# NLP’s Wars of the Roses: how GPT-3 sparked a fierce divide in language model research



Between 2020 and 2022, the natural language processing (NLP) research community experienced a significant period of internal conflict and transformation, often referred to as “The Wars of The Roses” within the field. This period was marked by intense debates over the capabilities and limitations of large language models (LLMs), epitomised by the release of OpenAI’s GPT-3 model in June 2020, which significantly advanced the field's technical landscape.

The debates centred around the fundamental question of whether such models genuinely “understand” language or simply mimic it through statistical patterns. Emily M. Bender, a professor in linguistics at the University of Washington and 2024 president of the Association for Computational Linguistics, played a key role in voicing scepticism about claims of true understanding in LLMs. Alongside computational linguist Alexander Koller, Bender co-authored a peer-reviewed paper introducing the “octopus test” analogy, which argued that models trained only on linguistic form without grounding cannot truly comprehend meaning. Bender described the relentless public arguments on this topic, noting the repetitive nature of the discourse: “It seemed like there was just a never-ending supply of people who wanted to come at me and say, ‘No, no, no, LLMs really do understand.’”

Opposing this viewpoint was Julian Michael, who recognised the importance of the discussion at the time, though he later described critical responses to Bender and Koller’s paper. Other researchers, such as Ellie Pavlick from Brown University, characterised this period as a reckoning for the field, where long-standing assumptions were challenged.

The release of GPT-3 marked a watershed moment. The model's unprecedented scale—over 100 times larger than its predecessor—and capabilities shocked many researchers. Christopher Callison-Burch, a professor at the University of Pennsylvania, described the experience of using GPT-3 as a “career-existential crisis,” witnessing years of student research seemingly replicated in a moment. The model’s ability to learn new tasks from natural language prompts in just a few lines of text was described by researcher Nazneen Rajani as “mind-boggling,” despite some early issues with safety and bias. Some analysts initially sought to debunk GPT-3 as a mere “party trick,” but, as noted by Christopher Potts, an eventual admission emerged that its capabilities were far beyond superficial feats.

Nevertheless, the field was divided over the implications of GPT-3’s proprietary nature and the commercialisation of LLM research. Sam Bowman observed a schism between the industry-driven development of models, which were accessed through paid APIs, and the academic NLP community, which traditionally prioritised open scientific methods and reproducibility. This division was echoed by Anna Rogers and Julian Michael, with concerns expressed about the shift from fundamental research to what some considered “product testing” or “API science.”

During this time, ethical and social concerns also came to the forefront. Timnit Gebru, an AI ethics researcher then at Google, collaborated with Emily Bender to highlight the potential dangers of increasing LLM sizes without addressing critical issues such as bias and environmental impact. Their paper, “On the Dangers of Stochastic Parrots,” sparked a civil war-like debate across the community. As Kalika Bali from Microsoft Research India recalled, this period saw polarised reactions within the NLP community, with some researchers deeply stressed by the factionalism.

As the debates grew sharper, younger researchers like Julie Kallini reflected on the difficulty of navigating the polarised landscape, where established figures in the field took opposing stances. Liam Dugan, a Ph.D. student at the University of Pennsylvania, noted the pressure to align with one side in order to pursue lasting research influence.

By mid-2022, the tension was such that the NLP community conducted a broad survey on controversial positions within the field, including the role of linguistic structure and whether scaling alone could solve core problems. The results revealed a fragmented field: a traditional linguistic camp sceptical of raw scaling approaches, a middle-ground faction, and an optimistic group that believed scaling could lead to general intelligence. As Dugan reflected, initial dismissal of the more radical views changed with the emergence of ChatGPT, which demonstrated practical applications of GPT-3-style models to the public.

Sam Bowman highlighted a key source of division: limited interaction between industry research advancing large-scale models and academic NLP, leading to misunderstandings about each group’s goals and achievements. This disconnect contributed to the crisis atmosphere, described by Julian Michael as a “field in crisis,” underscored by the difficulty in reconciling experimental method, commercial priorities, and foundational linguistic theory.

In summary, the 2020-22 period in NLP research was characterised by profound debate, rapid technological advances, and growing pains as the field sought to balance scientific inquiry with the realities of commercial-scale AI development and ethical considerations. These tensions both reflected and shaped the evolving landscape of language model research and its broader implications for artificial intelligence.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.noahwire.com> - This is the source of the article discussing the internal conflicts and transformations in the NLP research community between 2020 and 2022, often referred to as 'The Wars of The Roses' within the field.
2. <https://www.time.com/6308275/emily-m-bender/> - This article provides insights into Emily M. Bender's critical analysis of large language models and her skepticism about claims of true understanding in LLMs, aligning with her role in voicing skepticism about LLMs' capabilities.
3. <https://en.wikipedia.org/wiki/Stochastic_parrot> - This Wikipedia page details the 'stochastic parrot' analogy introduced by Emily M. Bender and Alexander Koller, which argues that models trained only on linguistic form without grounding cannot truly comprehend meaning.
4. <https://www.macloo.com/ai/category/nlp/> - This article discusses the debates over the capabilities and limitations of large language models (LLMs) in the NLP community, particularly focusing on the release of OpenAI’s GPT-3 model in June 2020.
5. <https://www.jeanmonnet-code.eu/2023/04/15/you-are-not-a-parrot/> - This article discusses the controversy surrounding the 'On the Dangers of Stochastic Parrots' paper, co-authored by Timnit Gebru, and the subsequent reactions within the NLP community.
6. <https://www.macloo.com/ai/category/nlp/> - This article explores the ethical and social concerns that emerged during the NLP community's debates, including issues related to bias and environmental impact, as highlighted by Timnit Gebru and Emily Bender.
7. <https://news.google.com/rss/articles/CBMilgFBVV95cUxOcjM1S2RoWWtWSExma3BFR0hUenYxcjN1WWZGUVNWcXUzTldJRlZGeHZrNTRMeWZTVUdzSUpsc3FBbU4wcmRRM3c5Zm5pdXR5d0JIVmE1alNKUUdFZ0RrVXZWaGFpTXgyajBNTUtqU2VTOWVFWHhkckxfYUtwVVJwYTNpXzduZ2EzUm1LRFRYYmhZZEE3eUE?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data