Natural Language Processing vs Large Language Models: this is the end of the world as we know it, and I feel fine

Bertrand De Longueville
European Commission
Joint Research Centre
Brussels, Belgium
bertrand.de-longueville@ec.europa.eu

Abstract

As practitioners in the field of Natural Language Processing (NLP), we have had the unique vantage point of witnessing the evolutionary strides leading to the emergence of Large Language Models (LLMs) over the past decades. This perspective allows us to contextualise the current enthusiasm surrounding LLMs, especially following the introduction of "General Purpose" Language Models and the widespread adoption of conversational chatbots built on their frameworks. At the same time, we have observed the remarkable capabilities of zero-shot systems powered by LLMs in extracting structured information from text, outperforming previous iterations of language models.

In this paper, we contend that that the hype around “conversational AI” is both a revolution and an epiphenomenon for NLP, particularly in the domain of information extraction from text. By adopting a measured approach to the recent technological advancements in Artificial Intelligence that are reshaping NLP, and by utilising Automated Socio-Political Event Extraction from text as a case study, this commentary seeks to offer insights into the ongoing trends and future directions in the field.

1 WHAT – the significance of the “ChatGPT revolution” on NLP

To start this commentary in the context of a renowned workshop on Automatic Event Extraction from Text, let’s ask ourselves a rather philosophical question: what distinguishes a socio-political event of rather anecdotic importance (e.g. “Donald Trump tweeted he is unhappy about XYZ”) from an event that is most likely to mark history (e.g. “WHO Declares COVID-19 a Pandemic and calls to act in consequence”)? Natural Language Processing (NLP), our area of research, is unlikely to provide a scientific answer to this question, so I would rather propose an empirical one, a twofold method for measuring the historical significance of a socio-political event. Firstly, it is interesting to observe that individual people witnessing events that make history have very accurate personal memories of what they were doing when it happened. For example, people born in the 1950’s or earlier, when being asked what asked they did on 20 July 1969, when Apollo 11 landed on the moon, often provide a precise narrative about their activities, even decades after the fact. There is a second method to recognise key events: they render obsolete almost immediately common beliefs and thoughts that held authority before them. For example, the permanence of USSR as a political entity could have been considered as obvious for most of its citizens … until the collapse of the Berlin Wall, and subsequently of the PCUS regime a couple of years later.

I am always reluctant to use the expression “ChatGPT revolution” to designate the hype that followed, in the fall/winter 2022-2023, the launch of OpenAI’s conversational AI Chatbot powered by the GPT3.5 Large Language Model. After all, Generative Pretrained Transformer (GPT) Models are the fruit of a decades-long continuum of technical evolution, from Support Vector Machines to Deep Learning, from early efforts to perform statistical machine translation to massive training of general-purpose language models (Johri et al., 2021). If we look at the performance of NLP applications we, experts, must acknowledge that the turning point has probably occurred several years earlier, with the discovery of the Transformer architecture, unlocking the efficiency of machine learning models for natural language understanding, thanks to the mechanism of attention (Vaswani et al., 2017). But for the general public, the

Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts associated with RANLP 2025, pages 32–37, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-099-1-004

"revolution" has happened when such transformers became conversational. I must admit that when I saw GPT 2.0 generating fake research paper reviews (or even fake papers) (Bartoli & Medvet, 2020), I was not overly impressed. In my view, natural language understanding was where societally relevant use cases resided, not in the generation of ersatz human texts.

I dismissed generative LLMs, seeing them as useless stochastic parrots (Bender et al., 2021)… and the entire world proved me wrong. Of course, it is the nature of a hype is to feed itself. And the impressive uptake of ChatGPT[1] can be explained by cultural factors, rather than by its technical innovation: the myth of the talking machine, from Medieval tales of the Brazen Head to "2001: a space Odyssey" … But this is not my point. Let's simply take note that the advent of LLMs matches our twofold criteria and therefore qualifies as a "significant" historical event. Firstly because, if you ask colleagues and friends, most will be able to tell how they encountered for the first time an LLM-powered Chatbot (often, without knowing it was LLM-powered). Personally, I recall precisely the circumstances in which my hierarchical superior explained me (gently, but firmly) that, as the Head of a Text Mining Competence Centre, I could not ignore the advent of conversational AI, "as a matter of existential threat to my research team". And secondly because it made obsolete many widespread claims about AI and Language Technologies. Take for example the widely cited and seminal paper in our area of interest, from 2016 and titled "Growing pains for global monitoring of societal events" (Wang et al., 2016) : it claims – rightfully, then – "the text-processing systems used in event coding are still similar to ones developed more than 20 years ago". Could we say this about our event extraction systems in 2025? I do not think so.

2 SO WHAT – LLMs as "game changers" for NLP

In this context, we may wonder: are LLMs truly game changers for Natural Language Processing in general, and for Automated Event Extraction from text in particular? An abundant literature suggests so, which corroborates the intuitions shared in the previous section (Cronin, 2024; Törnberg, 2023; Yang et al., 2024). Let's reflect further, from an NLP practitioner's perspective, on the implications of General-Purpose LLMs for Automated Event Extraction.

At first glance, one may claim we are reaching "the end of history" (Chernyavskiy et al., 2021) for NLP… after all, LLMs act as remarkably versatile zero-shot machine learning models, being able to extract almost any relevant information from a piece of text, relying on almost human-level of text understanding in hundreds of natural languages, and on a "world model" derived from their training on a significant share of all human knowledge ever produced (in the form of millions of books, encyclopaedias, scientific articles, websites, conversations, blogs, etc.). So, "game over" for NLP scientists, let's all retrain as "prompt engineers" by practicing the art of asking the right question to General Purpose LLMs/oracles…

Well, it's not that simple.

First of all, let's not forget the inference cost aspects. In Socio-Political Event Extraction, real-life use cases often require the processing of vast amounts of raw text (typically, news articles or field reports), so the computing power to process them in near-real time can become a significant bottleneck. Based on my own experience, I would say there is a ratio of about 1 to 50, or even 100, in terms of computing power required to run a "good old" BERT-like model compared to state-of-the-art LLAMA 4 or Mistral 3.1 open weights models. Moreover, the latest models require costly and powerful GPU hardware cards that are on high demand, while BERT-like models run on older hardware that is likely to be already amortised in terms of cost, and more easily available for purchase. Literature shows that properly fine-tuned models of the BERT generation perform at very high levels for specialised tasks such as geocoding(Tanev & De Longueville, 2023), sentiment analysis (Di Nuovo et al., 2024), discourse analysis (Stefanovitch, De Longueville, et al., 2023), or topic mining (Stefanovitch, Jacquet, et al., 2023), which are all relevant for Automated Event Analysis purposes. So one may wonder: why would we need to invest in a Ferrari when we have a highly adaptable fleet of Land Cruisers at hand?

There is another reason why LLMs are not "the end of history" for NLP. If LLMs can provide an

[1] ChatGPT reached 100 million users in two month, while it took Instagram two years to reach this symbolic step (Deng et al., 2023)

answer to virtually any question, it is never guaranteed that such an answer – although remarkably crafted from a linguistic point of view – is factually correct. The problem of hallucinations is well known and widely discussed (Huang et al., 2025), but interestingly, the root causes of such behaviour are often overlooked. One of these reasons is sycophancy[2] (Malmqvist, 2024). The need to provide an answer at any cost, in order to please the interlocutor is deeply embedded in the LLM's training process, as their reward function includes some form of "user satisfaction". For this reason, even the best prompt in the world cannot completely avoid sycophantic behaviour and hallucinations. So when facts matter, like in NLP and a fortiori in its Automated Event Extraction use cases, LLMs can never be blindly trusted.

Another trustworthiness issue with LLMs is linked to their "knowledge" component: because they are so eloquent, and because they have been trained on much more information than we could possibly read in our entire lives, we assume they are almost omniscient. But in fact, the world knowledge they seem to feature is more a by-product of their next-word-prediction ability than the result of an accurate and fit-for-purpose world model. LLMs talk, they know and they even reason… but not in the exact same way we do (De Longueville et al., 2025). It is easy to arrive at a misunderstanding situation with LLMs; in brainstorming or creative use cases, that can even be an advantage. But in NLP, where the goal is precisely to extract accurate information from inherently ambiguous natural language, misuse of LLMs abilities can lead to disappointment.

To overcome the "knowledge" ambiguity of LLM's behaviours, the best solution resides in the engineer ever more complex systems that feed them with the right contextual knowledge, in a process called Retrieval Augmented Generation (Lewis et al., 2020). In the context of event extraction, a RAG pipeline can for example include some Gazetteer lookup to improve geocoding (Tanev & De Longueville, 2023).

But if LLMs "know", they also "reason": imagine a sentence like "the political meeting will take place in Zoom". A RAG-enabled AI system, designed to rigorously lookup places in a comprehensive gazetteer would probably geocode such an event in Zoom, a Village in Soreng Tehsil in West District of Sikkim State, India (lat 27.144465910353176, long 88.26329879818186), while we, humans, would rather infer "Zoom" designates an online video-conferencing platform[3]. This example shows that the "reasoning" component of LLMs cannot be blindly trusted either, even when LLMs are fed with the best data and follow the best-crafted prompt instructions. It is important to have that concept in mind, as NLP experts, since with the advent of Agentic AI systems (Chawla et al., 2024), we will increasingly rely on LLM's ability to reason.

Based on the above, one may conclude that since LLMs are not magically addressing any possible issue, then they are junk… Since we cannot trust 100% for a task, then we cannot trust them for any task – including our preferred one: extraction of spatiotemporal information patters from text. This would not be a rational approach to the promises of LLMs. As scientists, we should wonder: if I cannot trust 100% my LLM system for this task, then to what percent can I trust it? And there, we start thinking in terms of precision and recall … There we go again!

3 NOW WHAT – the new NLP that looks like the good old one

Everything changes, but nothing changes: on the one hand LLMs can act as prodigious zero-shot information extraction machines that open new perspectives for NLP applications, but on the other hand, their precision and recall need to be accurately measured…

Evaluating the performance of NLP software modules to perform specific tasks, like automatically extracting information about socio-political events from text is a classic activity for NLP scientists. It requires the creation and curation of "gold standard" corpora, where the expected outcome of a large number of instances of the same task is encoded (usually, by human annotators), and on which variants of the software module are tested, until the highest possible F-score (Derczynski, 2016) is reached, expressing the best possible compromise between precision and recall.

So, really, nothing new under the sun for NLP practitioners.

[2] According to the Cambridge Dictionary, sycophancy is defined as the "behavior in which someone praises powerful or rich people in a way that is not sincere, usually in order to get some advantage from them".

[3] This is not a fictitious case: I saw it happening.

However, understanding the underlying reasons for LLMs successes and failures to provide results corresponding to gold standards opens new research perspectives. For example, there is a need to better understand and assess spatiotemporal reasoning abilities of AI systems based on LLMs, and how formal ontologies (like gazetteers for place names, or named entities databases) can complement LLM's internal (and perfectible) world model for tackling hallucinations and supporting entities disambiguation. In other words, there is a need to further explore hybrid approaches aiming at developing NLP processing pipelines that involve LLMs, advanced Retrieval-Augmented Generation technique and more deterministic approaches like rule-based on symbolic AI components. Interesting developments have been recently published that go in that direction for geoparsing (Halterman, 2023) or epidemic events detection (Consoli et al., 2024). These are examples to follow while exploring other epistemic tasks related to event extraction..

Also, while the founding principles of NLP task-specific evaluations remain valid, the scientific methods to measure the efficiency of non-deterministic AI pipelines executing complex event-extraction processes remain to be studied, thus paving the way for next-generation socio-political event extraction research. Inspiration could come from similar – but distinct – language technology research areas. For example, studying how to improve the knowledge extraction component of an LLM pipeline with a carefully engineered RAG component (Ceresa et al., 2025), or by developing integrated "software + datasets" bundles to well targeted task evaluation of specialised NLP software packages (Bassani & Sanchez, 2024).

To achieve scientifically reproducible results in this novel area of research for LLM-ready NLP, it is essential that academic organisations have the ability to run their own LLM inference systems. With the trend of ever larger LLMs[4], and given the IT infrastructures constraints described above, there is an increasing trend to rely on LLM-as-service provided through Application Programming Interfaces. This creates an additional difficulty for scientists, as such models, often provided commercially, do not fully disclose their detailed systems specifications (e.g. input filters, output filters or system prompts which have a proven strong impact on an AI service behaviour (De Longueville et al., 2025), and may change without prior notice, making previous NLP task evaluations obsolete. As a consequence, LLM-as-service is even more a black box than any Deep Learning model, as the LLM itself is surrounded with undisclosed technical components that influence its output.

It is thus a matter of independent science – and ultimately of Sovereignty – that Academic organisations remain capable of fully controlling the execution environment of the LLMs they base their research on. The availability of state-of-the-art open weight LLMs is therefore crucial for academia, and will become of paramount importance as the advent of Agentic AI will introduce novel paradigms for knowledge workers, and among them scientists especially, for interacting with data and information, using AI systems as "mediators" (e.g. when using an LLM-powered tool to perform systematic literature reviews).

In the light of the above, we may draw this oxymoronic conclusion: for NLP, the advent of general purpose LLMs is both a revolution and an epiphenomenon.

Been there, seen that: as a geospatial scientist, I saw in the early 2000's the combination of cheaper GPS devices, pervasive Internet connections and web 2.0 technologies like AJAX lead to a paradigm shift in my research area (De Longueville et al., 2010). The release of the Google Earth to the wide public in 2005 embodied this revolution for the general public and created a hype similar to the one around ChatGPT nowadays. Faced with such technologies enabling interoperable analysis and visualisation of geospatial data on a smooth Digital Earth interface, some may have wondered: is it the end of history for geospatial sciences? Yet, this research area remains vibrant 20 years later, increasing our Earth Observations capabilities and refining our common understanding of complex planetary phenomena.

Will the same happen to NLP with the advent of LLMs and Agentic Systems in the 2020's? In other words, dear NLP scientists, are you ready to cope with the rollout at large scale of "GPS and Digital Earth, but for the knowledge"? Your answers to these questions will shape the future of NLP in the next decades.

[4] Although this trend is perceived as plateauing already (Villalobos et al., 2024), the size of current top-performing models (which is only based on assumptions for commercial, non-open-source models) already exceeds the IT infrastructure capacity of most Universities and Research Centres for running them at large scale for inference.

References

Bartoli, A., & Medvet, E. (2020). Exploring the Potential of GPT-2 for Generating Fake Reviews of Research Papers. In *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/faia200717

Bassani, E., & Sanchez, I. (2024). Guardbench: A large-scale benchmark for guardrail models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18393–18409.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Ceresa, M., Bertolini, L., Comte, V., Spadaro, N., Raffael, B., Toussaint, B., Consoli, S., Piñeiro, A. M., Patak, A., Querci, M., & others. (2025). Retrieval Augmented Generation Evaluation for Health Documents. *arXiv Preprint arXiv:2505.04680*.

Chawla, C., Chatterjee, S., Gadadinni, S. S., Verma, P., & Banerjee, S. (2024). Agentic AI: The building blocks of sophisticated AI business applications. *Journal of AI, Robotics & Workplace Automation*, *3*(3), 1–15.

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers: “The End of History” for Natural Language Processing? In *Lecture Notes in Computer Science* (pp. 677–693). Springer International Publishing. https://doi.org/10.1007/978-3-030-86523-8_41

Consoli, S., Markov, P., Stilianakis, N. I., Bertolini, L., Gallardo, A. P., & Ceresa, M. (2024). Epidemic Information Extraction for Event-Based Surveillance Using Large Language Models. *International Congress on Information and Communication Technology*, 241–252.

Cronin, I. (2024). *Decoding large language models: An exhaustive guide to understanding, implementing, and optimizing LLMs for NLP applications*. Packt Publishing.

De Longueville, B., Annoni, A., Schade, S., Ostlaender, N., & Whitmore, C. (2010). Digital Earth’s Nervous System for crisis events: Real-time Sensor Web Enablement of Volunteered Geographic Information. *International Journal of Digital Earth*, *3*(3), 242–259. https://doi.org/10.1080/17538947.2010.484869

De Longueville, B., Sanchez, I., Kazakova, S., Luoni, S., Zaro, F., Daskalaki, K., & Inchingolo, M. (2025). *The Proof is in the Eating: Lessons Learnt from One Year of Generative Ai Adoption in a Science-for-Policy Organisation*. https://doi.org/10.2139/ssrn.5141665

Deng, Y., Zhao, N., & Huang, X. (2023). Early ChatGPT User Portrait through the Lens of Data. *2023 IEEE International Conference on Big Data (BigData)*, 4770–4775. https://doi.org/10.1109/bigdata59044.2023.10386415

Derczynski, L. (2016). Complementarity, F-score, and NLP Evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)* (pp. 261–266). European Language Resources Association (ELRA). https://aclanthology.org/L16-1040/

Di Nuovo, E., Cartier, E., & De Longueville, B. (2024). Meet XLM-RLnews-8: Not Just Another Sentiment Analysis Model. In *Natural Language Processing and Information Systems* (pp. 24–35). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70242-6_3

Halterman, A. (2023). *Mordecai 3: A Neural Geoparser and Event Geocoder*. https://arxiv.org/abs/2303.13675

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In *Lecture Notes in Networks and Systems* (pp. 365–375). Springer Singapore. https://doi.org/10.1007/978-981-15-9712-1_31

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-

Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Malmqvist, L. (2024). *Sycophancy in Large Language Models: Causes and Mitigations*. https://doi.org/10.48550/ARXIV.2411.15287

Stefanovitch, N., De Longueville, B., & Scharfbillig, M. (2023). TeamEC at SemEval-2023 Task 4: Transformers vs. Low-Resource Dictionaries, Expert Dictionary vs. Learned Dictionary. *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2107–2111. https://doi.org/10.18653/v1/2023.semeval-1.290

Stefanovitch, N., Jacquet, G., & De Longueville, B. (2023). Graph and Embedding based Approach for Text Clustering: Topic Detection in a Large Multilingual Public Consultation. *Companion Proceedings of the ACM Web Conference 2023*, 694–700. https://doi.org/10.1145/3543873.3587627

Tanev, H., & De Longueville, B. (2023). Where “where” Matters: Event Location Disambiguation with a BERT Language Model. In A. Hürriyetoğlu, H. Tanev, V. Zavarella, R. Yeniterzi, E. Yörük, & M. Slavcheva (Eds.), *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text* (pp. 11–17). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.case-1.2/

Törnberg, P. (2023). *How to use LLMs for Text Analysis* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2307.13106

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Position: Will we run out of data? Limits of LLM scaling based on human-generated data. *Forty-First International Conference on Machine Learning*.

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science*, *353*(6307), 1502–1503. https://doi.org/10.1126/science.aaf6758

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1–32. https://doi.org/10.1145/3649506