Posudek bakalářské práce Allegry Stodolsky: Walter Benjamin’s “aura” in the age of AI

Vypracovala: Mgr. Dita Malečková, Ph.D.

Práce Allegry Stodolsky je nejen originálním příspěvkem do debaty o umělé inteligenci (AI) a odkazu Waltera Benjamina - ale především do debaty o vztahu mezi nimi. Samozřejmě, že Walter Benjamin neměl žádnou zkušenost s tím, čemu souhrnně říkáme AI - a co autorka stručně, ale přesně popisuje, a to včetně rozpaků a obav s tímto fenoménem spojených. Ale je to právě autenticita výroků AI, která je v hledáčku mnoha kritiků - a zdá se, že dílo Waltera Benjamina je i ve věku AI dobrým nástrojem pro detekování autenticity díla.

Základní otázka tedy zní: Může být dílo vytvořené pomoc AI autentické?

Autorka zvolila zajímavou metodu: co se stane, když do GPT (Generative Pretrained Transformer) vložíme tři dokumentární díla o životě Waltera Benjamina a necháme generovat nový výstup? Výstupy dále pečlivě analyzuje (jde především o poměr fakticky správných faktů, nesprávných a čistě fiktivních). Nabízí se však otázka: Je skutečně logická a faktická správnost výstupů měřítkem jejich autenticity?

Text se k odkazu W. Benjamina staví věrně a citlivě, ale přiznávám, že mě nejvíce zaujala poznámka v jeho závěru o novém typu aury (“digi-aura”). Jakkoli je totiž kritika generovaných materiálů na místě - nejsou koherentní tak, jak jsme zvyklí u lidských textů, jsou “mimo realitu a fikci” atd., jsou zde jiné klíčové rysy jako například citlivost na počáteční podmínky, respektive vstupy - tedy jistá věrnost zdroji, která by mohla být kritériem jejich hodnocení - hodnocení, které dle mého názoru musí být odlišné od posuzování lidských textů.

Doplňující otázky:

1) Dalo by se v nějakém smyslu říct, že díla generovaná pomocí velkých jazykových modelů (LLM) mají jiný typ autenticity, která má jiné zdroje a kořeny?
2) S tím souvisí i "rituálnost" těchto postupů: což jsou mechanismy spojené s rituály zcela transparentní a k dispozici racionálnímu rozvažování? Naopak bych řekla, že práce s AI vykazuje rysy "archaických" praktik, například praxe orákula.

Poznámky:

1) Mimochodem další model, GPT-4, vykazuje oproti GPT-3.5 značný pokrok, bylo by zajímavé vyzkoušet v rámci experimentu různé typy neuronových sítí.

2) A ještě k části textu: "Nobody, not even AI itself is able to explain how algorithms of the neural networks function."
   Open AI nechalo GPT-4 analyzovat fungování GPT-2 a došli k velmi zajímavým výsledkům, jakkoli samozřejmě naprostá transparentnost není docela dobře možná, vzhledem k objemu dat a rychlosti fungování těchto nástrojů. Je však jisté, že "vysvětlitelnost" postupů AI je žádoucí a snad proveditelná i v lidském měřítku.
   Link:
   https://openai.com/research/language-models-can-explain-neurons-in-language-models

Doplňující otázky mají sloužit k další debatě o tomto zajímavém tématu, nejsou znakem nedostatečnosti textu, který je soudržný a přehledně a logicky vystavěný kolem originálního nápadu. Práci doporučuji k obhajobě a navrhuji hodnocení B, nadprůměrný výkon s minimem chyb.

V Praze, 4. září 2023 Mgr. Dita Malečková, Ph.D.

English version

Allegra Stodolsky: Walter Benjamin's "aura" in the age of AI

Allegra Stodolsky's thesis is not only an original contribution to the debate on artificial intelligence (AI) and Walter Benjamin's legacy - but above all to the debate on the relationship between the two. Of course, Walter Benjamin had no experience with what we call AI - and what the author accurately describes, including the embarrassments and fears associated with the phenomenon. But it is the authenticity of AI statements that is in the crosshairs of many critics - and Walter Benjamin's work seems to be a good tool for detecting the authenticity of a work of art, even in the age of AI.

So the basic question is: Can a work created with the help of AI be authentic?

The author has chosen an interesting method: what happens if we put three documentary works about Walter Benjamin's life into a GPT (Generative Pretrained Transformer) and let it generate a new output? She then carefully analyses the outputs (mainly the ratio of factually correct, incorrect and purely fictional facts). However, the question arises: Is the logical and factual correctness of the outputs really a measure of their authenticity?

The text treats W. Benjamin's legacy with fidelity and sensitivity, but I confess that I was most interested in the remark in its conclusion about a new type of aura ("digi-aura"). The criticism of the generated materials is without any doubt adequate and necessary - they are not coherent in the way we are used to in human texts, they are "beyond reality and fiction", etc., but there are other key features such as sensitivity to initial conditions, or inputs - that is, a certain fidelity to the source that could be a criterion for their evaluation - an evaluation that, in my opinion, must be different from the evaluation of human texts.

Additional questions:

1) Could it be said, in some sense, that works generated by large language models (LLMs) have a different type of authenticity that has different sources and roots?

2) Related to this is the "ritualistic" nature of these practices: are the mechanisms associated with rituals completely transparent and available to rational deliberation? On the contrary, I would say that working with AI exhibits features of "archaic" practices, such as the practice of the oracle.

Notes:

Incidentally, the next model, GPT-4, shows considerable progress over GPT-3.5; it would be interesting to try different types of neural networks as part of the experiment.

One more part of the text: "Nobody, not even AI itself is able to explain how algorithms of the neural networks function."
Open AI had GPT-4 analyze the workings of GPT-2 and they came up with some very interesting results, though of course complete transparency is not quite possible given the volume of data and the speed at which these tools work. However, it is certain that "explainability" of AI procedures is desirable and perhaps feasible on a human scale.
Link:
https://openai.com/research/language-models-can-explain-neurons-in-language-models

The additional questions are intended to further debate on this interesting topic, they are not a sign of inadequacy of the text, which is coherent and clearly and logically built around an original idea. I recommend it for defense and suggest a grade B, an above average performance with minimal errors.

Prague, September 4, 2023 Mgr. Dita Malečková, Ph.D.