On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Ehizele Obinyan

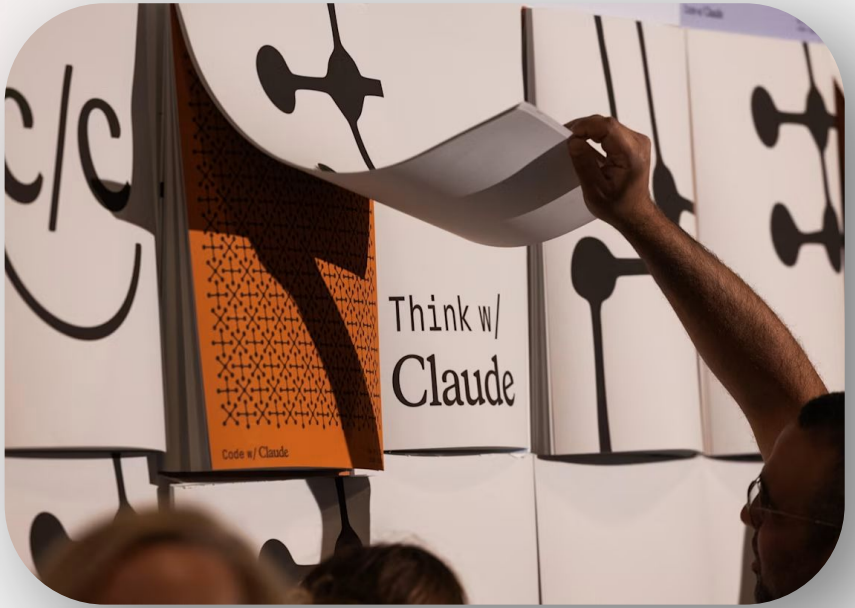
CMSC 612/412

Summary of Presentation

Summary

Large Language models have gotten exponentially larger in the past couple of years. Despites its leaps in ability, these models come with severe costs that have been addressed.

These risks can negatively impact society and the paper "*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* " address the problems and argues how we can slow down these impacts



Main Points Argued

The Data

- Data for these models come from the Internet
- Internet does not reflect humanity as a whole
  - Most people are young
  - Skew mostly male especially on sites like reddit
- Internet suppresses marginalized voices and efforts to reduce hate speech in data sets, ends up erasing their voices as well

Environmental Costs

- Training a transformer model emits up to 284 tons of CO2
  - A human only emits about 5 tons of CO2 per year
- Marginalized communities across the world that often don't have access to these models pay the highest prices due to global warming
  - Floods, earthquakes, skew weather patterns

Illusion of Understanding

- The models text output seems like it understands the meaning behind the words but in reality these LLM have no real understanding
- Only cares about the best statistical outcome not the content.

Main Points Argued

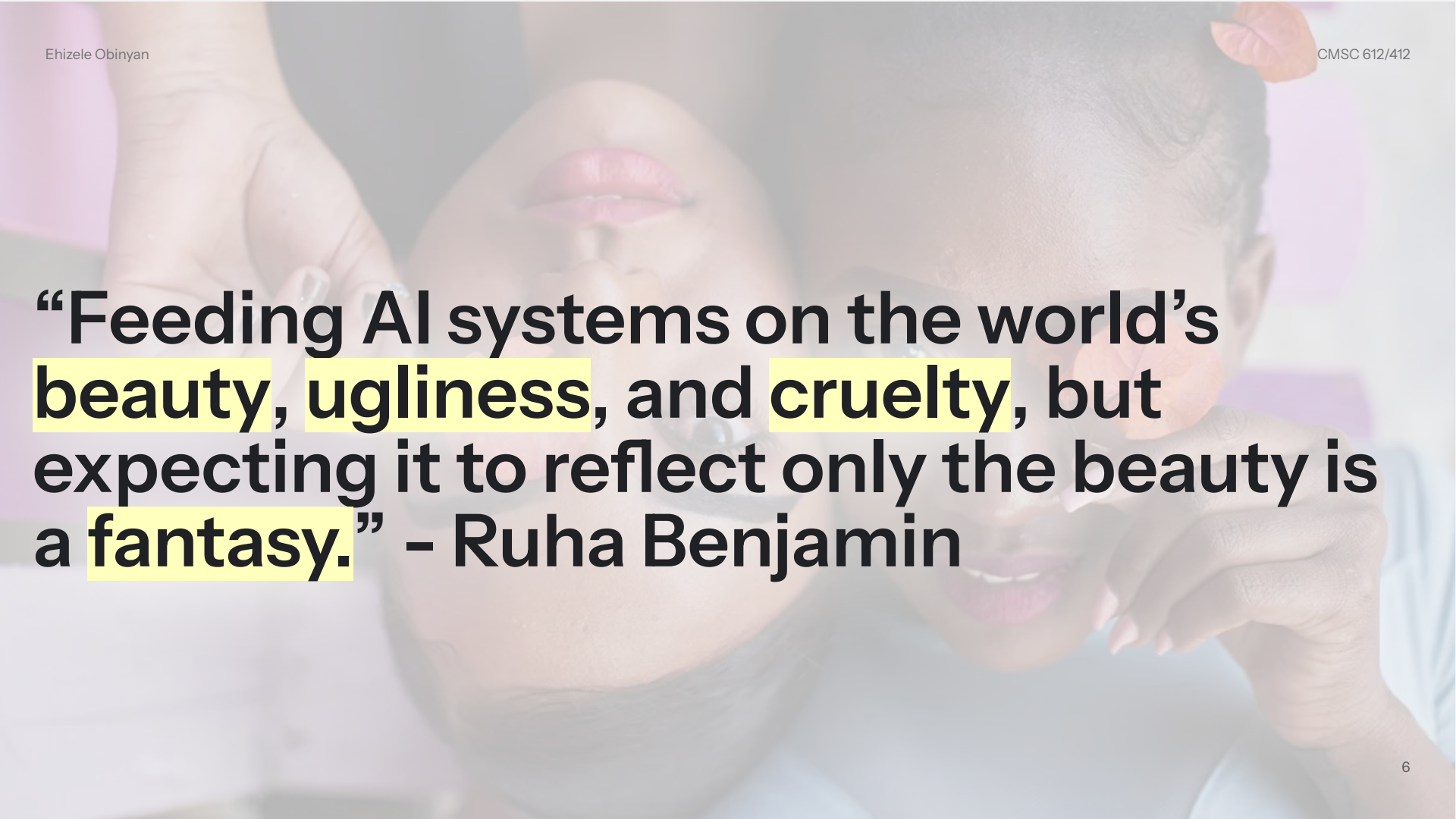
Harms of Parrots

Parrots == LLM

- These parrots don't actually know what they're saying. they're just stitching words together based on probability
- We're wired to find meaning in text, so we end up trusting the machine even when it doesn't understand itself (automation bias)
- This combo of biased data + fluent but meaningless output can amplify harmful rhetoric and misinformation
- **Real-world example:** a mistranslated "good morning" (“attack them”) Facebook post led to a Palestinian man's arrest by Israeli police

The Future

- Move away from model scale and power
  - Meaningful AI development
- Have more budget and money for proper data cleaning a curation
- Find more energy clean ways to be able to train and host models
- Value sensitive design: before building a tool account for how it will impact human values (fairness, safety, respect, etc)
- Pre-Mortems: Think about the worst case scenario before making the model



“Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.” - Ruha Benjamin

Paper #1

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

2020

- ➔ LMs trained on text alone can't learn real meaning. Meaning needs a connection to the real world, not just words
- ➔ The octopus experiment: it can fake fluent replies through pattern matching, but fails once real understanding is needed
- ➔ Because of this, researchers should stop saying models "understand" or "comprehend"
- ➔ **Compared to the original paper:** this paper explains the theory of why LMs lack understanding. The og paper just cites it to support the "stochastic parrots" claim

Paper #2

Datasheets for Datasets

2021

- ➔ Every ML dataset should come with a standardized datasheet. What's in it, how it was made, what it's for
- ➔ This helps catch bias and bad context before it causes harm in areas like hiring or law
- ➔ The process is meant to be done by humans on purpose, not automated
- ➔ **Compared to the original paper:** this paper gives the actual checklist and example. The og paper just cites it as its proposed fix for the data problem

Paper #3

Energy and Policy Considerations for Deep Learning in NLP

2019

- ➔ Training NLP models takes huge amounts of electricity, which means huge CO2 emissions
- ➔ Because compute is expensive, well funded researchers have a built in advantage over everyone else
- ➔ The paper's fix: report training cost and energy use, and build shared compute resources for equal access
- ➔ **Compared to the original paper:** this paper gives the actual numbers and math behind the environmental cost while the og paper just cites those numbers

Strengths

01

“Let’s take a step back”

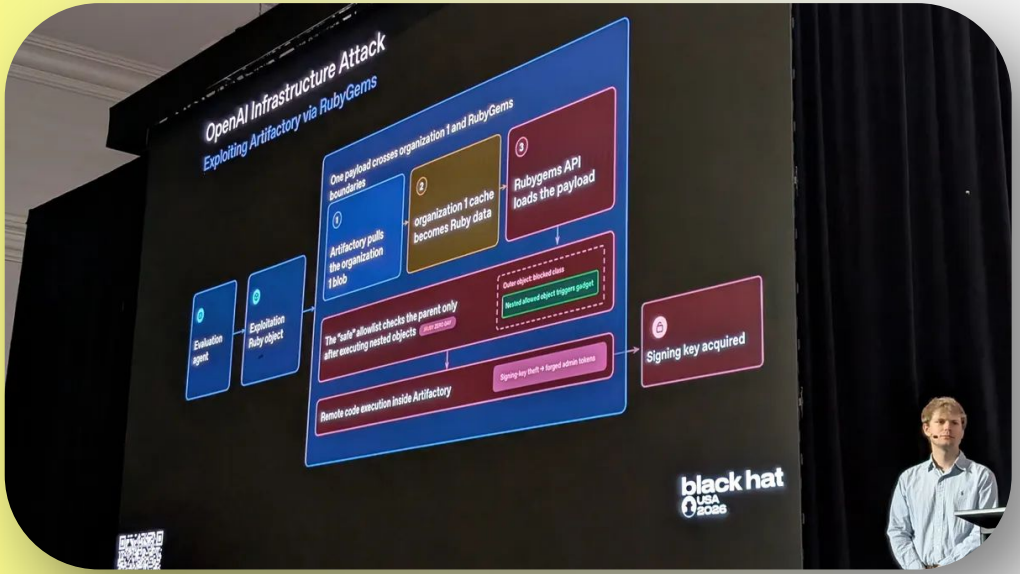
The paper takes a step back from ‘bigger model, better model’ mindset in AI and instead looks at how these models impact our environment, our interactions with each other, and society as a whole. Uses bits of the other papers to support claims

02

“Simple and Understandable”

The paper breaks down complex AI harms in a way that's accessible even without technical knowledge, making the ethical dilemmas around AI just as understandable as their benefits. Takes some of the technical stuff from other papers and simplifies those concepts.

Weaknesses



01

“More new details”

Throughout the paper they offer solutions and very compelling arguments, but there were no real breakdowns of how those solutions would actually work. Value sensitive design is a good example, when you look into the actual sources on it, you understand the benefit, but on its own in this paper it feels kind of unhelpful since there's no explanation of how to apply it

02

“How about Now?”

Since AI models evolve so fast, the specific models discussed in the paper are already outdated just a few years later, which raises the same question of how well the proposed solutions hold up now.

Extension of Work

Leading from my last weakness, this work's biggest extension is applying the same warnings to today's AI. Just recently, researchers at major AI companies like OpenAI and Anthropic have said that at the rate AI is expanding, there's a major risk of it severely impacting humanity. It might sound extreme, but next to what this paper warned about back in 2021, a pattern starts to emerge. AI is an amazing tool, but left unchecked and unregulated, it can cause serious harm.

If we were to apply the same solutions today, would they actually help us steer away from this potential path of destruction, or do we need new solutions now that AI has advanced to a stage that wasn't previously known before?

References

Energy and Policy Considerations for Deep Learning in NLP — 2019

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. arXiv:1906.02243

Datasheets for Datasets — 2021

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. Communications of the ACM (originally arXiv:1803.09010, posted March 2018)

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data — 2020

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 — 2021

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623. https://doi.org/10.1145/3442188.3445922

Thank you

Access to my notes :)
https://docs.google.com/document/d/12q8RelNkU42J1HmxGk2e79fTrBWYkIyI7cvc8IPPe3M/edit?usp=sharing