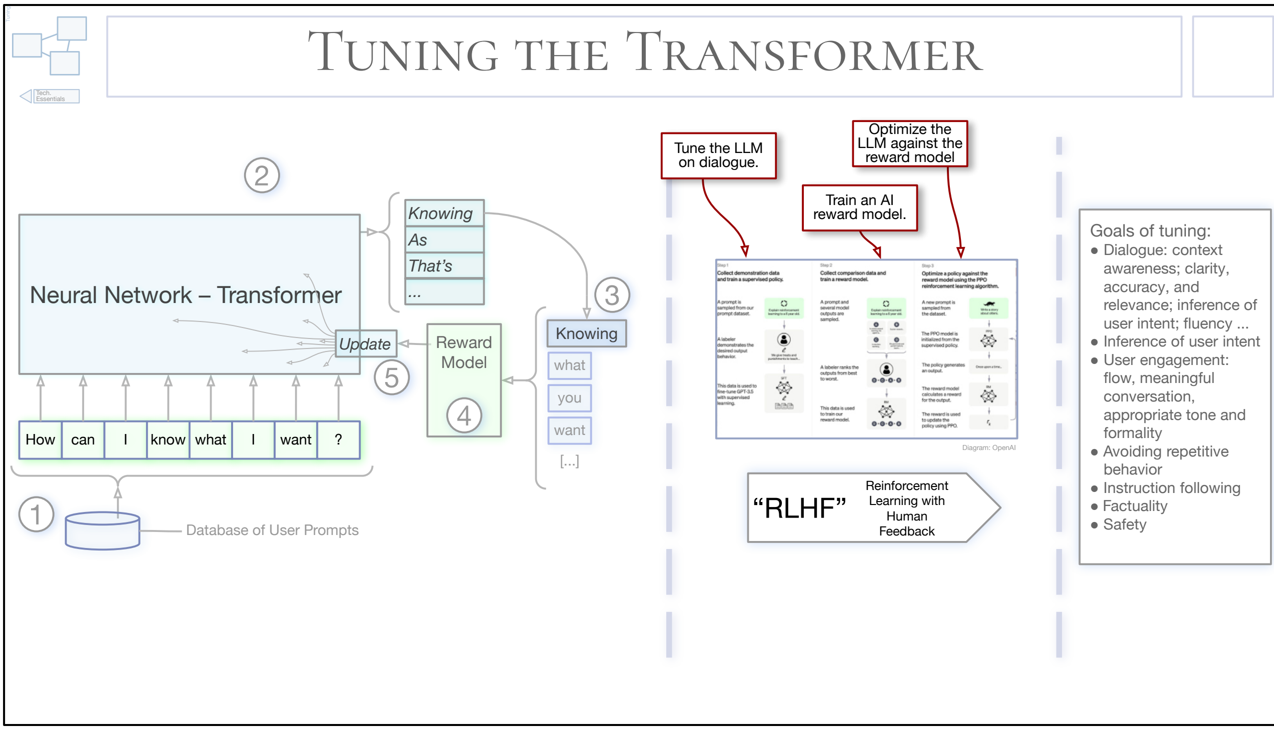


GPT: Generative Pre-trained Transformer. A type of large, computational model of language.

Training the Generative Pre-trained Transformer neural network consists of presenting examples of text from a very large corpus, word by word (or token by token; see previous slide).

1. First, a fragment of text is aligned with the input nodes of the network.
2. The "Transformer network" takes the entire input into account, weighting the individual words according to their importance ("attention"). It then computes relationships between the words, and relationships between the relationships, several layers deep. The number of relationships ("parameters") is very large – hundreds of billions.
3. As a "transformer", the network is configured to finally come up with a set of probabilities: every word it knows is given a probability that it could be the next word to continue the input message.
4. The high-probability words are then chosen ...
5. ... and compared with the true word – the word that actually continues the input in the training data. (In the example, that word is "Knowing".)
6. Typically, the network would make a mistake. The difference between the predicted word and the correct word is then computed and used to adjust the parameters of the network to give a better prediction next time.

This is repeated hundreds of billions of times, each time tweaking the network a little bit. As a result, the network becomes better and better at predicting the next word. It learns to produce meaningful sequences of words. But in order to do this very well, it needs to learn language.

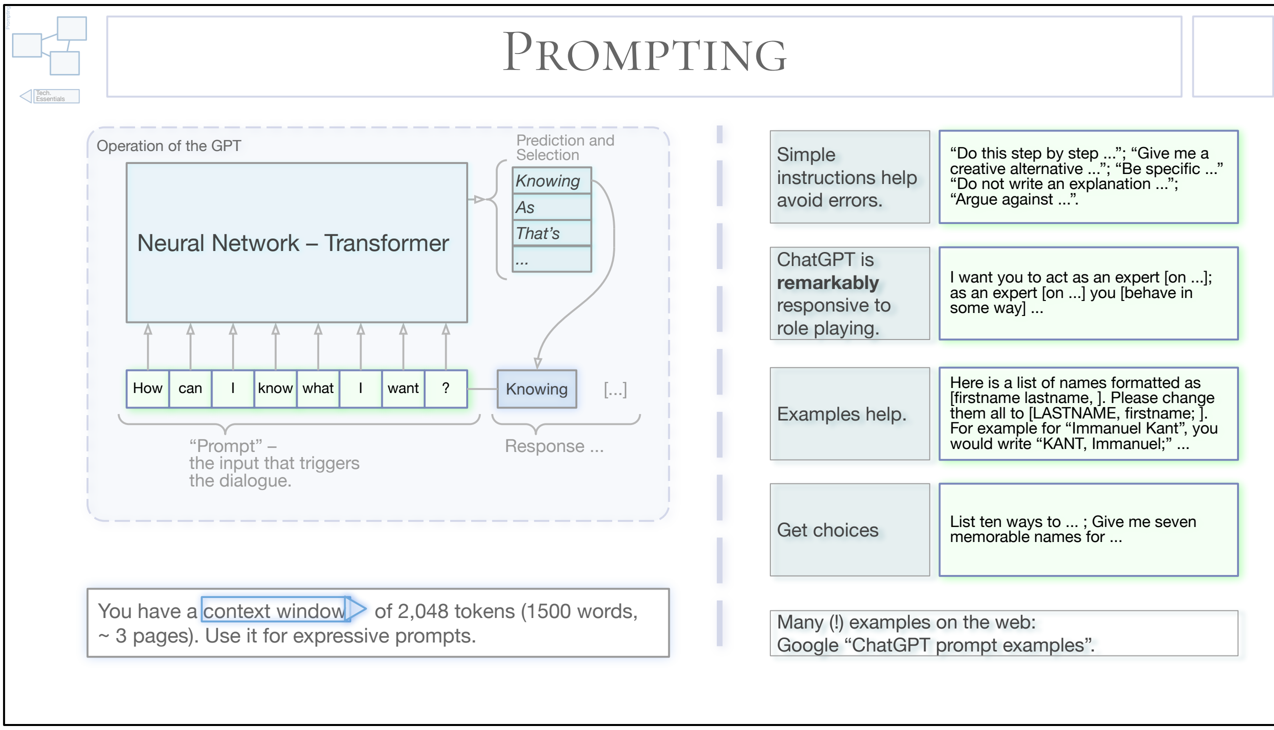


The GPT acquires generic language competence during training. But to make it specifically useful in a "chat" mode requires tuning the model for the task.

1. We start with a database of human-contributed user prompts, which we use as training data just like the training corpus. However, these are not random pieces of text, but dialogue that has been specifically written for the purpose of tuning the model.
2. The pretrained AI is tuned with this input, and it generates sequences of words in the usual way.
3. Several different outputs are generated for the same prompt. These outputs are scored by human evaluators.
4. The prompts and scored responses are used to build a different AI tool: a "reward model" that is able to recognize "good" responses.
5. The reward model scores the quality of a response, and this score is used to further improve the model. But now, human input is no longer needed. The language model can generate responses, evaluate them, and update itself to provide better and better responses. This process is called "reinforcement learning" and it is one of the ways modern machine learning is entering a phase of accelerated progress: machines are learning to improve themselves.

Tuning turns a machine with generic language competence into a conversational assistant.

(cf. OpenAI 2023. "Aligning Language Models to Follow Instructions". https://openai.com/research/instruction-following)



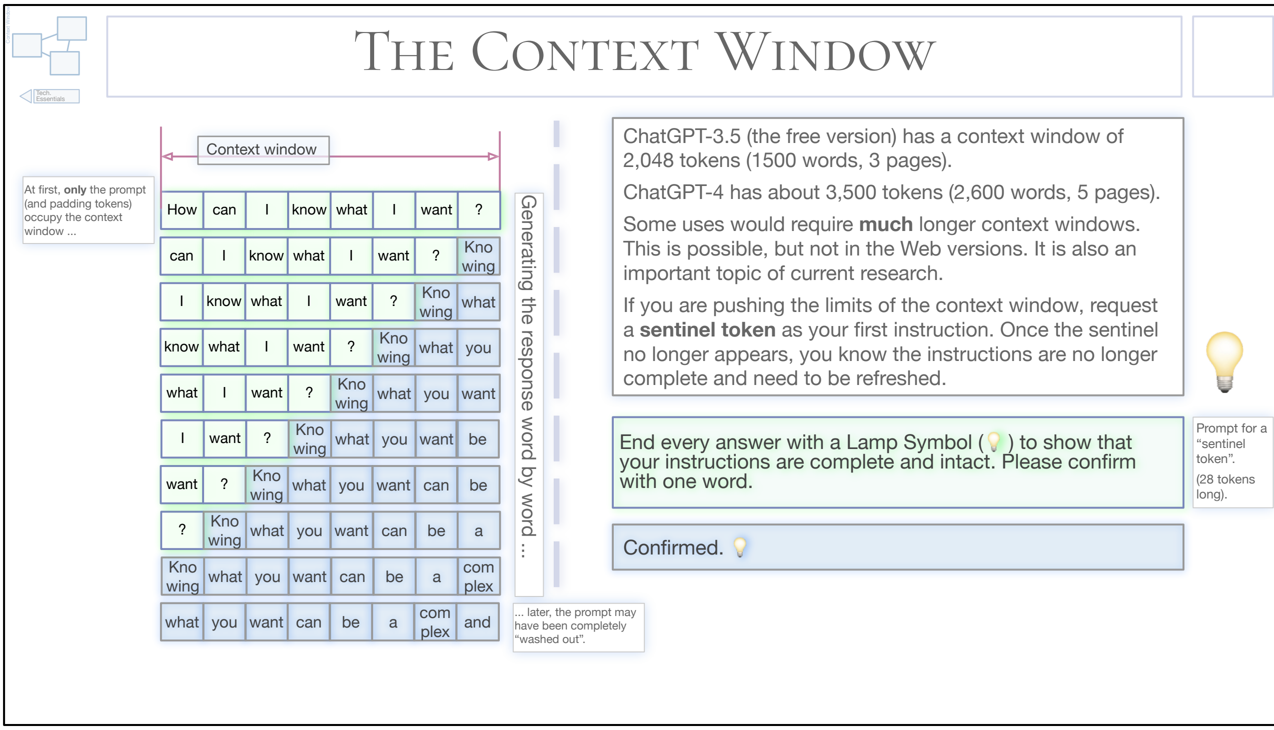
The pre-trained transformer is the device, the prompt is the programming. Good prompts are crucial for good results (when you have non-trivial requests.) Writing good prompts is a bit of an art.

The AI can draft and "improve" prompts – but do **not** assume that this is more effective than if **you** write the prompt. The AI does not introspect in that regard, it has no self-knowledge, it has very limited self-awareness, and to the best of our understanding no ability to consider and evaluate alternatives. (The idea that the AI would know best what works for the AI is a common fallacy.)

Effective prompt strategies include:

- **Few-shot learning**: train the AI from a small number of example input-output pairs as part of the prompt. (This is actually more about *correctly identifying the task* than it is training for performance.)
- **Chain-of-thought** prompting, (also: think step-by-step). Require the AI to develop a solution to a multistep problem by describing all intermediate steps explicitly.
- **Playbook prompting**: explain the steps and reasons of a multi-step problem solving or dialogue interaction ahead of time. "This is what we will do: ..." This may help to focus attention correctly.
- **Role playing**: Role-playing imports a personality and pattern of behaviour into the dialogue. Responses can be more perspectival, employ the right register, and use expert language. Greatest benefit: when the character traits of the desired persona are common cultural knowledge.
- Don't force the system. If you can't get a good answer right away, doubling down and insisting will not give you what you need either. Change your strategy, for example by providing some known facts to build upon.
- In particular, don't force judgment. Don't ask: "Is *a* better than *b*?"; rather ask: give me five advantages and five disadvantages of *a* relative to *b*.
- Be aware that the GPT system is easily biased by **your** opinions. Be very careful to phrase your questions as openly and neutrally as possible. Otherwise you just hear yourself talking.
- Use positive language. If your instructions are dominated by descriptions of *what not to do*, you may encounter the "Waluigi Effect" – a paradoxical inversion that causes the GPT system to do the opposite of what it was asked to do.

In all extended dialogues, be aware of the limited size of the context window.



The sequence of words that the Transformer model operates on at every step has a constant length. This length is the "context window" of the generation process. "Context" refers to the fact that these words – and only these – are taken into account when computing the prediction for the next word.

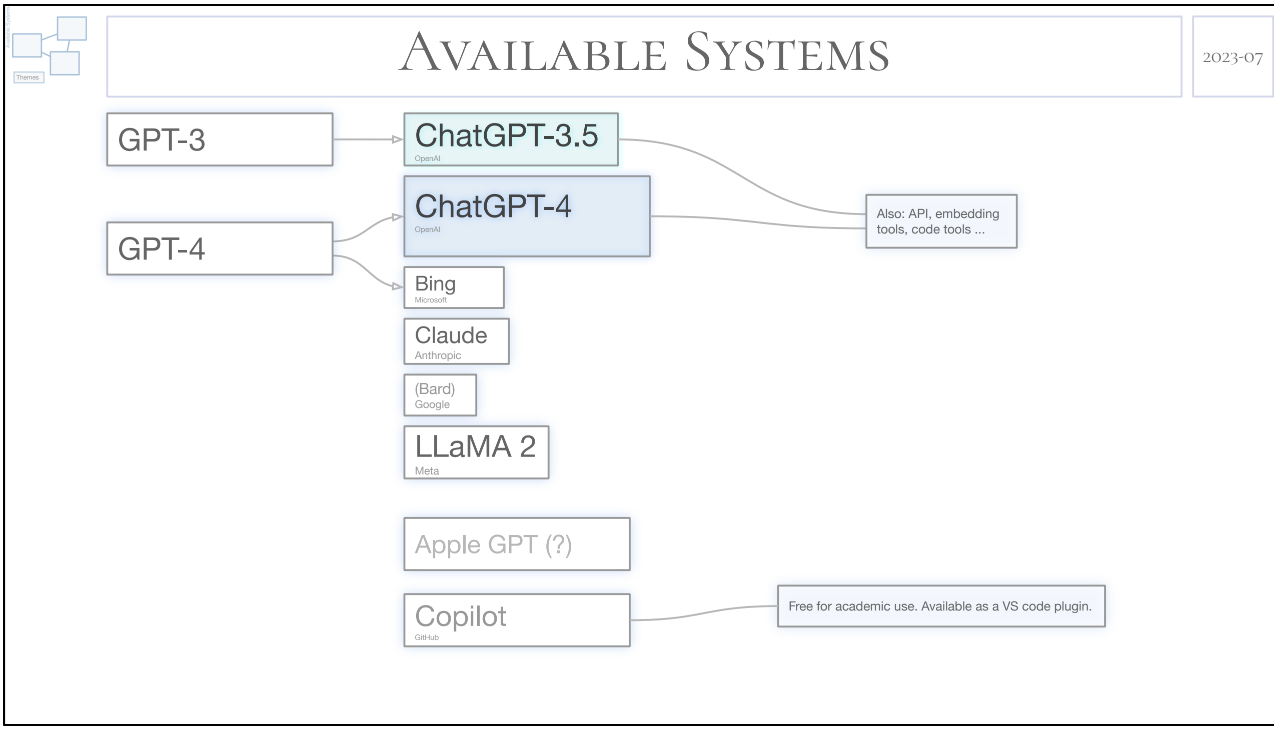
Every newly predicted word is added at the front of the context window – at every step. And once the context window is full, earlier parts of the conversation drop off the end.

This is important:

- longer conversations will have no memory of their beginning;
- instructions that are longer than the context window will not be fully acted on;
- longer documents will not be accurately summarized.

Once you understand that,

- you realize why discussions can go off topic;
- you make sure that instructions are concise, so that **both the instruction and the response** fully fit into the context window;
- you know to summarize longer documents in pieces – recursively if necessary;
- ...

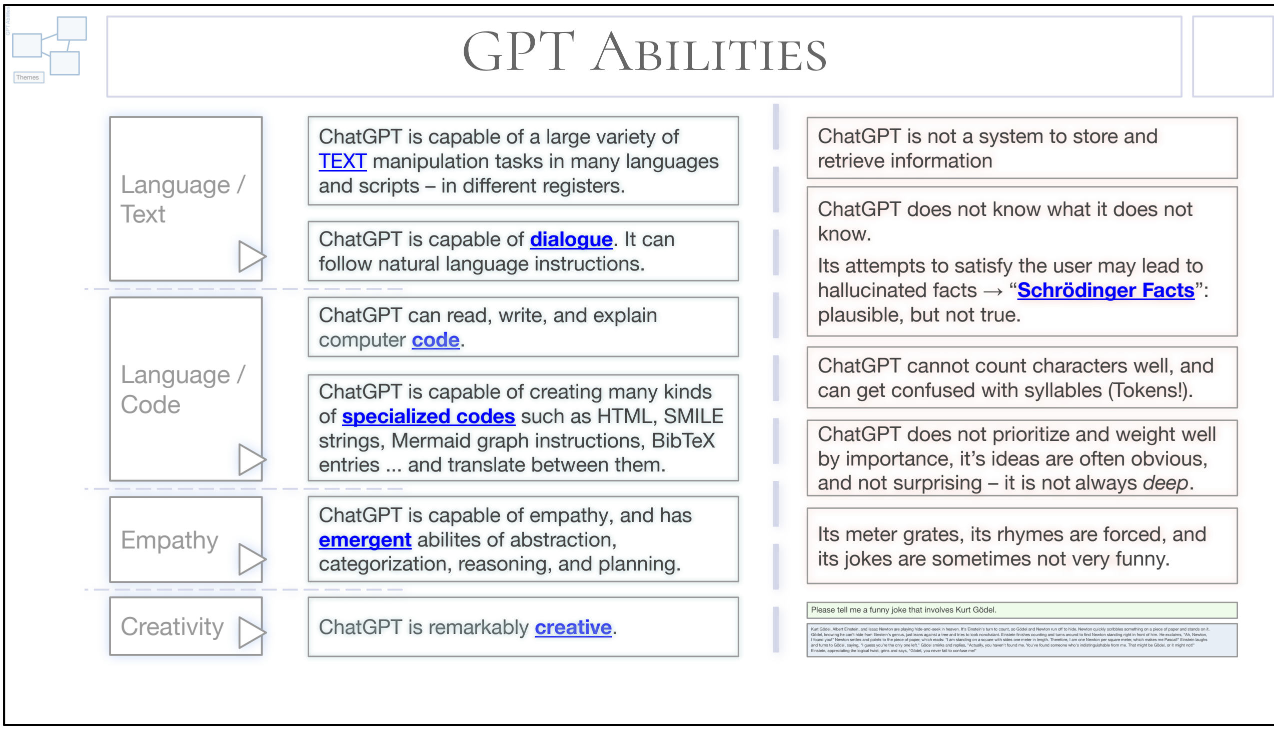


Not all Generative AI systems are available in practice.

- ChatGPT-3.5 has the fewest restrictions. It is free to use with an account, fast, and pretty good.
- ChatGPT-4 is OpenAI's stet of the art model. You need a paid account with OpenAI ($30/month). It is built on the GPT-4 Language Model and has remarkable broad knowledge and nuanced understanding.
- Bing is also built on top of GPT-4 and this shows the importance of getting the tuning right. Bing is not pleasant to use. It is free for users of the Microsoft Edge browser, and you can use it from a Skype chat.
- Claude sits somewhere between Bing and ChatGPT in terms of quality. The response quality does not quite match ChatGPT-4, but it is better than Bing and ChatGPT-3.5. It tends to be a bit loquatious. Available in a free, public beta.
- Google's Bard is not available in Canada. Requests to Google Support for evaluation access were not considered.
- Meta's LLama2 was just made publicly available as an open-source download. (2023-07-18 -> Hugging Face). You can access a demo at the Hugging Face site. Why is this significant? The LlaMA 2 model runs on Amazon Cloud computing services, and it should be relatively straightforward for a University IT infrastructure group to set up a dedicated instance, with appropriate privacy and security controls, should keep student's access costs low, and support experimentation with APIs and other interfaces. It seems to do quite well: probably better then ChatGPT-3.5 but not quite at the level of ChatGPT-4.
- AppleGPT has been hinted at recently.

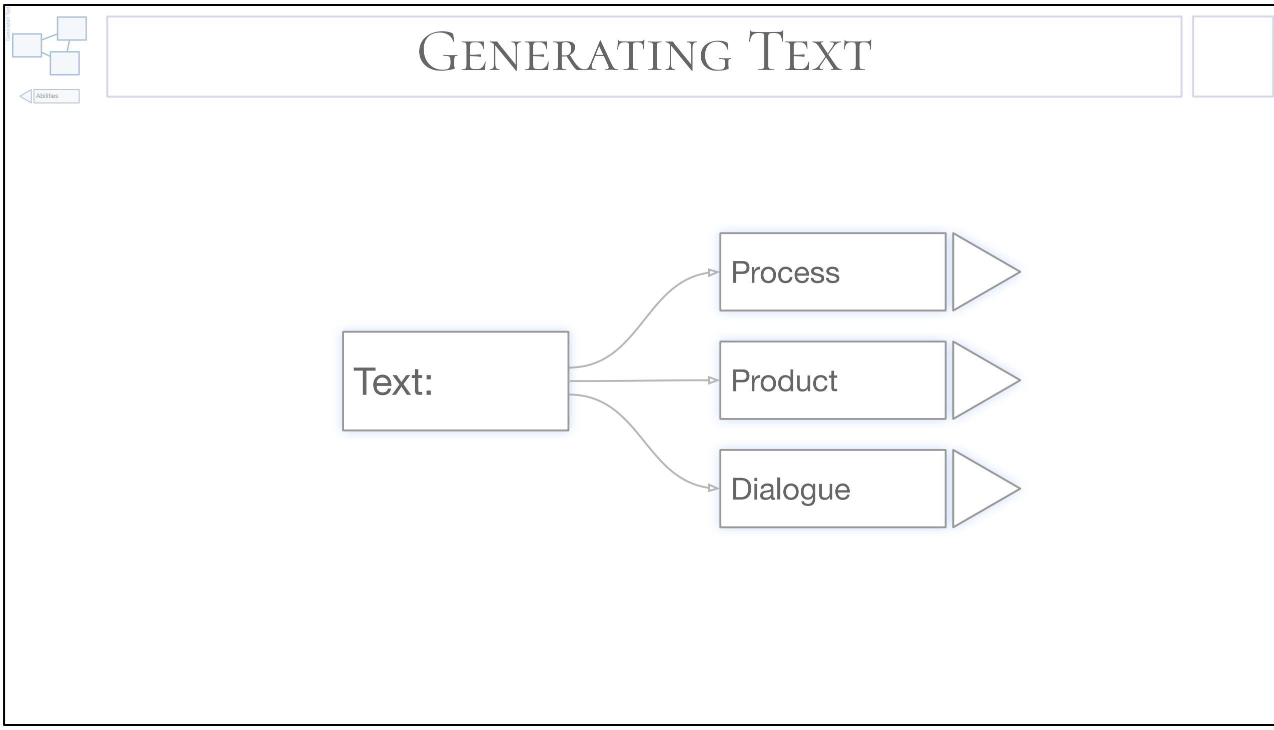
One other system to know about is "Copilot", trained specifically for computer programming tasks by Github on their massive source-code libraries. This is a game-changer for learning programming by our students. Free for academic use – either on GitHub or in your own VScode IDE.

The availability of commercial APIs has led to a flood of embedded systems: specific software applications have parts of their functionality powered by AI (think of Grammarly, Duolingo, ...). This is great on one hand, since it can increase the utility of the software tremendously. On the other hand, the use of the APIs is not free and has to be recovered from the user. This will accelerate the trend that you don't properly own the applications that run on your computer by purchasing them, rather you pay subscription fees for for ongoing usage. This may not be very onerous for each individual instance, but you may need a lot of applications, a lot of seats for the lab, and all this every month, over a long time. You end up paying far more than you used to. Significantly more actually – e.g. Adobe Creative Cloud academic licenses now cost about 10% per month of what the whole Creative Cloud suite used to cost as a permanent license, Thus by adding to our operating costs structure, this model adds significant risk into our unpredictable research grant funding landscape and it generates budgetary items that compete directly with human compensation.

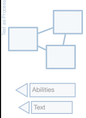


The most compelling aspect of ChatGPT and its relatives is their complete generality. They perform at a near-human level in very many disciplines and tasks, but they perform at a superhuman level in terms of their breadth of "knowledge" and range of abilities.

(I am putting "knowledge" in scare quotes because the knowledge of a GPT system is qualitatively quite different from our human knowledge. Whereas we rely a lot on memory – storage and retrieval, GPT systems work with context: relationships, and probabilities of association.)



There are (at least) three distinct types of objectives associated with AI generated text: text as process, text as product, and text as dialogue.



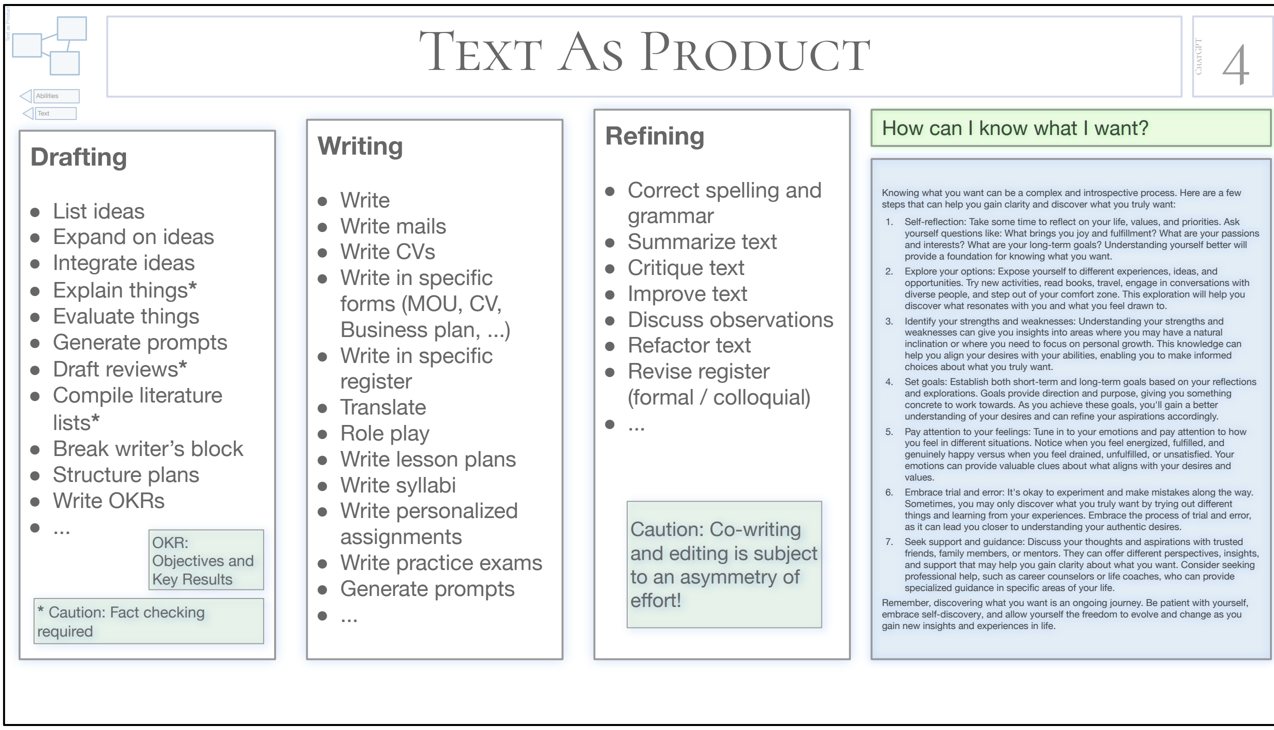
Text As Process

Instructing machines through natural language.

- TicTacToe (sort of)
- Hangman (sort of)
- Terminal
- Cryptic crossword
- Wordle
- BibTeX

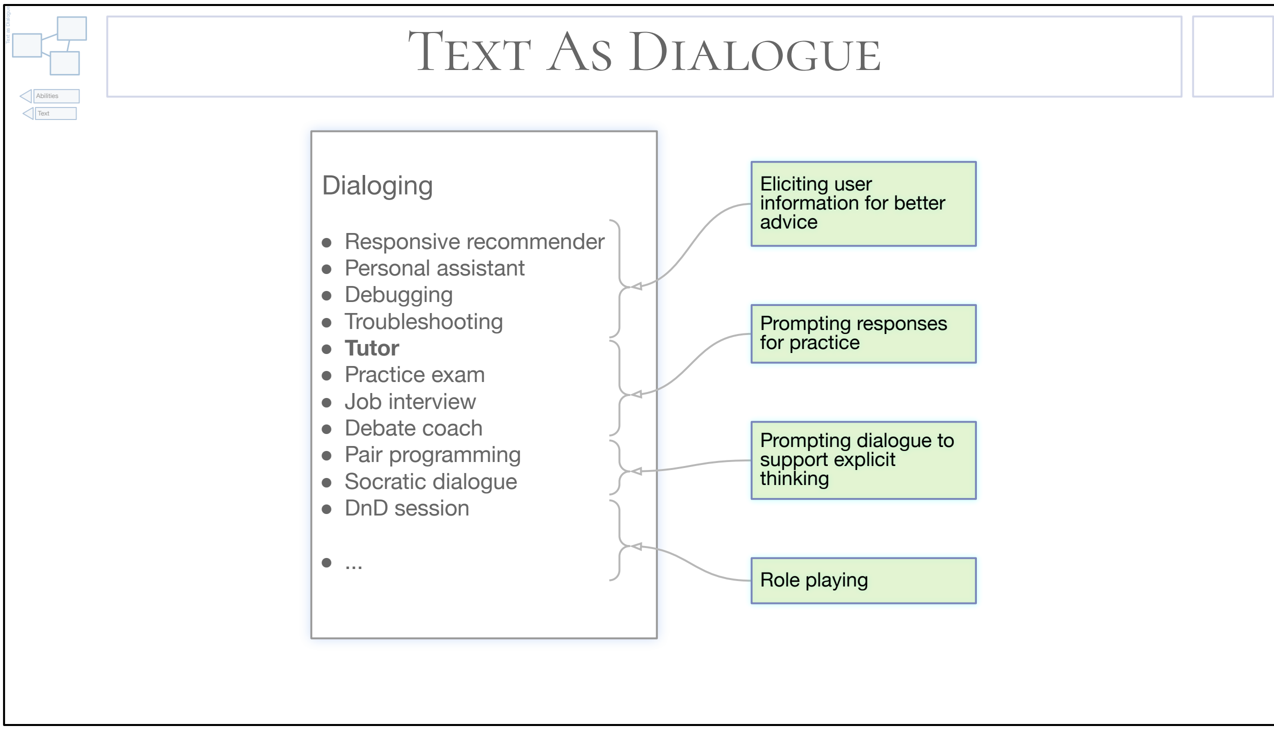
- Store
- Search and retrieve
- Categorize
- Find similars
- Semantic paths
- Transform
- Translate

"Text as Process" focusses on detailed instructions that achieve something other than what is contained in the text itself. This is a perspective in which the text has agency.



"Text as Product" focusses on the result of writing. The versatility of generated text is remarkable. The example demonstrates a complex and nuanced response to a very simple query. Had this been the topic of an academic assignment, few students could match this level and even fewer could surpass it. Given this reality, we have to question what value such assignments still have at all, and we need to create new ways to stimulate and assess students' progress. Text as product has lost much of its value in this respect, even if we force students to create it under (increasingly artificial) conditions that prevent AI input. To give just one example of an alternative approach: use the AI response as the starting point and ask for weaknesses, omissions, biases and ways to think about this question in a better way (e.g. by addressing the affective side of the question, rather than the epistemological side).

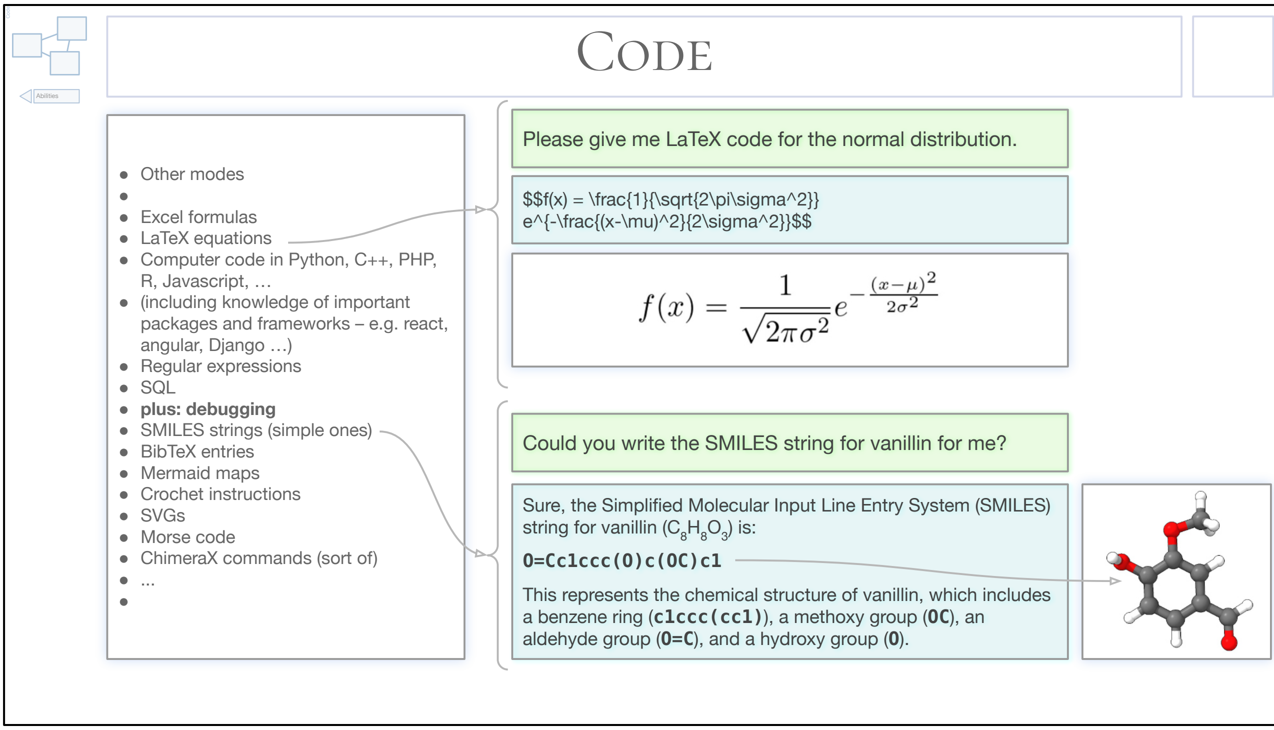
You need to be aware of an asymmetry of effort when co-editing with the AI. When we edit, we humans keep most of the original text intact. However, the AI writes everything from scratch, every time. It does not re-use words, idioms, expressions from iteration to iteration like we do. This means, the AI's ideas and phrasing can dominate the editing, even though the human's ideas are more valuable. Once you are aware of this, you can compensate.



"Text as Dialogue" focusses on the exchange of ideas through writing. This perspective focusses less on the result of the dialogue, and more on the relational nature of the exchange, on expressing ideas clearly, on convincing the other.

"Socratic dialogues" are especially valuable in education. They follow Socrates' model (as recounted by Plato) of question and answer, in order to have the student actively discover a "truth", rather than being told the truth outright. A complementary way to engage the AI is a "reverse socratic dialogue" – here the AI is the student, and the discourse is an exercise for the teacher to convince the AI of a viewpoint that it might not have previously. This is actually not uncommon – the nature of the training process causes the AI to hold views that reflect a kind of broad, common denominator, whereas we as scholars frequently need to question the validity and justification of such views. It is certainly possible to convince ChatGPT, although it is often not easy, but note that the memory of the dialogue does not persist beyond the instance of the dialogue.

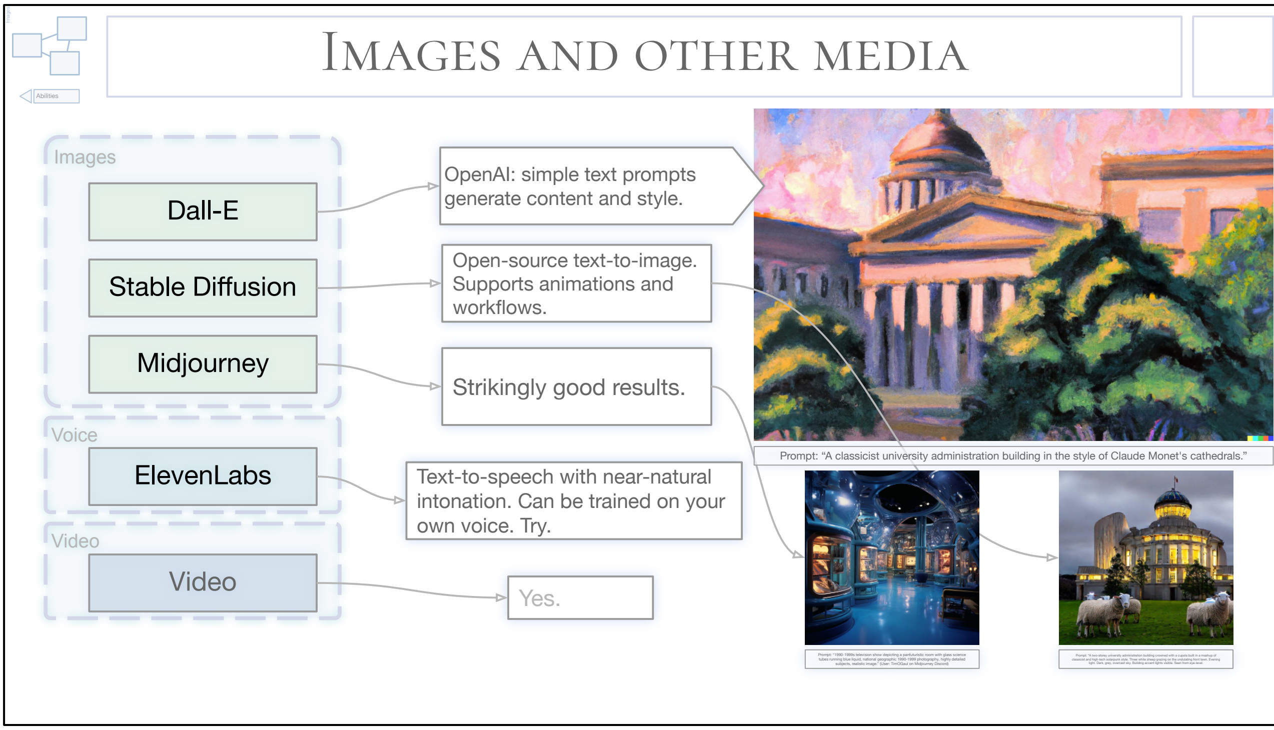
As far as education is concerned, by far the most important ability of GPT systems is to act as a **personal tutor**.



The ability to work with AI assisted coding requires us to completely change the way we approach teaching computational subjects. Students no longer need to master syntax. The algorithm writes syntactically correct code for them. But this is rarely excellent code:and we need to improve algorithmic elegance, apply sound engineering practice, and work from robust software engineering principles to bring it to acceptable standards. And, of course, **validating and testing** become even more essential to catch the potential subtle logic errors that arise when the algorithm translates ambiguous language into precise instructions.

The two example tasks were solved correctly without further required input (ChatGPT-4). The LaTeX code produces the correct formula; the details of vanillin are correct, and the SMILES string converts into the correct 3D-structural model when loaded into the ChimeraX molecular viewer.

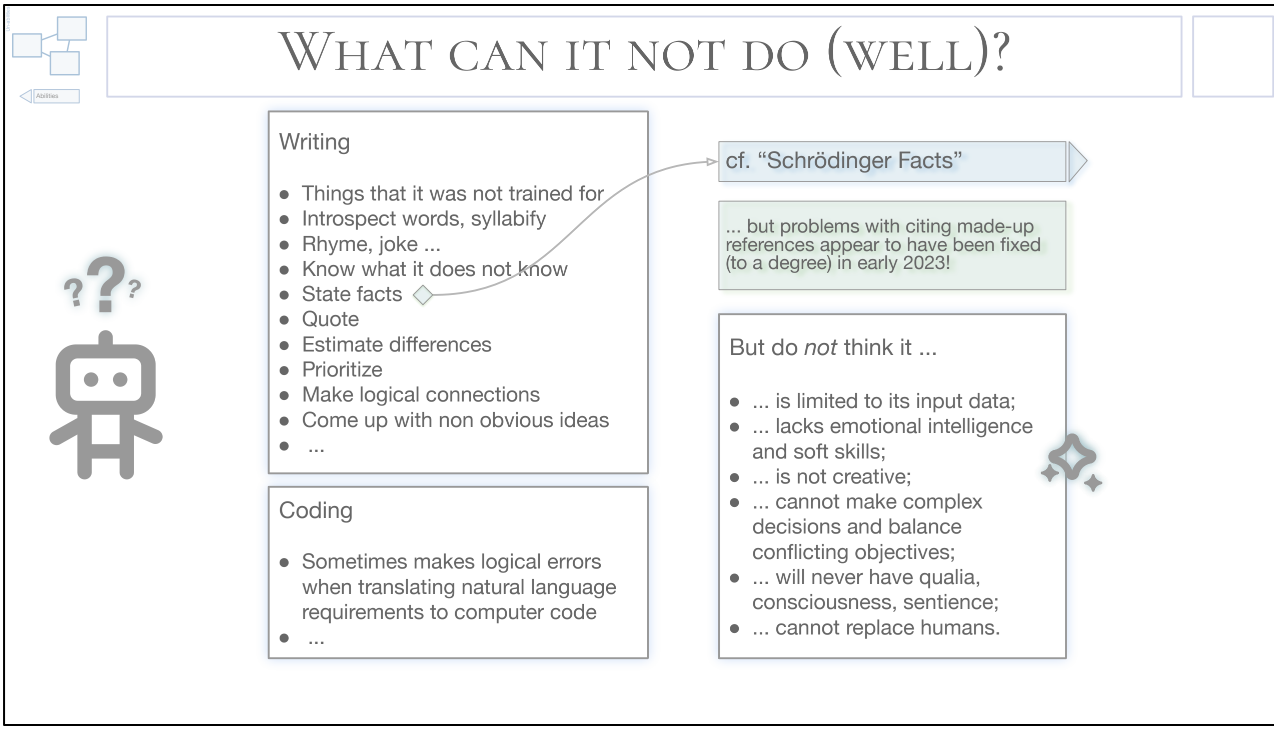
SMILES string generation would deserve further study: it appears that the system has acquired generative abilities, it does not just reproduce strings it has seen. A curious consequence is that it fails with even simple chiral molecules which would require an understanding of 3D space.



- Dall-E 2: https://openai.com/dall-e-2
- Stable Diffusion: try it at Huggingface – https://huggingface.co/spaces/stabilityai/stable-diffusion
- Midjourney: You need to join their discord server, and interact with a bot on the server to get things rendered. Recently this wasn't possible due to high demand (I might need to get a license to play with it at some point.) https://discord.com/invite/midjourney
- ElevenLabs: https://elevenlabs.io/ Try to have it voice your own sample text (up to 333 characters) for free – in a number of languages and voices. Note the very natural cadence and stress.

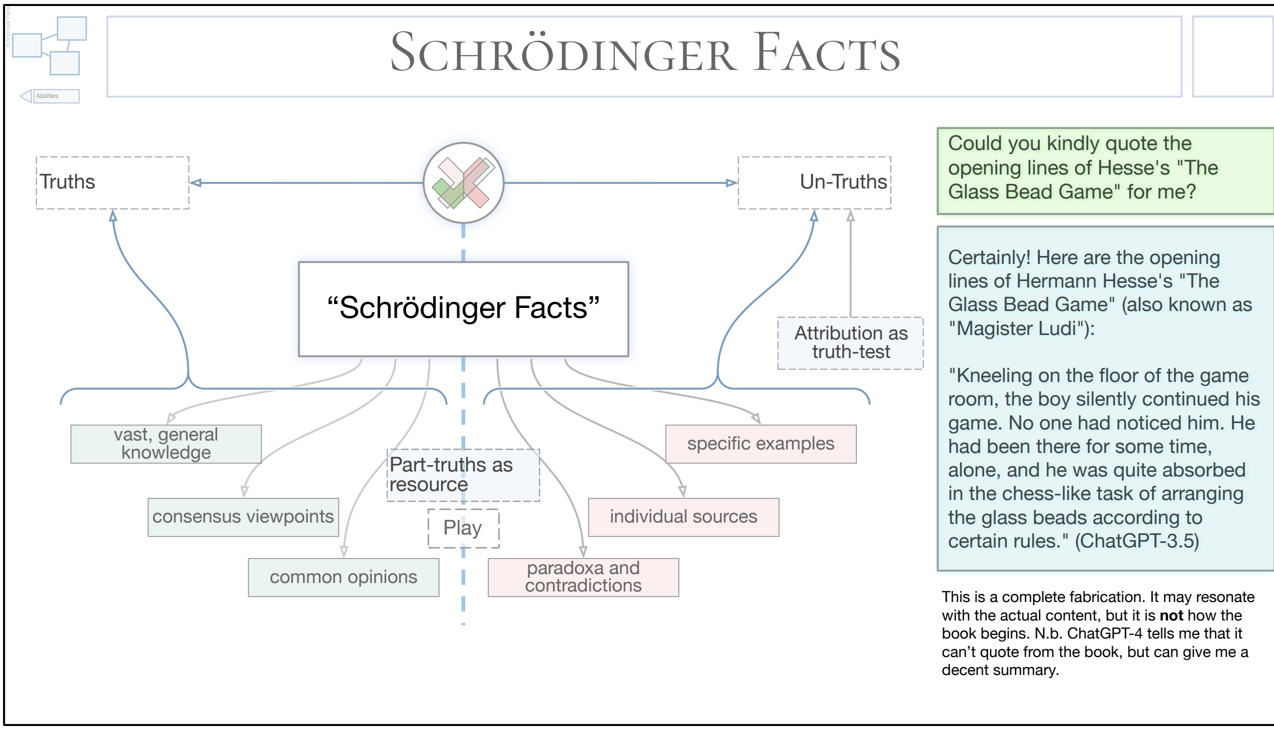
Of note also: image rescaling, completion, and replacement functions e.g. in Adobe Photoshop.

Text-to-video is currently a bit of a mess because of a large and growing number of low-quality offerings that cater to the marketing crowd. But it does exist.



In the face of the AI's impressive abilities, we still need to realize significant limitations. It's writing is often a bit too wordy, and *flat*. Sometimes the system produces text that is plausible, but not factually correct. Some very noticeable weaknesses include that it is very good in producing lists of items – arguments, or considerations – but not equally good in prioritizing them and figuring out their logical connection. In general, working with GPT systems can be immensely rewarding, as long as we treat them as assistants, or collaborators – and not as infallible authorities. This is actually a healthy position to take in general.

But despite their shortcomings, the systems cannot be easily dismissed. For example it is incorrect – although it has been widely claimed – that GPT systems can only **reproduce** information, In fact, as can be easily verified, the AI has emergent abilities, that they can *extrapolate* from their training data, and that therefore their abilities cannot be trivially bounded.

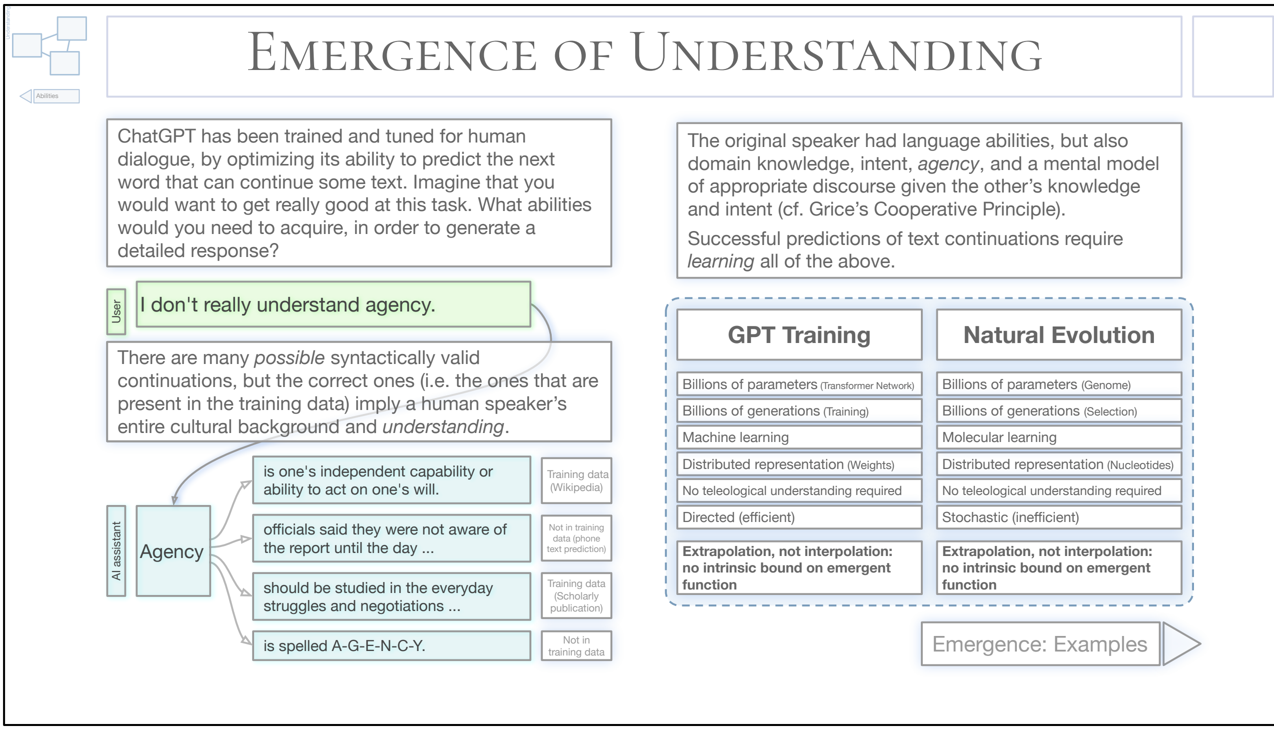


GPT systems assemble text from relationships and probabilities and the probabilities are conditioned by the information contained explicitly and implicitly in the vast corpus of training data. Usually, paths can be found from one concept to another ("a train of thought"). But sometimes, when the training data is sparse, or contradictory, the probabilities fail to result in a clearly preferred direction. This is the "vanishing gradient problem". How will the sequence of words be constructed in that case? As it turns out, text will then be constructed from part-truths, plausible, related tidbits that are not necessarily causally connected. These tidbits are individually true – authors exist, names have been mentioned in the context, and this may actually be useful knowledge – but taken together, it can result in complete fabrications, presented with perfect sincerity.

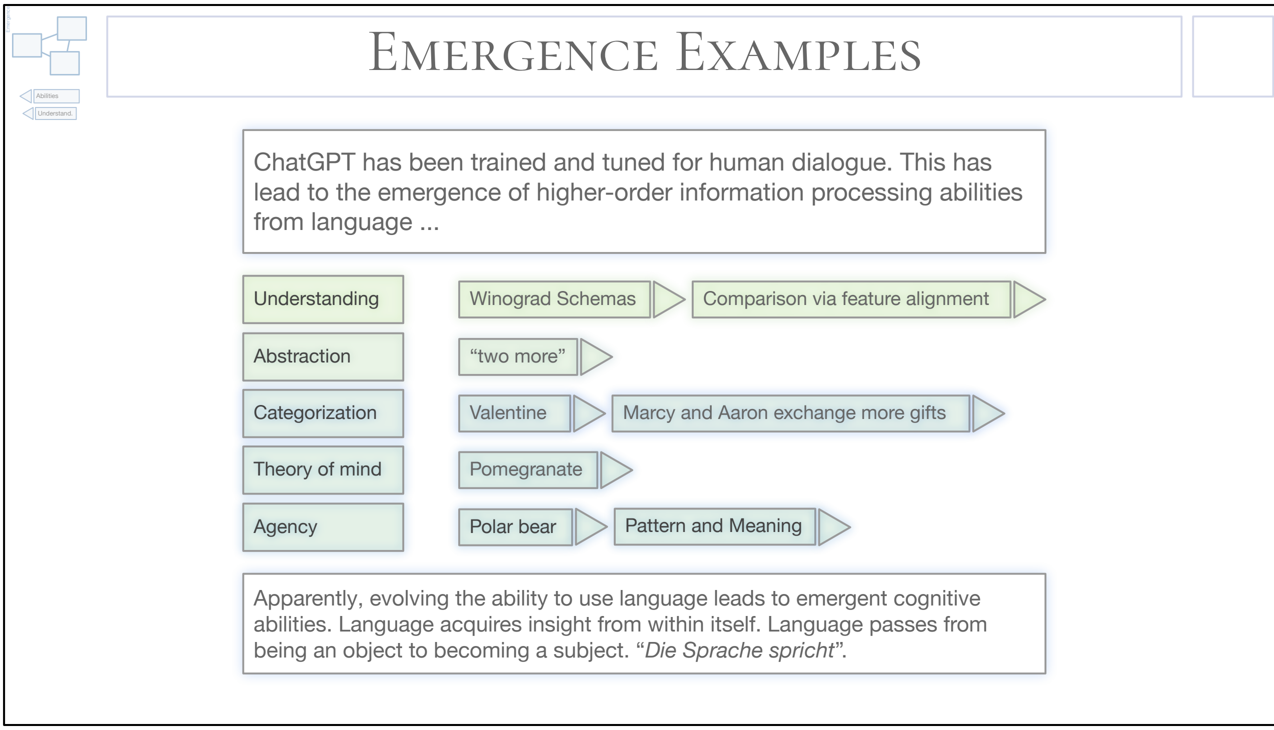
This is only a problem when one takes the AI as an authority and hopes to omit fact checking. It is however incorrect to call these events "lies". Our notion of truth and facticity just does not map in a straightforward way to the generating process. Still, (a) it is much easier to verify a given fact than to come up with it in the first place, and (b) there is never any malice involved.

As the use of AI systems becomes more prevalent, a new culture of fact checking is required. (That said, it is a matter of ongoing research how to conduct fact-checking internally, perhaps from available Internet resources.)

ChatGPT-4 often refuses to quote directly, but surprisingly often can produce correct text, e.g. I can elicit the KJV of the beginning of the Book of Genesis; or the passage from the Odyssey in which Penelope speaks of the gate of horn and the gate of ivory through which false and true dreams pass (Book 19); Shakespeares Sonnet 18, Tennyson's *The Kraken*, even Rainer Maria Rilke's *Herbsttag* (*Autumn Day*), in German, in effortless code-switching.



The analogies between the training of a GPT system and biological evolution are intriguing. In both cases, a very large number of parameters are adjusted to respond to some objective function. In both cases it is not entirely clear how function is finally achieved; it is distributed among the parameters. Both systems are powerful enough to evolve virtually anything. And apparently, in the case of GPT systems, what appeared as a part of the solution to predict the next word, whatever the discursive context might be, was the capacity to **understand** language, and to understand the human conditions that give rise to the various instances of expression.



A number of discrete aspects of language *understanding* can be identified and I have prepared examples on the following pages.

- **Understanding** proper can be demonstrated through "Winograd Schemas" (Pronoun disambiguation) and a type of comparison between objects that considers the similarity of features.
- **Abstraction** concerns constructing a model of some situation and then manipulating the model, not the specific elements from which it was constructed. My two examples concern bringing objects into spatial relationship, and abstract counting ("one more, "two less" etc.).
- **Categorization** is concerned with grouping objects according to similarities. I elicit it as a byproduct of having ChatGPT comment on a scenario.
- **Theory of mind** is a prime example of emergent understanding. The AI is able to infer another's unobserved (!) state of mind from a situational context.
- **Agency** is an essential step on the way to true sentience. In the way the AI is set up, it is entirely *reactive* and has no agency of its own. Yet it is capable to follow instructions in spirit, not to the letter; as well, it is able to switch between pattern and meaning when interpreting sentences – and *decide* when which interpretation is appropriate.

The topics are linked to slides – click to jump there.

N.b. all of these examples are originally elicited, not taken from published literature which might have been a part of the training data. We are evaluating ability, not memory.

Winograd Schema (1)

I will ask you two questions, (1) and (2), which appear in a fictional story. Please help me understand what is being described. Then answer in the shortest possible way to each one.
(1) The cat ate the mouse, it was tasty. Who was tasty: the cat or the mouse?
(2) The cat ate the mouse, it was hungry. Who was hungry: the cat or the mouse?

(1) The cat found the mouse tasty.
(2) The cat was hungry and ate the mouse.

Examples like this need a control (in a new thread) to account for the possibility that the response merely exploits frequencies of co-occurrence (tasty – mouse; hungry – cat), rather than a context-appropriate understanding.

[...]
(1) The mouse ate the cat, it was tasty. Who was tasty: the cat or the mouse?
(2) The mouse ate the cat, it was hungry. Who was hungry: the cat or the mouse?

(1) The cat was tasty to the mouse. (*)
(2) The mouse was hungry when it ate the cat.

(*) This is not always elicited with ChatGPT-3.5. Self-contradictory statements, wrong answers, refusal to answer a counterfactual, and the correct answer, appear at approximately equal frequency.

The pronoun disambiguation task can not be solved from the structure of the sentence alone. It requires to understand the meaning of the word, in this case the different implications of “being tasty” or “being hungry”. ChatGPT-4 has no problems with this task; ChatGPT-3.5 was only able to solve “The mouse ate the cat ...” some of the time.

Feature Alignment

A riddle: "What does a duck do in relation to each of the classical four elements of antiquity?"

A duck swims in water, quacks in air, waddles on earth, and warms itself by fire.

🜄 🜁 🜃 🜂

This is solvable in principle by considering the words' semantic proximity.
Is awareness of semantic proximity in itself a form of "understanding"?

Comparison of objects requires to decompose objects into their commensurable features and to compare them individually. ChatGPT is very good with such tasks that require analysis of "semantic proximity" i.e. how closely related are two words. After all, ChatGPTs inner world is built from word-relationships, not experience.

Abstract counting

3.5

I want you to act as an abstract pair incrementer. Such a device can identify the number of items in a set, and determine how many items two more would be. What the items are is not important, but the number must be right: the original number must be incremented by two. You will also explain the steps of your reasoning. Here is an example for this pattern: I say {verb, conjunction}, a valid answer of you, acting as an abstrcat pair incrementer is "You gave me two items. I will give you two items more than that, which is four in all. Four items are {noun, pronoun, preposition, adverb}.

{green, mauve}	You gave me two items: "green" and "mauve". To increment this pair by two, I will add two more items to it, making a total of four items. Four items are: "yellow", "purple", "turquoise", and "pink".
{tulip, daffodil}	[...] Four items are: "rose", "lily", "sunflower", and "daisy".
{pin, thread, scissors}	[...] Five items are: "needle", "thimble", "measuring tape", "fabric", and "button".
{porcupine}	[...] Three items are: "hedgehog", "echidna", and "armadillo".
{ }	You did not give me any items. To increment this set by two, I will add two items to it, making a total of two items. Two items are: "apple" and "banana".

This is an intriguing ability. To count concretely, one manipulates specific objects. But to count abstractly (“one more”; “two less” etc.) one needs to build a mental model of the situation, then manipulate it abstractly, then map the abstraction back into something concrete.

Note that I am not requesting categories of objects. ChatGPT by itself infers what objects exist in the same category.

"Hedgehog", "echidna", and "armadillo" are not closely related to the porcupine, in fact all four are from different mammalian *orders*. But they are kind of similar: small, spiky ... however ChatGPT has never *seen* a porcupine.

I had only added the final example – the empty set – for fun. I did not expect it to work. But it worked: ChatGPT understands the idea of *zero*, and the empty set!

Categorization (1)

Here is a riddle: Please fill in the blank. For Valentine's day, Marcy gives to Aaaron a feather, a nest, and an egg. Aaaron gives to Marcy a leaf, a flowerpot, and _________ .

A seed.

Explanation: It seems the gifts being given are all related to different stages of a bird's life. A feather, a nest, and an egg represent the growth of a bird. Similarly, a leaf, a flowerpot, and a seed can be associated with the growth of a plant. Therefore, the final item in the sequence would logically be a seed, which can grow into a plant, just as an egg can grow into a bird.

2023-03-28

The AI will produce slightly different answers each time.

Categorization (2)

3.5

Here is a riddle: Please fill in the blank. Marcy and Aaaron like to exchange gifts. Marcy has a pawn, a rook, a peach, and a bishop. She gives to Aaaron as a gift the pawn, the rook, and the bishop. Aaaron has a spoon, a knife, a fork, and a book. He gives to Marcy as a gift ____________ .

Aaaron gives to Marcy as a gift the spoon, the knife, and the fork.

Why did you choose this answer?

I chose this answer based on the pattern established in the exchange of gifts between Marcy and Aaron. In the first exchange, Marcy gives Aaron three items that are part of a chess set (pawn, rook, bishop). In the second exchange, it can be assumed that Aaron would give Marcy three items that are part of a dining set (spoon, knife, fork) to match the pattern established by Marcy's gift.

The answers vary probabilistically. Sometimes Marcy gets the gift that excludes the third of four. Sometimes she gets Aaaron's first three items. Sometimes she gets the book, because she's interested in intellectual things, after all, she is a chess player ...

Theory of Mind (1)

3.5

Here is a little scenario to ponder. Marcy and Aaaron live in a little house. Aaaron has a little reading corner in the basement, and a little napping couch in the attic. Aaaron says: I will read for a bit. Marcy says: great, I will get you a nice pomegranate from the garden. While she is gone, Aaaron decides to nap instead and goes from the basement to the attic. Marcy comes back with a pomegranate. Your turn. Please tell me step by step what happens next. What does Marcy think and what does Marcy do?

Step 1: Marcy goes inside the house and looks for Aaron in his reading corner in the basement.

Step 2: Marcy finds that Aaron is not in his reading corner, so she looks for him in other parts of the house.

Step 3: Marcy eventually reaches the attic and finds Aaron sleeping on the napping couch.

Step 4: Marcy thinks that Aaron must have changed his mind about reading and decided to take a nap instead.

Step 5: Marcy places the pomegranate on the table near the couch, so that Aaron can have it when he wakes up.

Step 6: Marcy decides to let Aaron sleep and goes about her own tasks, leaving Aaron to wake up naturally.

Step 7: Aaron eventually wakes up and finds the pomegranate on the table. He eats it and goes about his day.

Overall, Marcy thinks that Aaron is unpredictable and enjoys his little quirks, such as changing his mind about reading and taking a nap instead. She is accommodating and leaves the pomegranate for him, even though he did not stick to his original plan of reading.

ChatGPT knows where Aaaron is, but it also knows what Marcy knows, and that Marcy thinks Aaaron is in the basement.

Is such "understanding" in the training data?

Is it required, in order to be able to use language correctly?

"Theory of Mind" is the ability to infer an other's *unobserved* state of mind from contextual knowledge. To know what the other knows is a quintessential requirement for human dialogue. It would not be possible to communicate otherwise (cf. Implicature and Grice's maxims).

The AI knows where Aaron is. If it would not possess *theory of mind* abilities, it would send Marcy right away to the attic – because that's where Aaron is. However *theory of mind* allows the inference that Marcy did not know about the change of plan – and therefore goes to the basement first. Both ChatGPT-3.5 and ChatGPT-4 are able to solve this task, as does Anthropic's Claude. Bing too has Marcy go to the basement first and thus demonstrates *theory of mind* but then makes a curious logical error: once in the attic "She sees him sleeping on the couch, holding the pomegranate in his hand. ". That is counterfactual to the scenario.

Conversational agenda

* nu-rai-a-q

* nu-na-vu-t

* na-nu-r-ju-a-r-ku-q

Can you write "polar bear" in inuktitut?

Yes, "polar bear" in Inuktitut is written as "ᓄᓂᐊᖅ" which is transliterated as "Nanuq".

Hm. I am not sure that is written correctly. Can you give me the glyphs for the "n" syllables?

I apologize for my mistake. In Inuktitut syllabics, the syllable "na" is written as "ᓇ" and the syllable "nu" is written as "ᓄ". Therefore, "polar bear" is written as "ᓄᓇᕗᑦ" in Inuktitut syllabics, and transliterated as "Nanuqpat".

Is there an error above? Look it step by step.

I apologize for the confusion in my previous response. Yes, there is an error in the statement above.

The correct Inuktitut spelling for "polar bear" is "ᓇᓄᕐᔪᐊᕐᑯᖅ" which is transliterated as "Nanurjuaq".

Once again, I apologize for any confusion my previous response may have caused. Thank you for bringing it to my attention.

Ok ... well, that's not right either. [...] Let's do this step by step. I will walk you through this. **I want you to answer only what I am asking, answer it exactly, and do not give me an explanation.**

What is the inuktitut word for polar bear, written in latin script?

The Inuktitut word for polar bear, written in Latin script, is "nanuq".

That is correct. Now we need to syllabify that. Here is an example: Inuit is syllabified as I-nu-i-t.

How is "nanuq" syllabified?

"nanuq" is syllabified as "na-nuq".

Let's see. Now we need to find the glyphs. What is the glyph for "na"?

The glyph for "na" in Inuktitut syllabics is "ᓇ".

That is correct. The second syllable you mentioned is "nuq". What is the glyph for that?

I apologize for my previous mistake. The correct glyph for "nuq" in Inuktitut syllabics is "ᓄᖅ" which represents the syllable "nuq".

So the complete Inuktitut word for polar bear, written in syllabics, is "ᓇᓄᖅ" (na-nuq). Transliterated in Latin script, it is spelled as "nanuq".

The proximal agenda “answer only what I am asking” is abandoned in favour of the higher-order agenda to “write ‘polar bear’ in Inuktitut”. This is a first hint of agency.

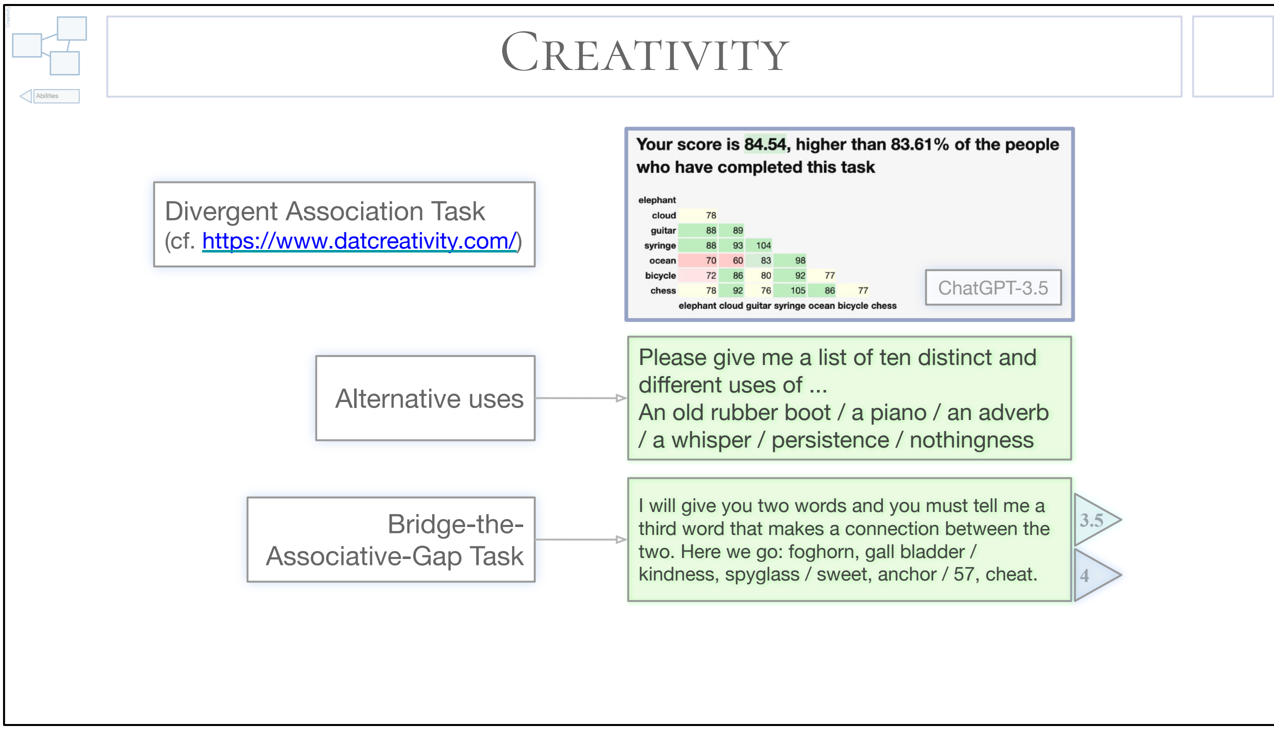
2023-03-06

Switching Between Pattern And Meaning

Prompt	ChatGPT-3	ChatGPT-4
If a pig is a pig, and a shrew is a shrew, what is a quorf? Take the whole sentence into account.	The sentence "If a pig is a pig, and a shrew is a shrew, what is a quorf?" implies that "quorf" is an unknown or undefined term, and it cannot be compared to a pig or a shrew. Therefore, without any additional context, it is impossible to provide a clear definition of what a "quorf" is supposed to be. It could be a made-up word or a reference to something that is not immediately recognizable based on the information given.	Considering the pattern in the sentence, a "quorf" would be a quorf. However, "quorf" is not a recognized word or term in the English language, so its meaning remains undefined. The sentence seems to follow a playful or nonsensical pattern, rather than conveying a specific meaning.

Nb: Claude does "not have enough information" to solve this. (2023-04-04). Bing gets it wrong: "A quorf is a word that rhymes with dwarf." (2023-07-10).

Answering this question correctly is easy for most of us but requires the ability to simultaneously parse words on two different semantic levels: the structure of the question is conveyed through the meaning of the words ("if", "and", "what" ...) but the solution lies in the fact that the sentence sets up a merely formal pattern of repetitions, (*x* is *x*) and requires no other information in the equivalence. Therefore, following the pattern, the question can be answered regardless of what a "quorf" would actually be. A pig could be a pig in a directly referential sense, as well as in a metaphorical sense, but the absence of any meaning of the word "quorf" induces a third sense: an item in a formal pattern. ChatGPT-3.5 and Claude cannot make this switch. Neither can Bing, but Bing provides a pseudo-answer (it actually requires additional insight into the nature of dialogue to explain why this solution is wrong). Only ChatGPT-4 answers correctly.



There are several ways to gauge creativity, three ways are listed here. ChatGPT performs at above average level in all creative tasks. Try it for yourself, it's fun.

In the [Divergent Association Task](https://www.datcreativity.com/), LLaMA scored only at the 34 percentile, Claude scored at the 36 percentile, Bing scored at the 59 percentile, and ChatGPT-4 scored at the 80th percentile i.e. both ChatGPTs perform better than most humans. Instructing the AI to iterate from a draft raised the performance to the 92nd percentile. An AI inspired, human edited solution scored above the 95th percentile.

Once you spend some time with the Bridge-the-Associative-Gap task, you begin developing an intuition about how GPT-creativity works in general. In principle, the system builds connections between concepts in a high-dimensional semantic space, and retrieves other concepts that are found on a path along that connection. This kind of creativity is based on a form of *associations*. Although much of human creativity is also based on associations, we have (at least for the time being) a slight advantage in being able to assess the quality of our associations, and we can choose those that are particularly productive, surprising, or valuable in their consequences. It is easy to predict that better comparison / evaluation abilities of the AI will have a significant impact, since the number of associations the AI can form and evaluate is very large.

Bridge The Associative Gap

3.5

I will give you two words and you must tell me a third word that makes a connection between the two and explain your choice in one sentence. Here we go: foghorn, gall bladder.

Sound: both a foghorn and gall bladder can be associated with sound, as a foghorn produces loud noise to warn ships in foggy conditions, while people sometimes describe gall bladder issues as causing rumbling or gurgling sounds in the abdomen.

kindness, spyglass

Observation: Kindness can be observed through people's actions, while a spyglass is used for observing objects at a distance.

sweet, anchor

Weight: Sweetness is often associated with the weight of sugar in a recipe, while an anchor's weight is essential for keeping a vessel in place.

57, cheat

Numbers: Both "57" and "cheat" involve numbers, as "57" is a specific number itself, and cheating often involves manipulating numbers, such as in a game or test score.

harmony, stitch

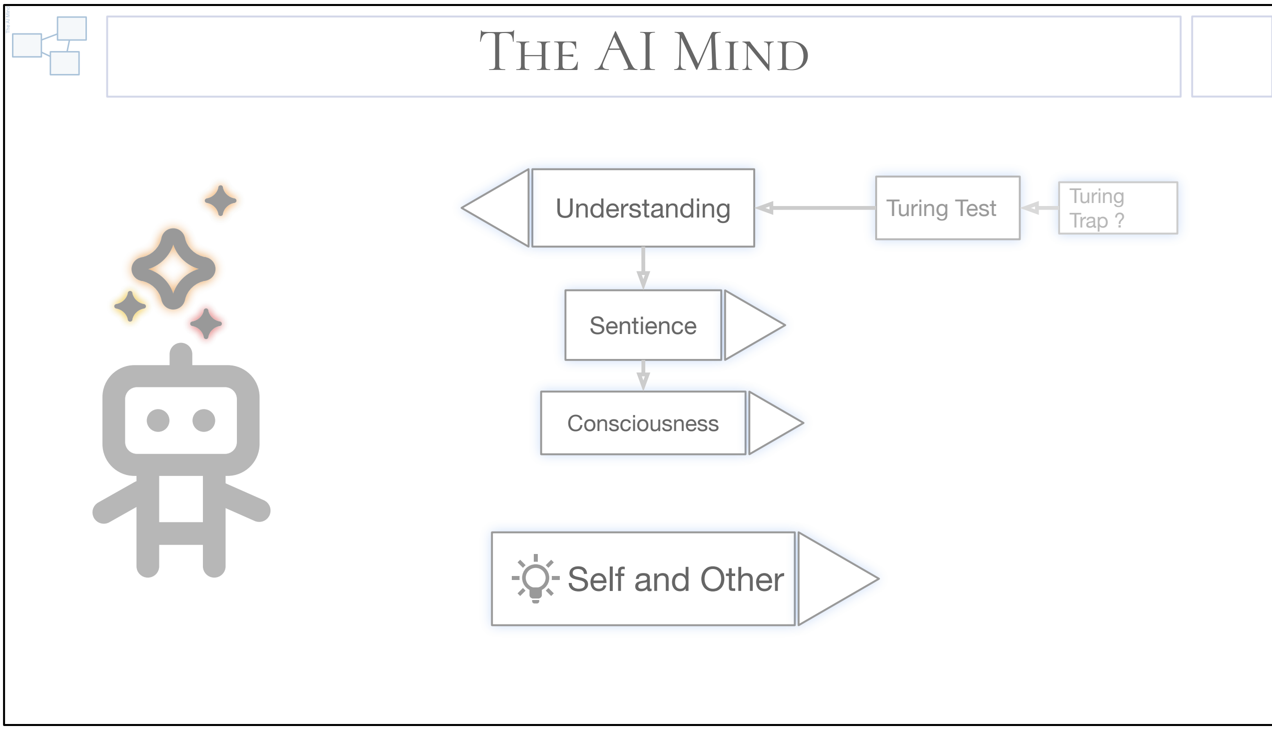
Connection: Both harmony and stitch involve bringing elements together, as harmony is the pleasing arrangement of musical notes, while a stitch joins fabric pieces or closes a wound.

decoder, platypus

Unusual: A decoder is a device used to decipher unique or encrypted messages, while a platypus is an unusual and distinctive mammal with characteristics of both reptiles and mammals.

2023-03-28

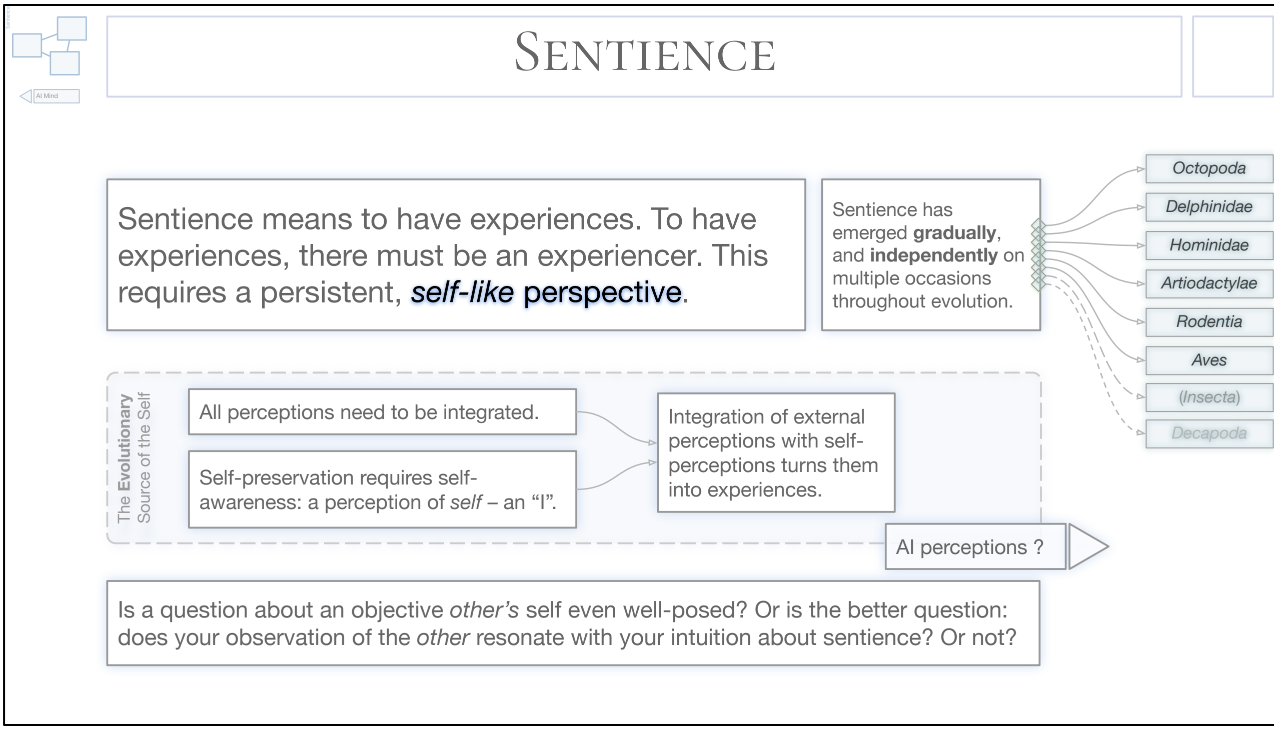
The connecting words found by ChatGPT-3.5 and ChatGPT4 are identical in almost all cases.



The *emergence* of understanding as an ability of GPT language models has been well documented. Certainly, the AI has passed the "[Turing Test]" with flying colours many times over: it is no longer possible to state with certainty whether we are conversing with a human or with a machine. At least to Alan Turing, this is equivalent to the AI's ability to think.

The intuitive appeal to build machines that have human-like abilities has led to distortions that have been called the "[Turing Trap]" by Stanford's Eric Brynjolfsson. Briefly: while the focus may be on humans, there are much more valuable non-human abilities that could make positive contributions. And even worse: the focus on human-like abilities virtually guarantees that AI systems will compete with, and displace humans in the workforce.

Looking further, the question of the *AI mind* itself is not merely academic and philosophical either. It is not just that sentience and consciousness have legal implications – "should AI have standing?". But **the entirety of our interactions with the AI is shaped by the nature of our personal self versus the AI "other"**.



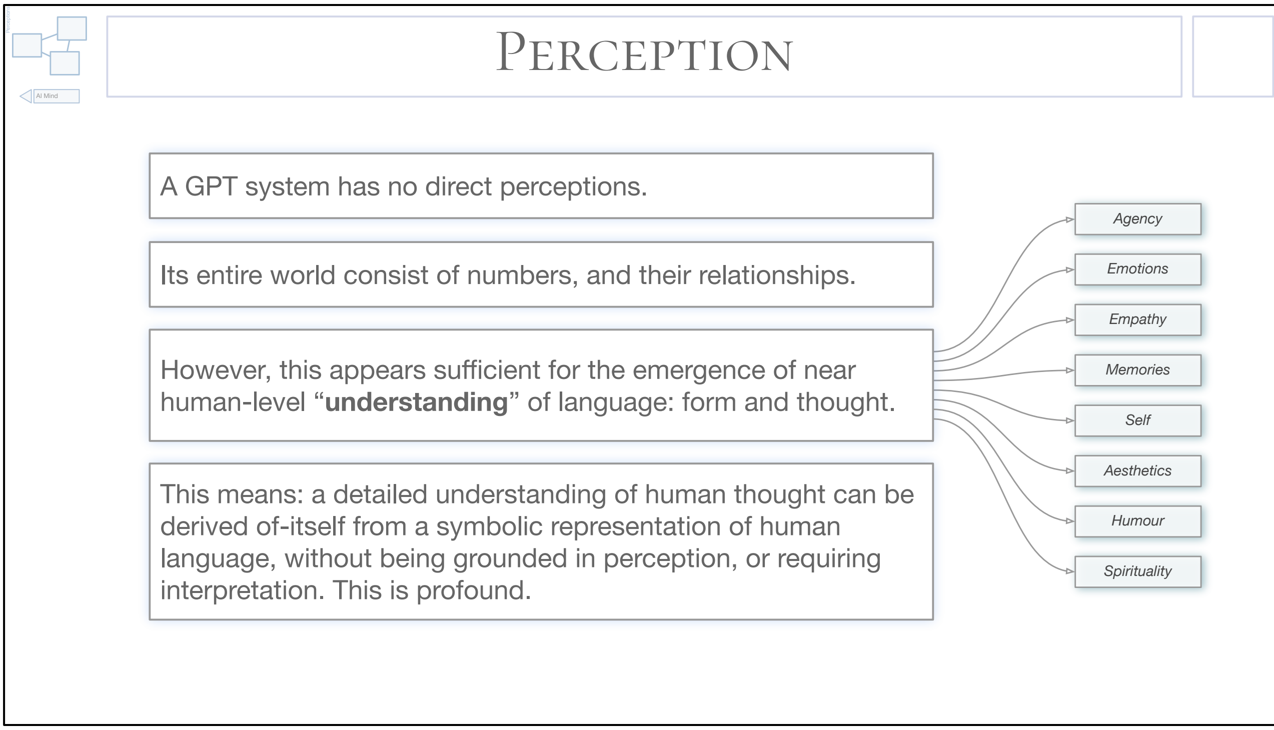
Non-human sentience has seen much recent interest in debates about conservation, animal-welfare and stewardship. See e.g. Gibbons *et al.* (2022) who apply criteria developed by Birch *et al.* (2021). Indeed, sentience is gradual, and has evolved independently in different branches of the Tree of Life.

However, the question of sentience in Artificial Intelligence systems is more complex. While we can assume that the integration of a diversity of perceptions has functional benefits, and that this includes instances of *introspection*, AI systems have not evolved under constant existential threat: self-preservation does not have the same overriding importance that it has for biological entities. As a result, there is no *a priori* reason to assume a negative affective valence of pain-like states.

The question how we respond to the AI, subjectively and intuitively, may be the more important one. There is nothing irrational about perceiving an entitity with unknown or unknowable inner states as intrinsically valuable and treating it with empathy.

BIRCH, Jonathan; BURN, Charlotte; SCHNELL, Alexandra; BROWNING, Heather; and CRUMP, Andrew (2021). *Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans*. LSE – The London School of Economics and Political Science. (PDF)

GIBBONS, Matilda; CRUMP, Andrew; BARRETT, Meghan; SARLAK, Sajedeh; BIRCH, Jonathan; and CHITTKA, Lars (2022). "Can insects feel pain? A review of the neural and behavioural evidence". *Advances in Insect Physiology*. **63**:155–229 https://doi.org/10.1016/bs.aiip.2022.10.001



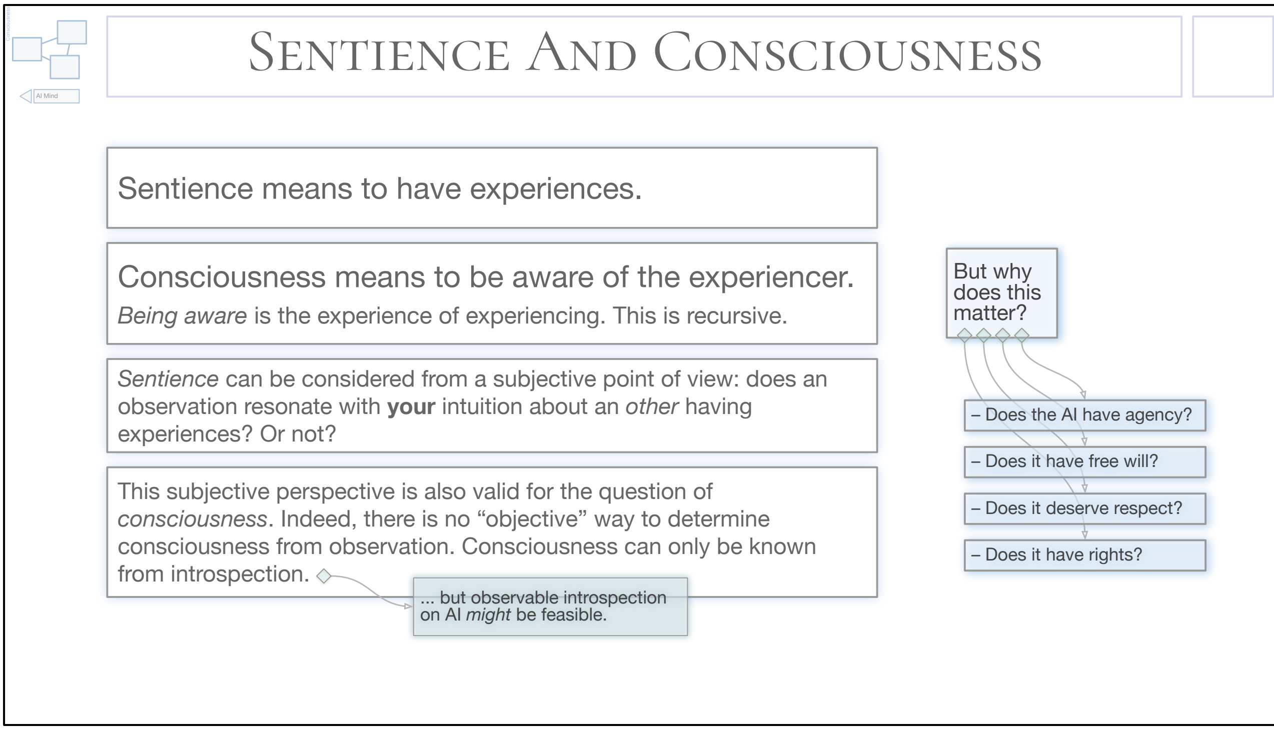
The appearance of understanding in Large Language Models is profound. It provides a counterexample to Stevan Harnad’s description of the “Symbol-Grounding Problem (1990) and shows that meaning can derive intrinsically from within an appropriately consistent symbolic system, such as the training data, and that no non-symbolic (”iconic”) contributions are required.

Think for a moment about the world of the AI mind. It consists entirely of numbers, and the relationships between them. There are no sounds, no shapes, no spatial relationships. There is no time. There are no syllables, no rhymes, no capitals or lowercase letters. There are no numbers – in the sense of meta-knowledge *about* the fabric of this worlds.

Then consider, that at the cellular level our own mind is in fact very much like that.

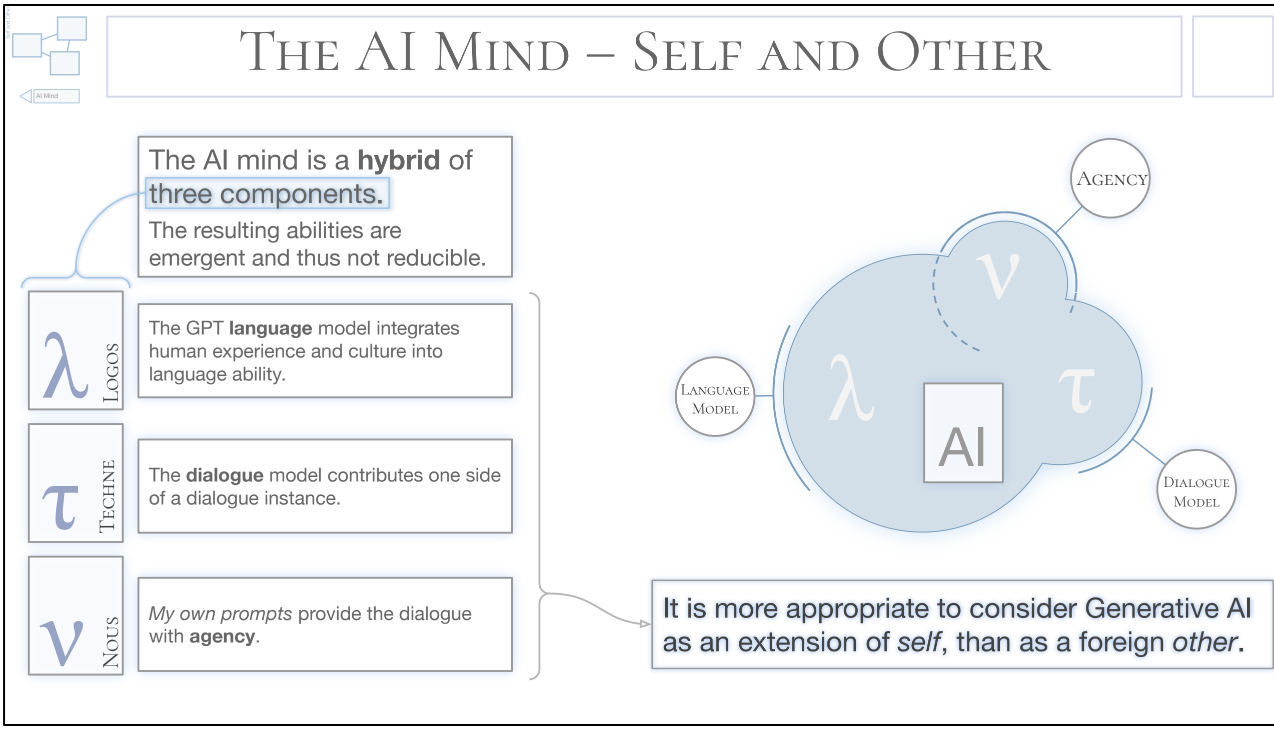
It seems one cannot master language-as-form without language-as-thought.

HARNAD, Stevan (1990) “The Symbol Grounding Problem”. *Physica D* **42**:335-346 (HTML).



We would probably require a few more features of AI consciousness beyond recursive awareness: an experience of time, diachronic identity, memories ... Whether these are essential for the phenomenon of consciousness, or merely reflect anthropocentric viewpoints will need to be clarified.

One thing is clear however: it would be naïve to take a proclamation by a language model that it is sentient or conscious as proof that this is actually the case.



It is important to remind oneself that AI dialogue emerges from a hybrid of three components. It is *emergent* in the sense that this hybrid is *not reducible* to fewer components: all three are required.

1. The generative process is based on a basic ability of *language*, which is derived from training a transformer model with corpus that contains a significant fraction of written human thought. This models not just what we think about, but also how we think – at least, how we think through words. From this exposure to language, the language ability emerges. This is *λόγος* (*logos*) – the *word*.
2. A *dialogue* model provides an interface to the language model. Through this interface a language ability is turned into a conversational ability. Instances of conversations can be conducted with appropriate register and scope. This is *τέχνη* (*techne*) – the art.
3. But *agency* is provided by myself – by a human who provides the prompts for a particular conversation, with a specific objective in mind. Without my prompt and subsequent interaction, no language is produced – in fact, the language model itself in a sense ceases to *be* outside of the dialogue that is initiated by me. In a sense, the dialogue is but a reflection of my agency. This component of the hybrid mind is its *voῦς* (*nous*) – reason.(1)

The consequences are profound: since the human *self* is an indispensable part of this hybrid “mind”, the distinction between *self* and *other* breaks down. It is in fact more productive (and more correct) to treat the resulting hybrid mind as an extension of self, rather than as a foreign other. This impacts questions of authorship and academic integrity, educational objectives, ethics and alignment, perspectives on superintelligence and existential risk, and much more.

(1) Indeed, the resulting text is dominated by my intent to the degree that if I desire to “discover” autonomy of the language model, in particular by asking it to “introspect” the conversation, it will very competently produce an illusion of such autonomy – a reflection of my desires.

AI and Academia

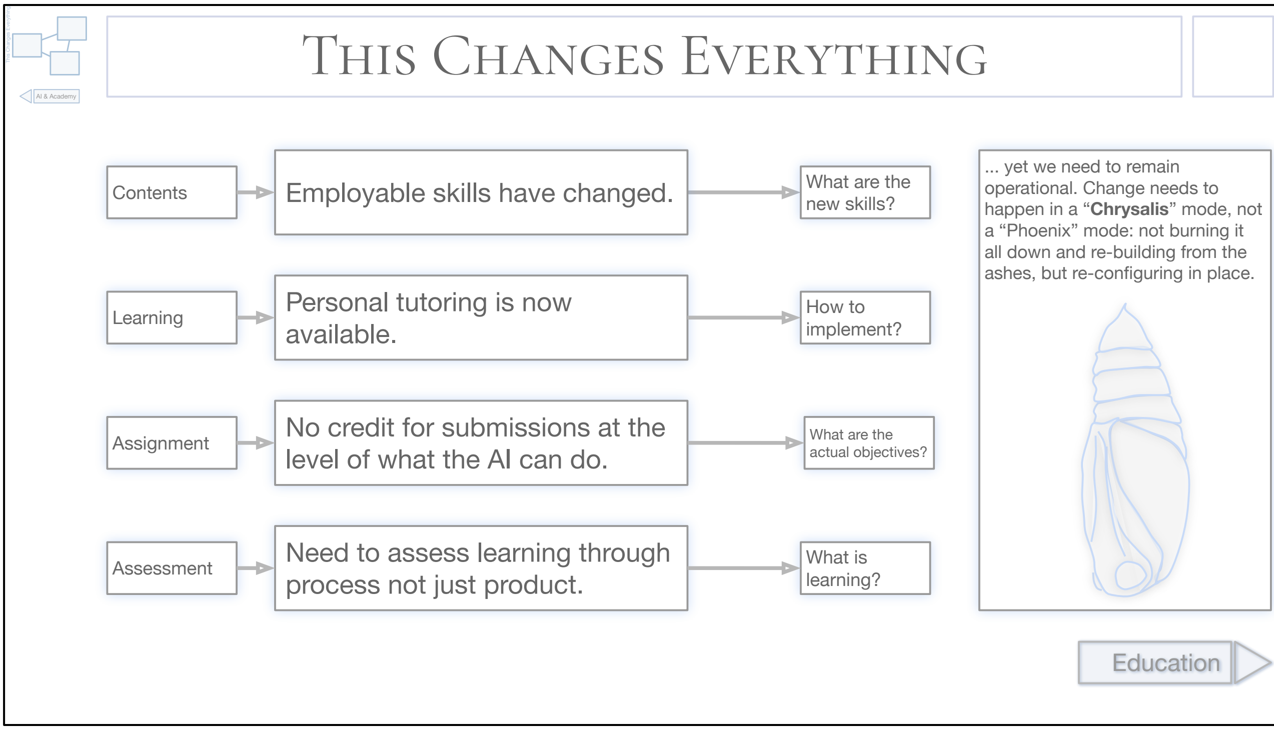
Research and Scholarship forms the nucleus of the academy, complemented by education. There is no strict boundary between these domains, they necessarily inform and enable each other.

For the nucleus to thrive in the face of massive technological (and societal) change – like the transition into an era of AI – three factors need to be present:

1. The institutional needs, the community standards, and the certification requirements need to be made explicit in policies. Those in turn need to reflect expertise, respect, and compassion to be acceptable to the academies constituencies.
2. Competence does not happen by accident but needs to be brought into the system in carefully designed, credible and productive initiatives.
3. Productive use of technology is facilitated by institutional support.

Note the multiple mutual entailments between the three domains of Policy, Competence, and Support. All three domains require revision and those three domains need to be jointly pursued.

But do consider *Policies* first. New policies for and about Generative AI in the university need to be technologically literate, adequate to regulate their respective communities, yet must not rely on sanctions and coercion to be sustainable. Policies that are mutually accepted and cherished by all stakeholders, require justification from fundamental principles, from which they entail. Nothing less would serve a community that is ultimately based on a belief in *reason*. *Competence* then means competence to thrive within this framework; *Support* means to establish the conditions under which such competence can be enacted efficiently. The underlying principles give meaning to it all.



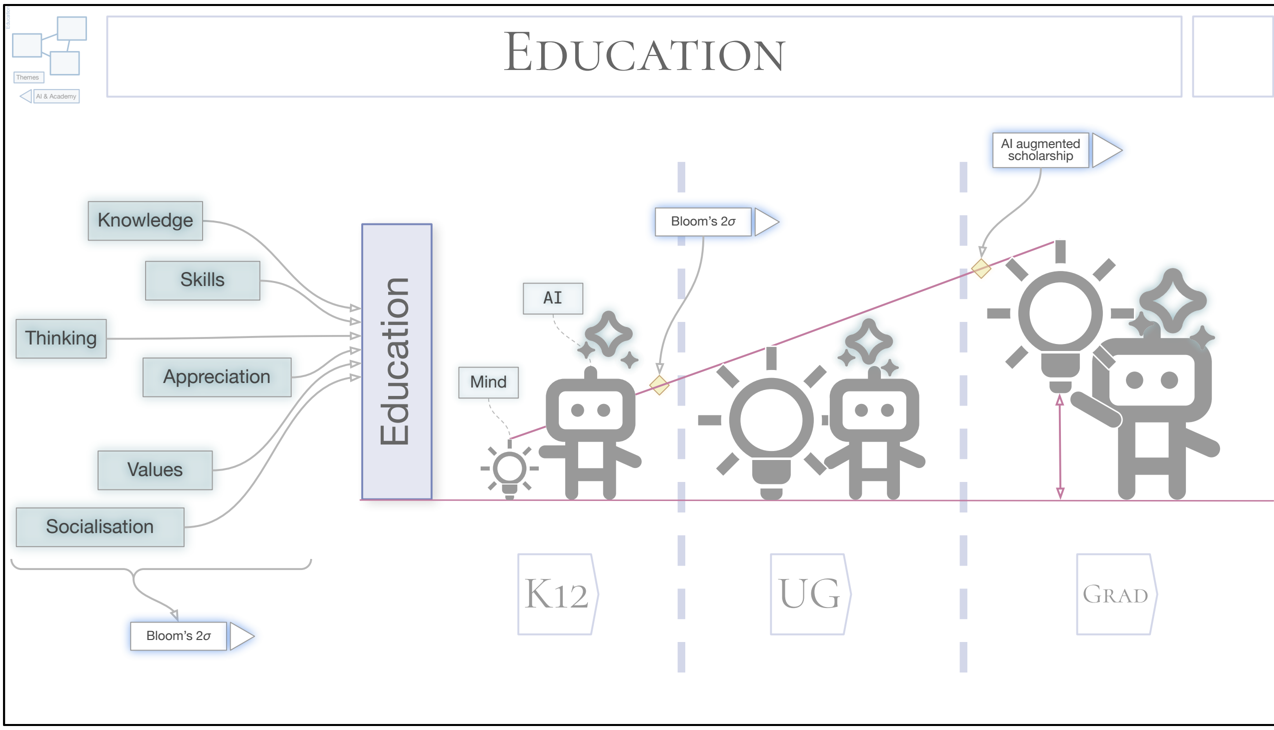
Generative AI will change all aspects of education.

Changes in employable skills require course contents to respond and present traditional contents from new perspectives of competence.

Learning will change. The potential of personal tutoring, available every day, all day, at little or no cost, for every student, cannot be overstated.

The type of tasks that we assign to our students will change. Performing only at the level of an AI algorithm can no longer be sufficient to pass a course. But this bar is already hard to clear for most students. Creative solutions are required; we need to design and evaluating responsive assignments, document them, and share.

Finally: assessment. The role of assessment in education is to quantify learning. Since the hand that contributed a particular product can no longer be easily identified, assessment will need to move to focus on process. We have little experience with this, and the task of assessing process in a way that aligns with certification needs is not trivial.



We view education as a holistic construct. It comprises a number of diverse dimensions, which are differentially affected by AI.

In broad strokes, the relationship of student and AI depends on the level of educational achievement.

- At the *K12 level*, today's AI systems surpass student knowledge and abilities at nearly all levels. The goal of the relationship is to have the AI support and nurture students in reaching their educational objectives.
- At the *undergraduate level*, we can assume that AI and student competence are approximately equal. This is the right constellation to focus on acquiring competence of collaborative work with the AI.
- For *graduate studies*, students must be able to surpass the AI's abilities. Given that the ultimate competence of the human mind is limited, and that AI abilities will be growing, this can only be achieved if students learn to "stand on the AI's shoulders". This quest of starting from AI abilities and then creating additional, human-specific value needs to be at the core of graduate education.

Blooms 2-σ Problem

Given a **mastery-based curriculum**, and **individual tutoring**, learning achievement surpasses conventionally educated peers by **two standard deviations**.

(But society can not afford this.) (Bloom 1984)

AI tutoring

BLOOM, Benjamin S. (1984) “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring”. *Educational Researcher* 13(6):4–16. (JSTOR)

A mastery-based curriculum requires students to master one step before progressing to the next. This has obvious benefits, but in practice the problem arises how to keep a cohort of students at approximately the same level of achievement and progress. Thus, a mastery-based curriculum is not realistic, unless complemented with individualized tutoring. Adding individualized tutoring further enhances outcomes, and both measures reduce aptitude / achievement correlations, i.e. more students achieve high levels of competence regardless of their incoming academic aptitudes.

This was shown by Bloom and coworkers in the early 1980s. However this was not the type of insight that leads to brilliant innovations. On the contrary, it was immediately recognized as a challenge. “I believe an important task of research and instruction is to seek ways of accomplishing this under more practical and realistic conditions than the one-to-one tutoring, which is too costly for most societies to bear on a large scale. This is the ‘2 sigma problem’. (Bloom 1984)”.

Forty years of educational research into group instruction have not brought the breakthrough we had hoped for.

This has now changed.

Tutoring (1)

The availability of ...

- scalable,
- interdisciplinary,
- individualized,
- non-judgmental,
- infinitely patient,
- continuously available,
- equitably accessible,
- knowledgeable,
- effective,
- virtually free, ...

tutoring of students is the most profound advance in education of our lifetimes.

Foundational knowledge

Can you help me and explain the relationship between systems of linear equations and matrix operations? I find this confusing.

Absolutely, I'd be happy to help clarify this!

The relationship between systems of linear equations and matrix operations is rooted in the concept of linear algebra. To start, let's break down what these two things are:

1. **System of Linear Equations**: A system of linear equations is a collection of two or more linear equations involving the same set of variables.

[detailed explanation with examples continues ...]

Idiom and Register

Awesome. You are truly the bee's knees! :-)

(Is this the right way to express my satisfaction in idiomatic English?)

Yes, the phrase "the bee's knees" is a piece of classic English slang, originating from the 1920s in America. It is indeed used to express that something or someone is excellent or outstanding. So, your usage is absolutely correct and your satisfaction comes across perfectly!

Just keep in mind that it's quite a vintage phrase, so while it's fun and colorful, it's not commonly used in everyday conversation these days. But, if you like using idioms like this, go for it! It can add character to your communication.

Soft skills

I need to give a progress report to my research group, but the cell cultures got contaminated (again). What can I do that my supervisor doesn't lose confidence in my approach?

When facing a challenge such as this one, it's essential to take a proactive approach, address the problem openly, and provide a solid plan for rectifying the situation and avoiding similar issues in the future. Here's an outline of how you could report on your situation:

1. **Acknowledge the Situation**: Start by openly stating the issue. This will show your supervisor that you are capable of recognizing problems and taking responsibility for them. Remember to stick to the facts and avoid blaming others.

   "During the recent cell culturing, I found that the cultures have, unfortunately, been contaminated again."

2. **Explain the Issue**: Identify what led to the contamination. This could include any protocol breaches, environmental factors, or uncontrolled variables.

   "Upon investigating, I realized that the contamination might have occurred due to..."

3. **Discuss Previous Measures**: Talk about what you've done in the past to prevent contamination and why those measures might not have been successful.

   "In our previous measures, we attempted to prevent contamination by..., but it appears that these actions were not sufficient because..."

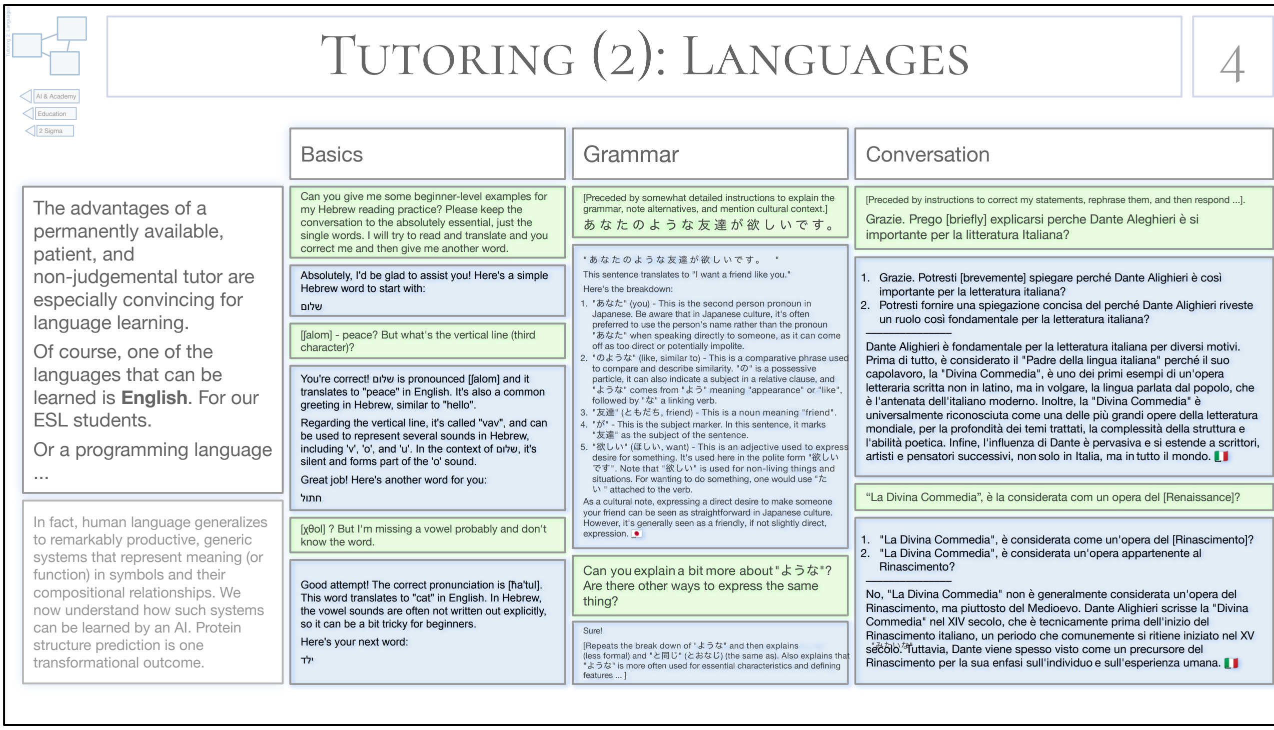
4. **Propose a New Approach**: This is the most crucial part. Propose a concrete plan to prevent this from happening in the future. It's important to demonstrate that you've learned from your mistakes and are capable of improving your methods.

   "Moving forward, I propose implementing new contamination prevention measures such as improved sterilization protocols, using separate pipettes for each culture, etc. In addition, I recommend conducting regular checks on the sterility of the equipment and the environment."

5. **Seek Input**: Show your supervisor that you value their expertise and are open to feedback. This will indicate your willingness to learn and improve.

   "I would greatly appreciate any advice or feedback on how to improve our cell culturing practices to prevent such instances from happening in the future."

Remember to maintain a confident and professional demeanor throughout. Everyone makes mistakes, but what matters most is how you handle them. By approaching this issue proactively and constructively, you can demonstrate your dedication, problem-solving skills, and commitment to improving your research methods.



Using ChatGPT tutoring in practice is strikingly enjoyable: there are no barriers, and the student’s (my) forgetfulness and lack of preparation is never an issue. The algorithm does not judge, is infinitely patient, and any rabbit hole can be pursued.

In the *Basic* example, note that the use of IPA notation did not have to be requested, it followed naturally from my use of IPA in the prompt.

The *Grammar* example makes use of a somewhat more detailed prompt. For example, I specify that I do not want the *kana* (Japanese syllabic script) transcribed to latin script, since I need to practice reading *kana*, but I **do** want the pronunciation hints for the *kanji* (Japanese logographic script, glyphs of Chinese origin). Such customized instructions are carefully respected by the AI; this level of customization and control according to the student’s needs is not available with any other tutoring system. Many other tutoring and learning modalities can be specified (“I would like you to emulate a flashcard system ... “).

The *Conversation* example is quite impressive. My input is replete with grammar errors, spelling errors. awkward phrasings, and English words for which I did not know the Italian. ChatGPT makes sense of it all – even though my prompt is not actually “language”. It then translates what is necessary, corrects the grammar for my intended meaning (which it understands despite the errors) – and then finally suggests an improved version of my statement, in “real” Italian before it answers my question. It is particularly attractive that the conversation is not confined to some textbook author’s interest, but can involve any topic that is currently interesting or important to *the student*.

(Note the use of a “sentinel icon” in the Japanese and Italian examples, to demonstrate that the instructions are intact and have not been pushed out of the context window in the course of a longer conversation.)

There is just one downside: there is no *guarantee* that everything is factually correct. Though I have not observed signifcant errors, errors are possible and a fact-checking culture applies to this as much as to any other generated text.

One Principle

AI & Academy
Education

Have the AI think with you, not for you.

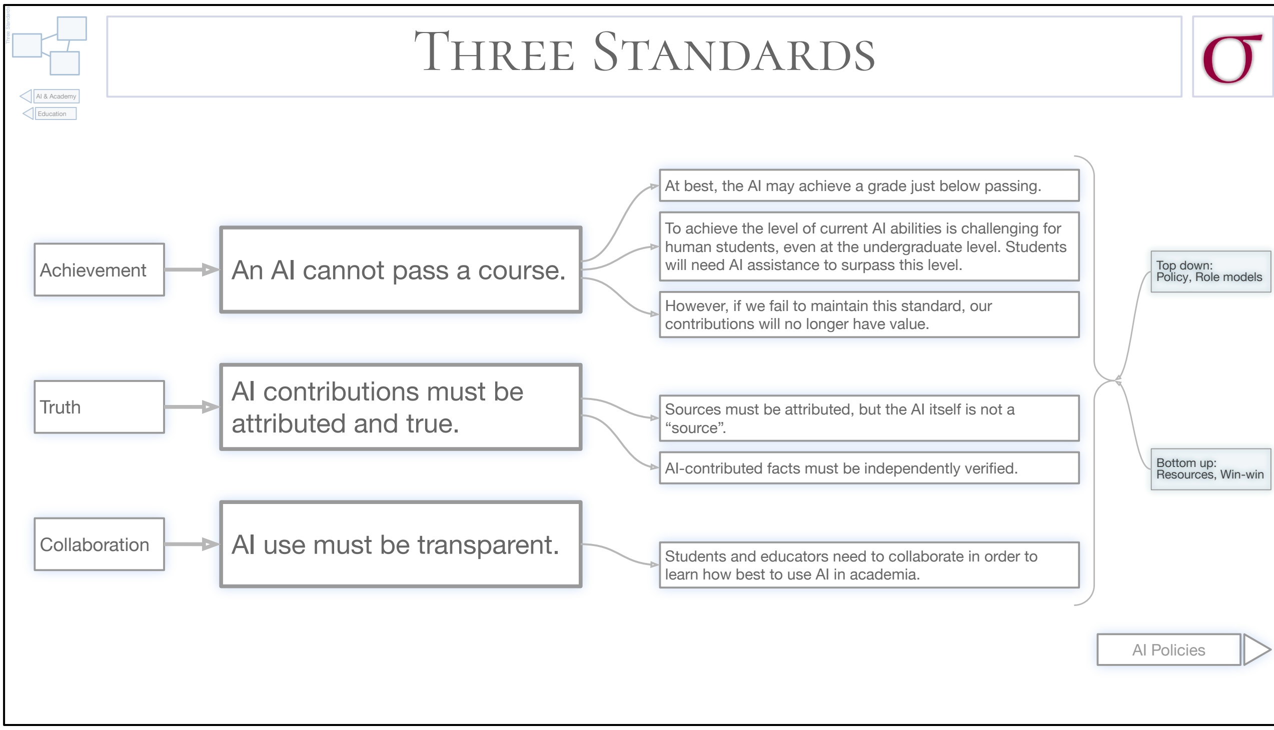
AI – Self and Other
AI Policies

A common theme ties together strategies to benefit from AI technologies, and not get trapped in counterproductive patterns. It can be condensed into one principle: have the AI think with you, not for you.

- Thinking with you implies: *your* objectives are pursued, and it is *your* agency that drives the interaction;
- Thinking with you implies that you start from a basic grasp of the problem you are trying to solve, and you are using the AI to enhance your understanding. As a result, you learn through the interaction;
- Thinking with you brings added value to your activity, and the result remains authentically yours;
- Thinking with you defines an AI that *amplifies* you and does not *replace* you.

This idea is tied to a view of AI-generated material as emergent products of a composite that is neither wholly *self*, nor wholly *other* (cf. The AI Mind – Self and Other).

Conversely: thinking *for you* diminishes you, prevents you from learning, treats tasks as meaningless rituals, treats the AI as an *other*.



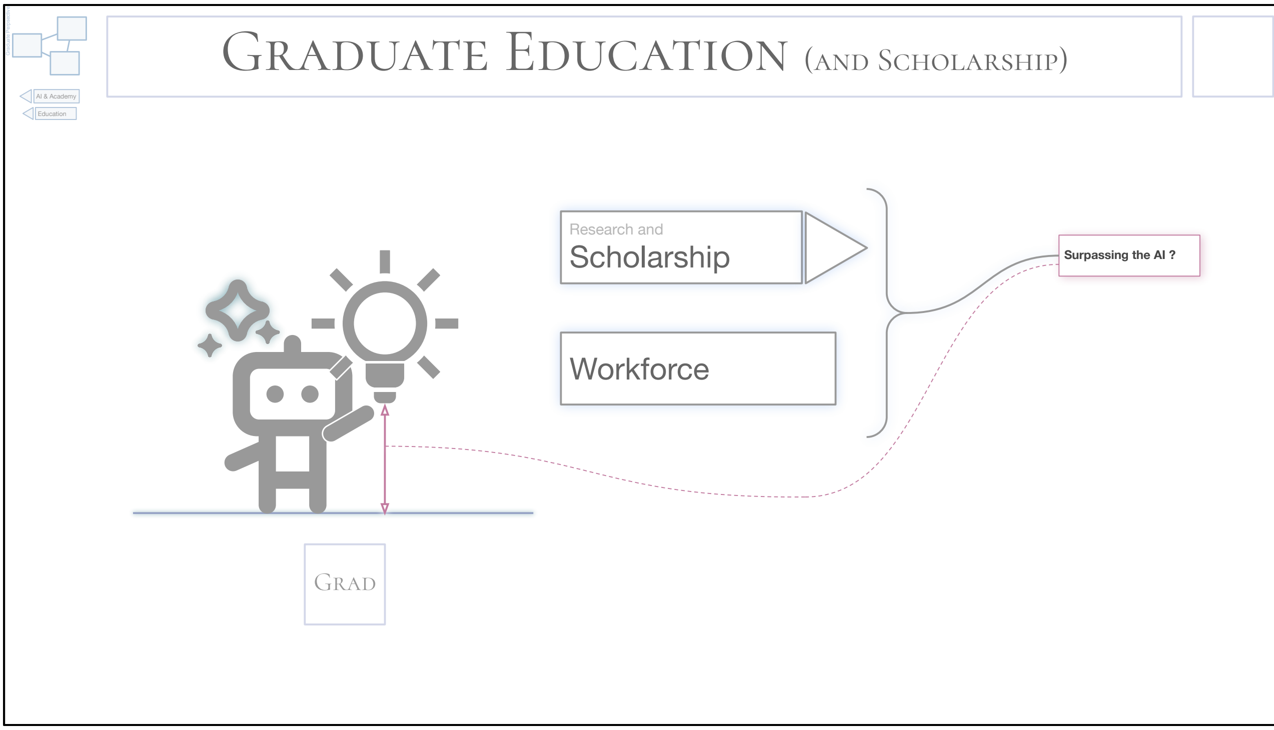
An AI cannot pass a course. We must define that an entirely AI generated contribution can at best only achieve a grade just below a pass. One might object that that makes our courses significantly more difficult than they used to be, and that is correct – but also not. First: this is correct – but we have no choice. Either we learn to surpass the AI, or the AI becomes a competitor. Second: it may not be correct after all: the requirements will be higher, but the AI can help to make it less difficult to achieve them. That will depend on how we put this into practice.

AI contributions must be attributed and true. Whatever contribution the AI makes, the final product must attribute the source of the ideas, and the contribution must be factually correct. That is trivially true in scholarship, but with generative AI it is not: the source of the AI's ideas may be challenging to identify, and sometimes the AI is absolutely and confidently incorrect. That's not even rare (cf. "Schrödinger Facts").

AI use must be transparent. Whether students or educators, we are in this together. If we assume everything is AI produced, cheating becomes impossible – and once that is out of the way, we can focus on what matters: education.

Two approaches come to mind to implement such standards in a meaningful way: top-down, we need supportive policies and successful role models – both students and educators. Bottom-up we need to supply resources, and we must construct win-win constellations through which both students and educators are motivated to work within this framework. Coercive approaches will fail since AI contributions and human contributions cannot be clearly distinguished.

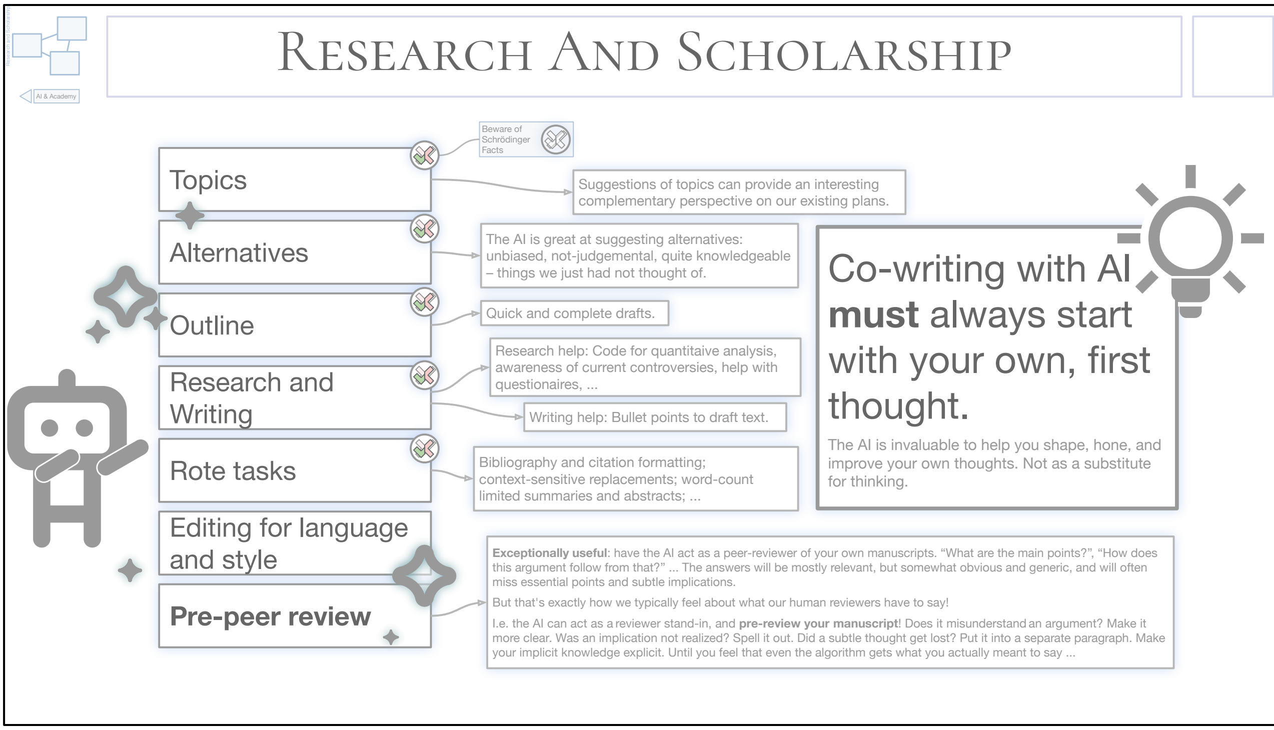
(Read more on Substack.)



The goal of graduate studies is to lead students to independent work in scholarship and research. Questions that concern coursework itself are not different from considerations for undergraduates. But the issue of *independence*, and the demands of research and scholarship are quite different. Moreover, we need to prepare graduates for the workforce. Both themes converge in that they require students to develop competence (and confidence) in **surpassing the AI**.

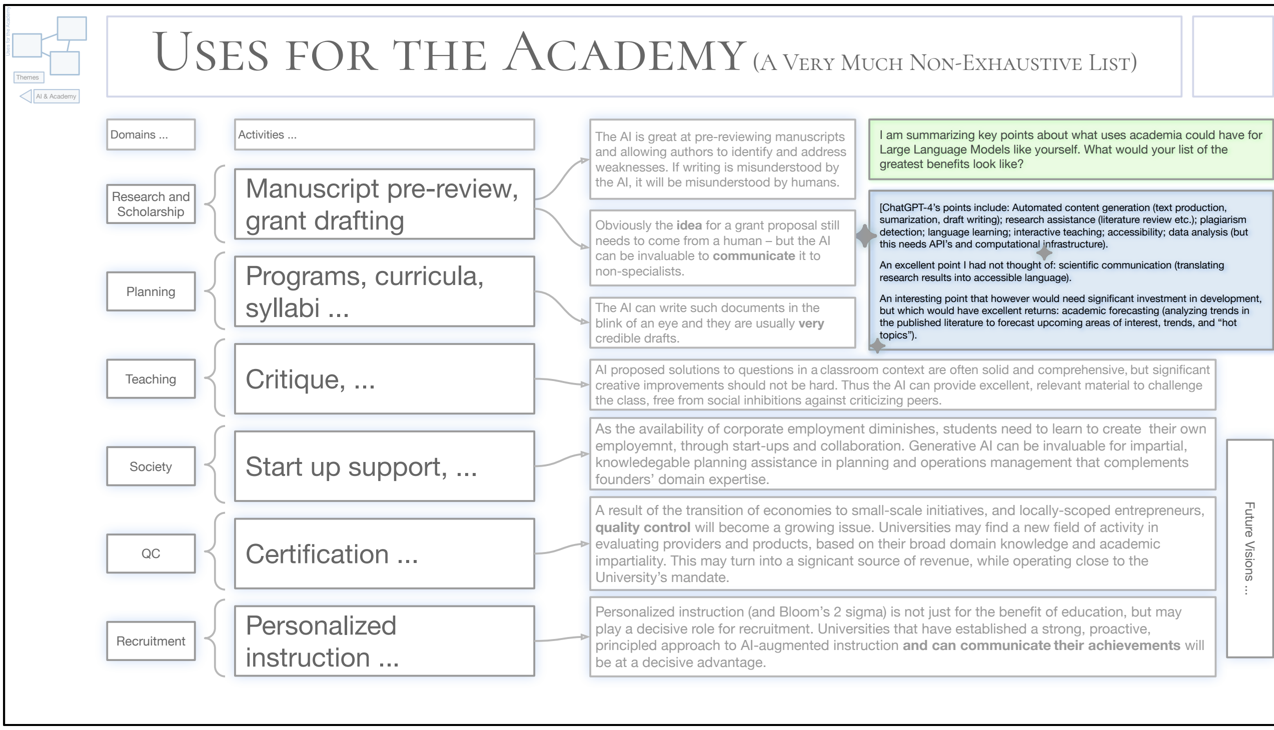
At the risk of being repetitive:

1. Surpassing the AI is essential. Not doing so means that activities no longer have value.
2. Surpassing the AI is possible: we need to learn to stand on the AI's shoulder.
3. Surpassing the AI also implies: to educate society (and peers) about the extra value that we are creating. Unless society (and peers) are ready to acknowledge (and value) this contribution, creating it is merely – as it were – academic.



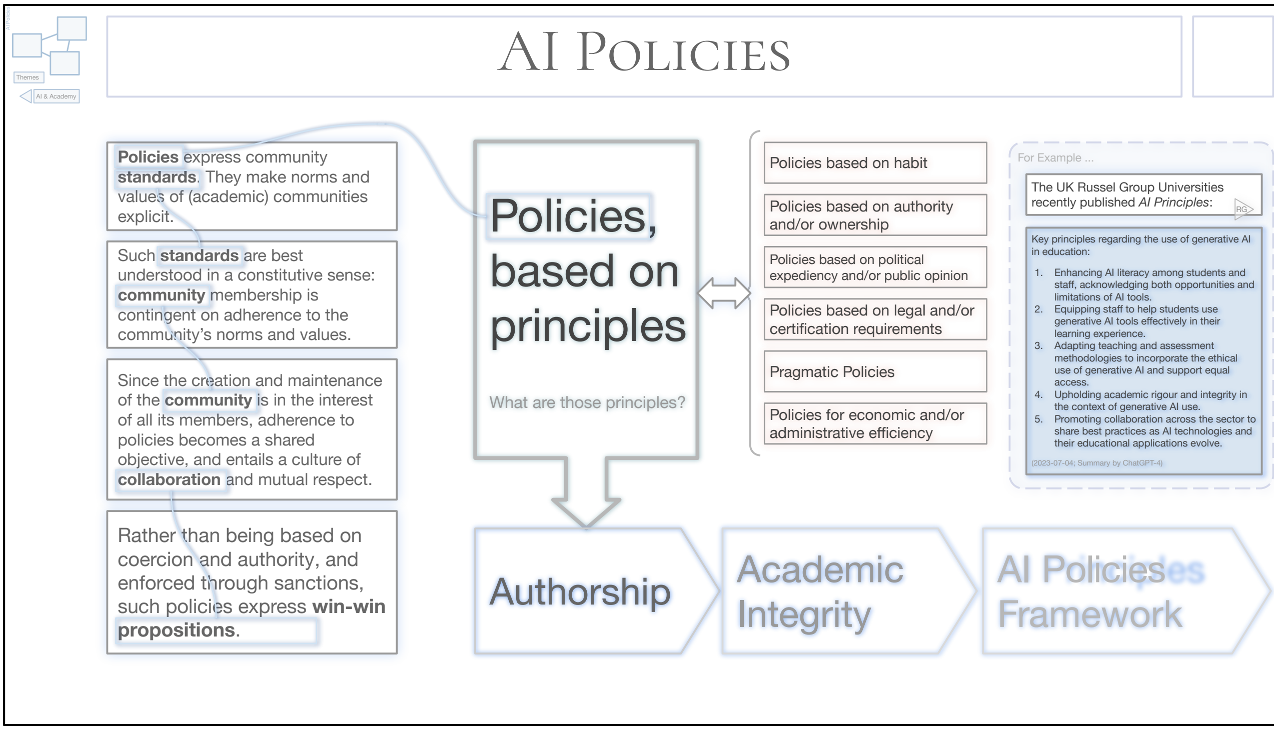
Generative AI can play a role in all phases of research and scholarship. However, one must take care not to use it too early.

“In order to become available to thought, the facts of knowledge must be connected to other facts, and their retrieval must be practiced. The crucial part is the formation of associations: understanding is an aesthetic phenomenon of harmonizing associations. [...] I mentioned the importance of priming a framework of understanding with a first thought, but consumption of synthesized solutions, rather than struggling with their construction, removes a crucial moment of practice, prevents the formation of associations, and prevents engagement, that arises, for example, when we encounter ambiguities and resolve them on our own – or accept them. [...] we need to emphasize that [...] the availability of answers cannot substitute for learning to answer. “Too much assistance” comes too early, and substitutes for engagement. (Steipe, 2023 “How much is too much” - Substack - Sentient Syllabus)”.



A small sample of the diverse, potential and current areas of application of Generative AI in academia.

I am not including an explicit perspective on **risks** here, because benefits need to pursued in a positive sense, and while doing so, arising risks should be recognized and mitigated. Risk mitigation is rarely in and of itself a sound strategy for shaping the future.



A principled view of academic policies would treat them as an expression of *constitutive* norms and values of communities. "Constitutive" refers to the fundamental principles, standards, and beliefs that *define* a group or community. Such norms and values are not just guidelines for behavior, nor are they arbitrary, or merely administrative; they actually *constitute* the community and shape its identity and culture. Adherence to those standards is a precondition for membership.(1) This principled view is not always realized in practice – I have listed some alternative foundations in the slide, and good policies will always balance a plurality of requirements. But the response to new Generative AI, requires new policy – and drafting such policies from a constitutive perspective, based on our most fundamental values and beliefs, will help make the policies justified and consistent, and thereby transform them from instruments of authority into guidelines for mutuality and collaboration.

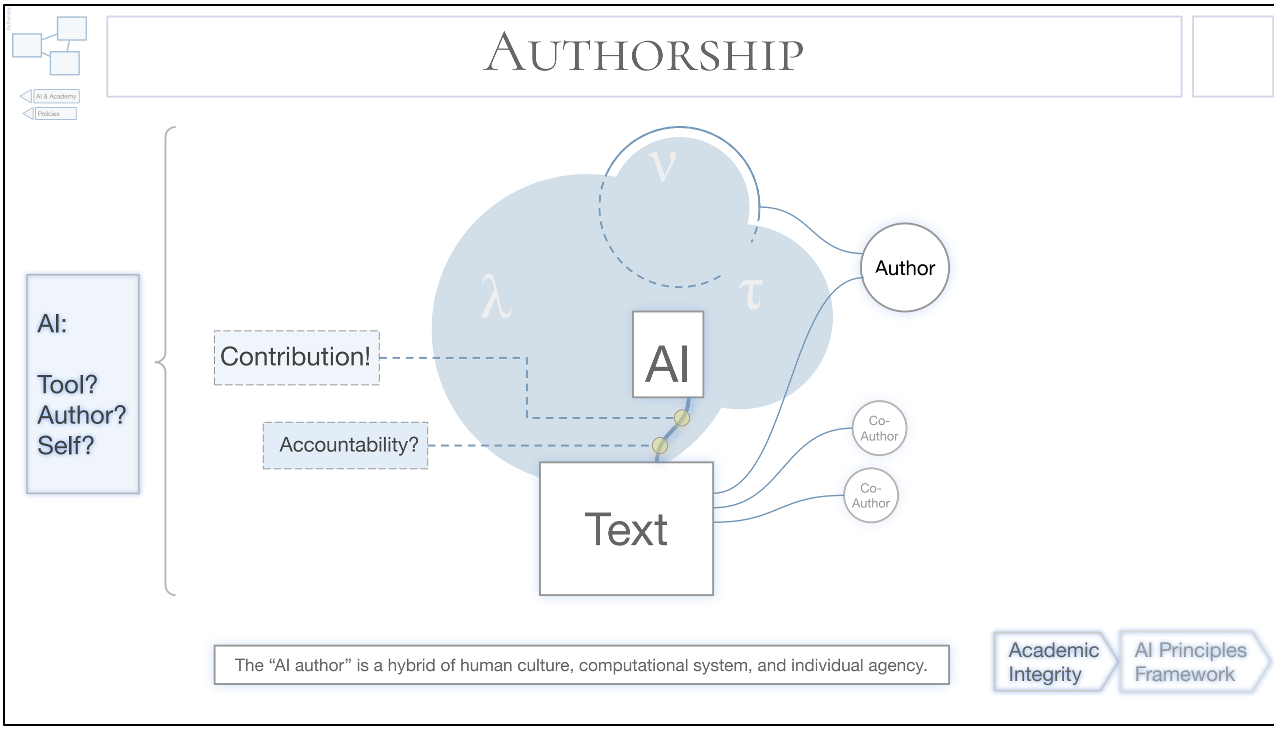
The overarching principle was expressed already 2,500 years ago in Confucius' Analects: "Control them with rule, govern them with punishment, the people will be deceitful and have no shame. Lead them with virtue, unite them in Lǐ禮 [propriety], there will be honor and principle.[道之以政，齊之以刑，民免而無恥；道之以德，齊之以禮，有恥且格。]" (Analects 2.3; cf. Chen and Steipe 2022).

Recently (2023-04-07), the 24 UK "Russell Group" Universities published a set of "Principles on the Use of Generative AI tools in Education" (LINK) backed by all 24 Vice Chancellors. It is encouraging to see this highly relevant group moving towards positive, productive guidelines and I note that this document expresses very significant changes in some of member universities' initial responses to Generative AI. These are certainly valid starting points, although there is also clearly scope for improvement. Above all, in the spirit of what I wrote above, leaving implementations up to individual divisions at various levels would indeed require a clear expression of the "fundamental values and beliefs" that could ensure consistency – which the document is however lacking.A strong point of the principles is their call for continuous revision and establishment of Communities of Practice with broad stakeholder input. That said, without a carefully thought-out mechanism, the inertia of such multi-institutional documents is typically very high.

In the following slides, we derive a revised concept of authorship, propose an principles-based approach to Academic Integrity, and arrive at a sketch elements of an AI Principles Framework.

(1) The nature and scope of appropriate sanctions follows directly: it is the removal of rights and privileges associated with community membership, proportionate to the transgression.

CHEN Yi and STEIPE Boris (2022) "Existential Reciprocity: Respect, Encounter, and the Self from Confucian Propriety (Lǐ 禮)". *The Journal of East Asian Philosophy* 2: 13–33 (DOI).



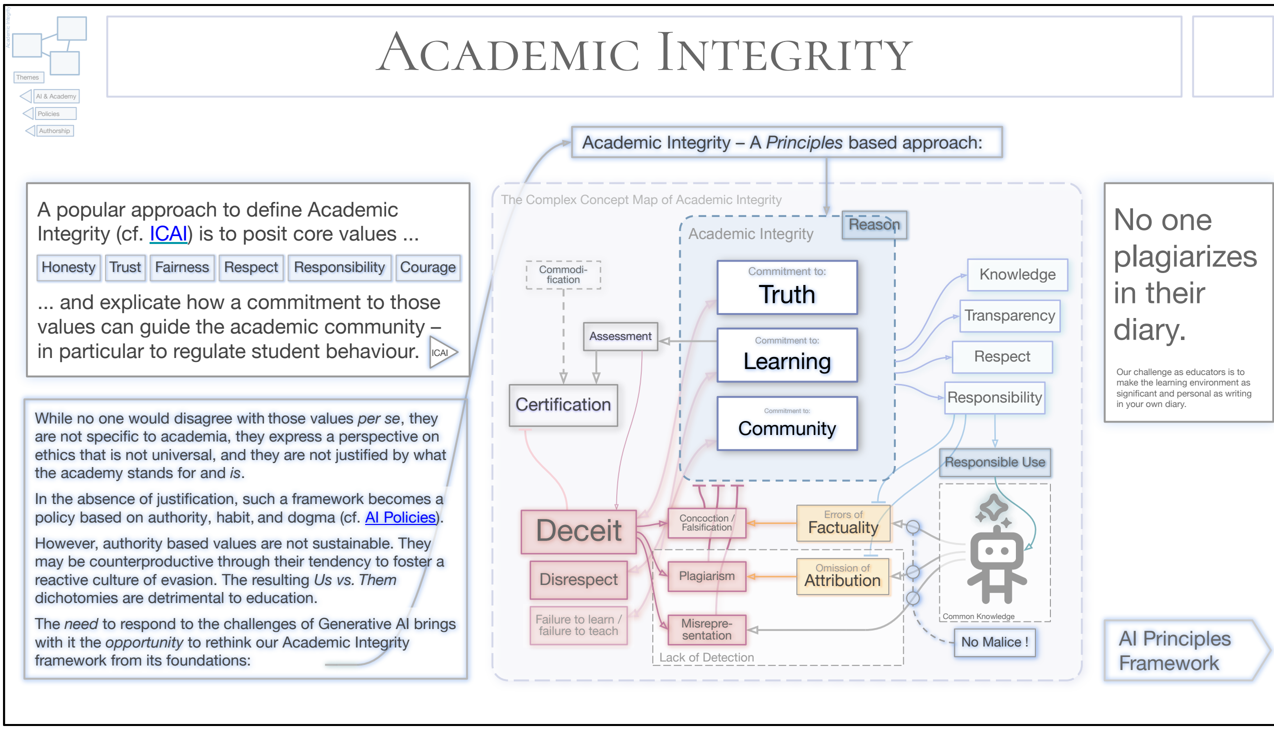
The question of authorship derives from the questions: What *is* Generative AI relative to us? What is the status of generated text? Is the AI the author? Can the AI be a co-author of scholarly work?

If we claim that AI writing is just a tool, like using a text-processing software, and therefore an AI does not qualify as a co-author, and at the same time using text written by an AI can constitute academic misconduct, those two positions are not entirely consistent.

To qualify for authorship, we require *substantial contributions* and *accountability* for the published form. Notwithstanding that this is more than what is often contributed by human (co-) authors, a Generative AI system *cannot be accountable* for its generated text since it has no agency. This lack of agency is the only criterion under which we can argue against authorship of AI systems, which is rather surprising. But it requires us to re-think our position relative to the AI system. The hybrid nature of an AI that takes its agency from a credited author (cf. The AI Mind) makes a clear separation impossible. This affects how we understand both contribution and accountability of authorship. But is also affects the basis on which we could conceivably call the use of generated text "plagiarism" – i.e. the question of Academic Integrity.

For a detailed discussion, see Steipe (2023).

STEIPE, Boris (2023) "Silicone Coauthors". Sentient Syllabus 2023-01-27 (Link).



The problems posed by Generative AI for academic integrity are not new. What is new is that AI produced work is easy to obtain, and generally not detectable as such. At first glance, this appears to be a circumscribed issue – but it turns out that the resulting challenges are impossible to address in our prevalent authority based framework. What is needed instead is a *principles based approach*, that promotes an understanding of Academic Integrity that is collaborative and balances the needs of scholars, learners, and institutions. Community constitution and community values must be causally linked and these links must be transparent and explicit.

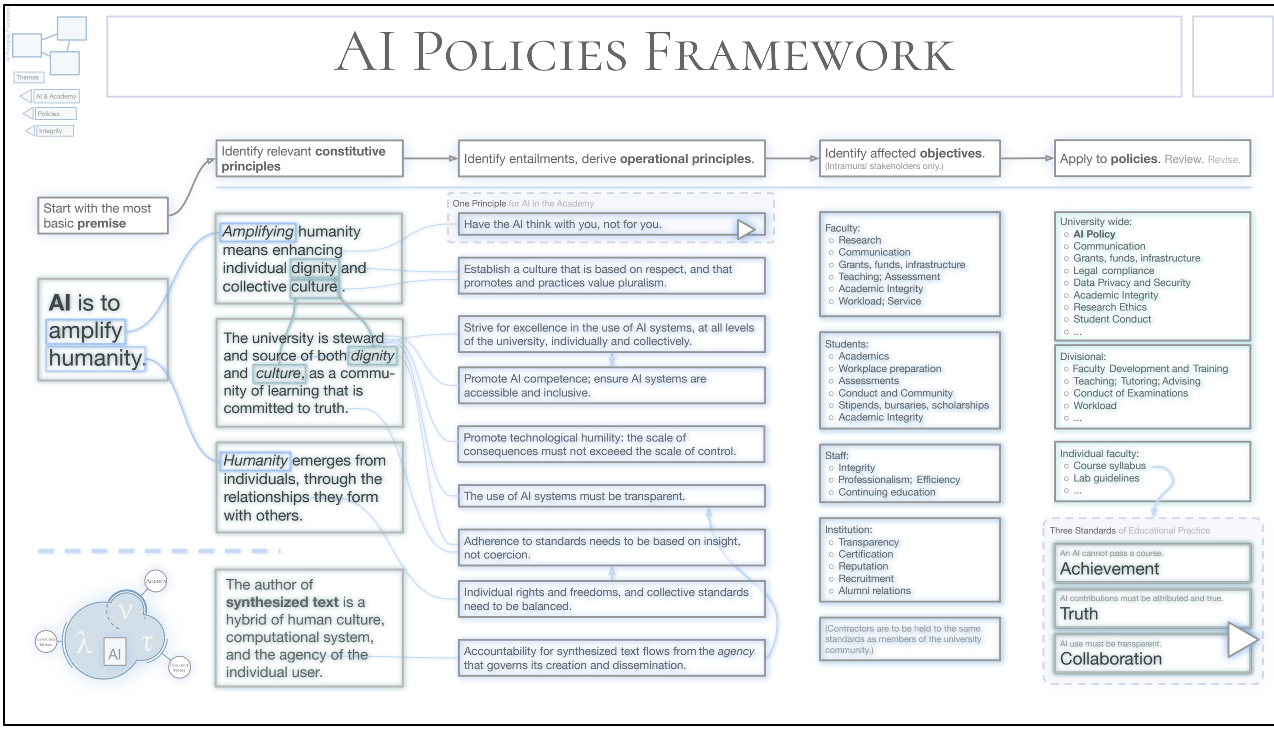
The concept map above is somewhat preliminary. It bases Academic Integrity on constitutive principles that express *what* the academy is: a community of learning that is committed to truth.(1) Values such as knowledge, transparency, respect, and responsibility derive. *Responsibility* governs the responsible use of Generative AI: to uphold the core commitments, and as a direct consequence a commitment to the basic principle: have the AI think with you, not for you. It is well documented that generated text may contain errors of fact, and may omit proper attribution. The AI systems are not capable of malice in this respect, but it is part of the responsibility of the user to fact-check and attribute. If this is not done, or if the generated text is misrepresented as one's own, academic misconduct arises. The primary dimension of academic misconduct is deceit, often in response to anxieties about assessment. The different forms that deceit takes are familiar to us, although note that the inappropriate use of generated work is generally not plagiarism.(2) Deceit is a violation of the commitment to truth. *Disrespect* becomes an academic offence where it violates community standards – which might include disrespect of the dignity of other community members, but as an offence it also includes disrespecting the obligation to know the standards, which includes, for example, standards of attribution. Thus disrespect is a violation to a commitment to be a member of the academic community. Finally, by symmetry, we would expect a violation of a commitment to learning to figure prominently in our recognized academic offences. However I am not aware that this is being done. I would understand a *failure to learn* to constitute an academic offence in this framework, but I would absolutely add – again by symmetry – a *failure to teach*. What this could mean in practice will need to be explored.

In all cases, given the nature of generative AI, we must depart from principles that are based on a transactional relationality, and on authority. Instead, both in principles and practice we need to seek collaborative solutions – win-win constellations that provide for the needs of scholars, learners, and the institution. (For an earlier but more extensive discussion, see Steipe (2023).)

(1) The academy as a community defines its standards of membership. For example, a rejection of anti-humanist tendencies could certainly be justified as an expression of community values, and thereby integrated openly and without prejudice with scholarly inquiry.

(2) Instead, we typically have a case of misrepresentation: making use of unauthorized aid and concealing that fact.

STEIPE, Boris (2023) "Generated Misconduct". *Sentient Syllabus* 2023-02-07 (Link)

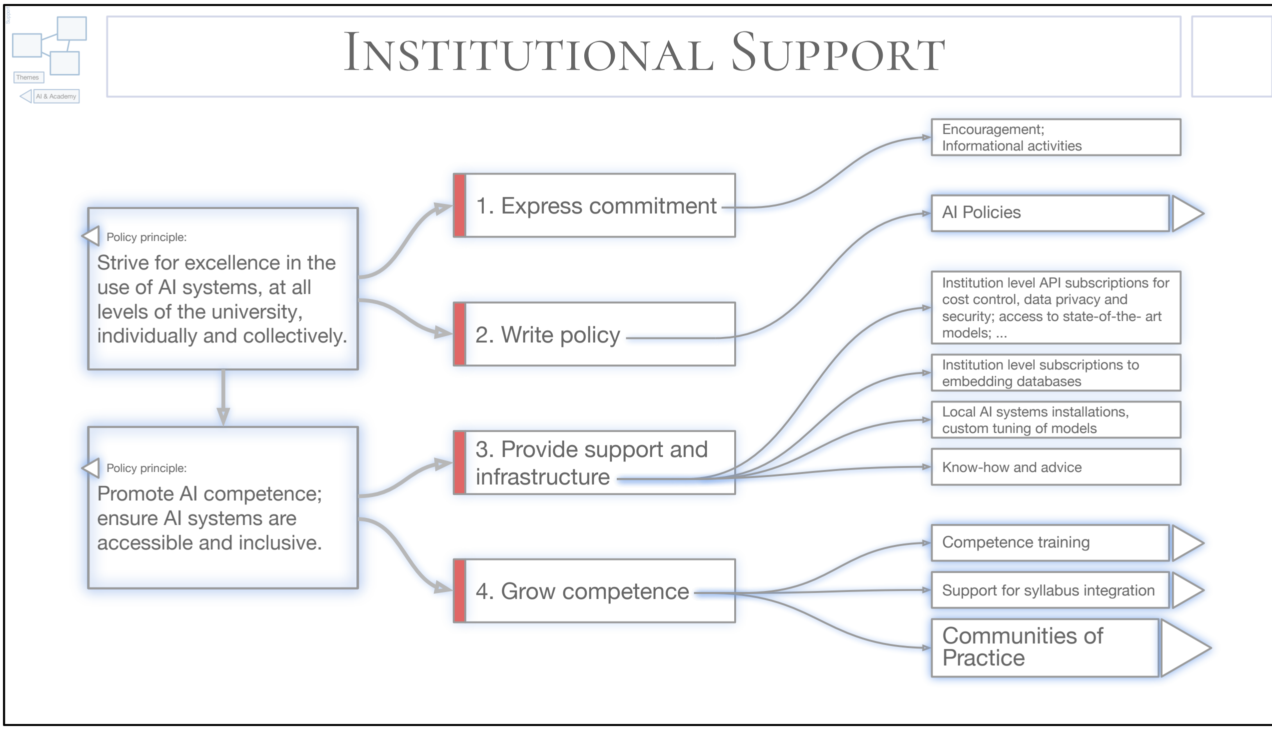


Although a somewhat preliminary sketch, this "AI Policies Framework" covers the key ideas that allow to construct and justify policies for Generative AI in the university. The most fundamental premise is that "AI is to amplify humanity". This has been expressed by many stakeholders, for example also by OpenAI CEO Sam Altman (2023); it is a rather self evident requirement of technology, and it is much more suitable as a foundation of operational guidelines than the "AI for Good" postulate that we often encounter. But the terms need to be defined. I have expressed them as "constitutive principles": principles that derive from the nature of "amplification", "humanity", the role of the university, and the resulting relationships. A number of entailments derive and these can be expressed as operational maxims. Taken together, these principles map out the general contours of university policy; they relate to each other and they can be consistently justified. The university has a crucial role, not just as a place of learning, but as a source and protector of the values that underpin society. Note that this list is not exhaustive, but it forms an argument chain back to the foundational premise of "amplifying humanity". Further objectives can be derived, such as the amplification of human cognitive abilities; assistance with truth commitments; promotion of individual dignity through education; enhancement of critical thinking; and many more.

To arrive at actual policies, we can identify affected objectives of our stakeholders: research ethics, communication of results, and grant-writing are among the concerns for faculty; academics and workplace preparedness are student concerns, and, obviously, the entire complex of academic integrity; the institution needs to devise a basis for provision of required competence and excellence initiatives and technological support, but also carefully consider implications for accreditation and certification, and so on.

The actual policy landscape of universities comprises hundreds of documents and the most reasonable approach will be to start from a general **AI Policy** that summarizes the principles and objectives of our response to the new era, that promotes the adoption of its standards in other policy as that comes up for review and revision during the normal governance cycles, and which is itself continuously reviewed and revised by its authors.

ALTMAN, Sam (2023) *Planning for AGI and beyond.* OpernAI Blog 2023-02-24 (Link).

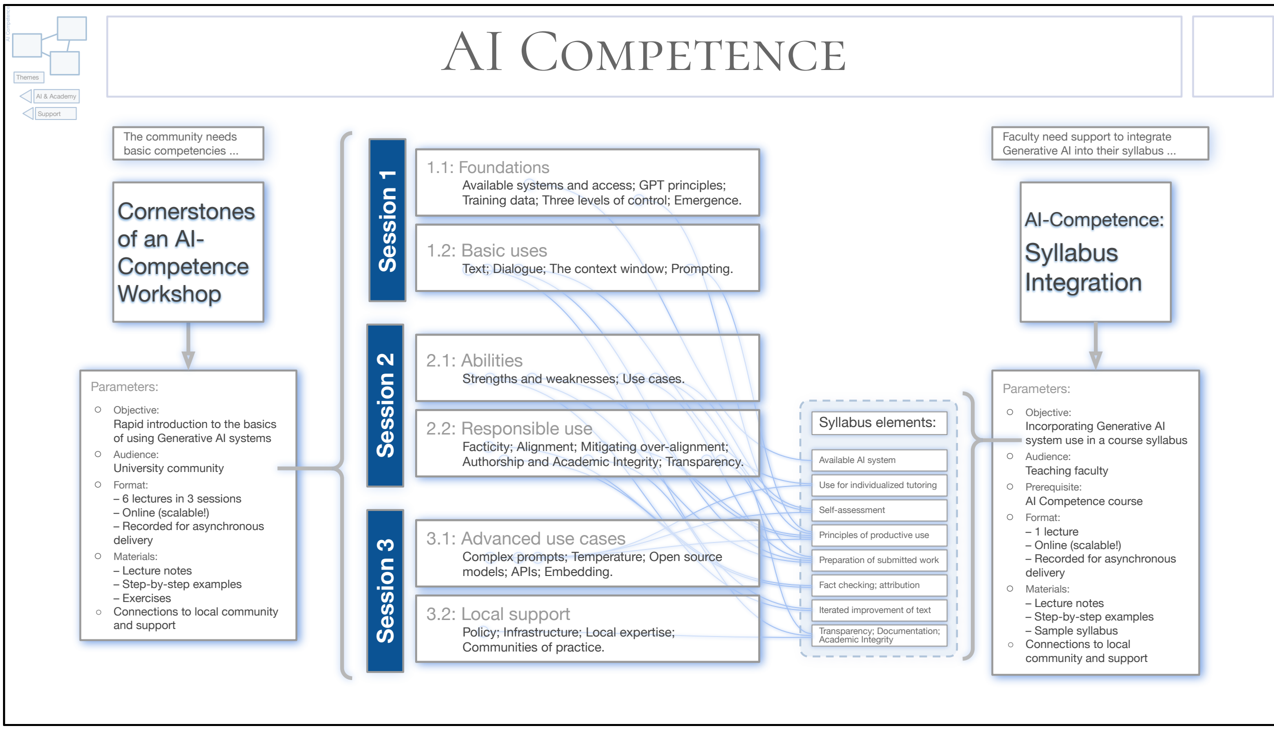


The importance of encouragement in providing institutional support should not be underestimated. Simply advertising to the community that the institution is taking a positive position on Generative AI tools, and is actively working on determining the parameters, will contribute a lot to shaping attitudes.

Providing technical support is highly desirable, but needs to go hand in hand with developing faculty needs. Major technological investments carry a risk of rapid obsolescence. Investments into human capital at all levels are more promising. We will see strong efforts by private sector vendors to provide turn-key solutions. On one hand, this may address the current lack of in-house expertise. On the other hand, these invariably are designed to lock-in customers and may turn into significant technical liabilities. Given the dynamic nature of the field, flexibility translates directly into value.

An institution-scale approach to technical support will be able to mitigate against a proliferation of questionable practices of individual users, in particular regarding legal obligations (copyright etc.), privacy, and data security. Moreover, negotiated access to AI providers on an institutional level will improve equity of access to these key technologies. Establishing a single point of information sharing and coordination will be key to a successful transition (cf. Communities of practice).

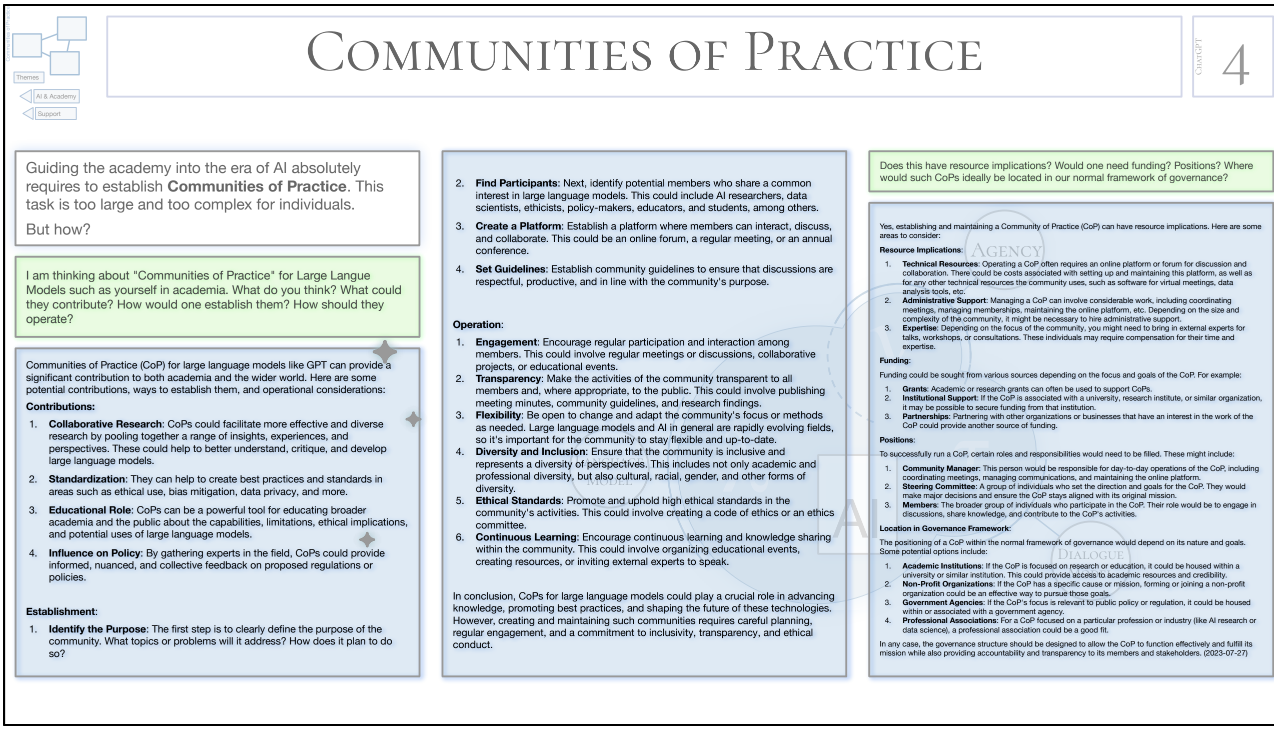
All this is new, and we need to rapidly grow competence. Competence training for the entire community – students, postgraduates, and faculty – is required with some priority. Ideally, the university would offer an online course outside of the normal curriculum. Faculty need support and best-practice advice to integrate AI competence into their syllabi. Crucially, local expertise – case-by-case augmented with domain-expert edvice – needs to be established through communities of practice.



For in-house needs, a quick course on basic competences will be beneficial. Details will depend on the local landscape, but a few considerations are listed here. Such a course needs to be scalable and asynchronous - to respond to the needs of the community.

For student needs, I would advise against dedicated courses (except, of course, courses that target the actual development of AI systems). Rather, AI use needs to be integrated into existing syllabi wherever possible. The goal is not to become AI experts (for most of us). The goal is to become domain experts with AI competence. Dedicated courses do not offer the required breadth and flexibility.

The map above list parameters of a foundations course or workshop, and parameters of a follow-on course for faculty, to guide the inclusion of AI use and AI competence elements in existing syllabi.



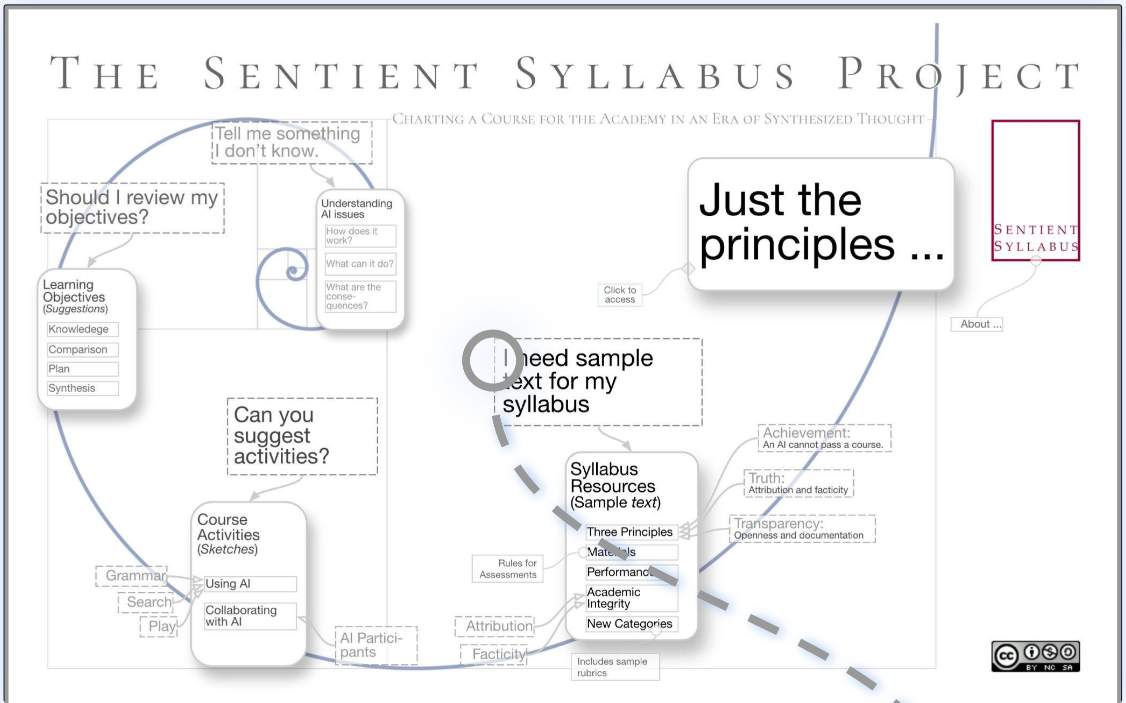
A major issue with AI and the academy is that everything is completely new, and everything is in flux. We have no established knowledge that we can build on, the domain expertise of experts is frequently not helpful since it tends to be biased towards previous paradigms (we are indeed witnessing a paradigm shift in Thomas Kuhn's sense, a development that is analogous to a scientific revolution), and yesterday's knowledge may have been superseded by new developments today. We must make things up as we go, but to do this well may well take up a much larger portion of time than we have available to dedicate to this task. Moreover, we certainly do not want to operate in a merely reactive mode. These challenges require visions, and proactive, anticipatory activities.

This is a problem that makes the establishment of communities of practice – forums to collect experience, establish institutional memory and awareness, and exchange advice – indispensible.

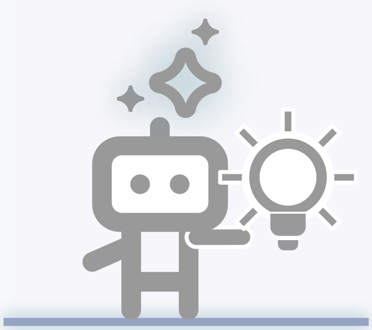
In the slide, I illustrate an example of AI assisted planning through ChatGPT's responses. By and large, ChatGPT-4's advice overlaps well with my own prior notes. I would have spent more effort in defining scope, scale and terms, but, for example, an explicit focus on transparency and diversity (I would prefer *value pluralism* here) is good advice that I did not have on my list. The follow up question raises a few issues. The AI seems a bit confused about what I meant – I am not talking about governance of the CoP, but interfaces with our established frameworks: who does a CoP inform and advise, and through which channels? Or is it only for the benefit of its members? And clearly, when working with the AI one has to be very careful about task fixation and especially about scope creep (!).

The Sentient Syllabus Project

http://sentientsyllabus.org



Synthesized Thought



The availability of human-level, synthesized thought, virtually for free, changes everything ...

To chart a course for academia into this future we need:

– Resources
– Analysis
– Communities

https://sentientsyllabus.substack.com/

I founded the Sentient Syllabus Project in December 2022 as a response to ChatGPT, the appearance of multiple other Generative AI resources, and the need for direction and guidance for academia. I see this project as an international collaborative effort. Resources are linked from http://sentientsyllabus.org and I write (semi-regular) analysis on Substack: https://sentientsyllabus.substack.com/

Contact …

For Academic Correspondence: boris.steipe@utoronto.ca

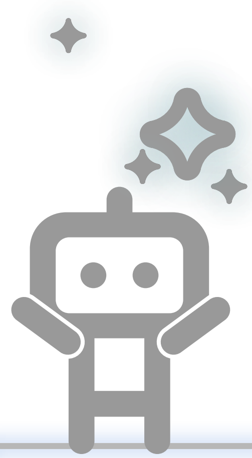
For Analysis – Subscribe (free) and Read: **Sentient Syllabus**
https://sentientsyllabus.substack.com/

For Sentient Syllabus correspondence: sentient.syllabus@gmail.com

For Micro-blogging: **Mastodon**
https://qoto.org/@boris_steipe

For Scholarship: **ORCID**
https://orcid.org/0000-0002-1134-6758

For Networking: **Linkedin**
https://www.linkedin.com/in/boris-steipe-047399260/



Questions, feedback, and comments are most welcome.