

Considerations after the impact of Large Language Models and the arrival of Large Agent Models: towards a revival of multi-agent systems in AI?

Mots/Machines #6, Brest

Christophe Servan, PhD

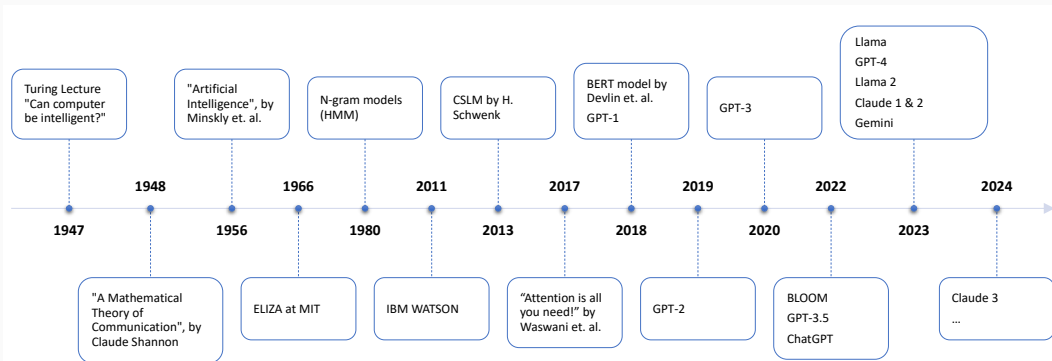
08th of March 2024

Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France, christophe.servan@lisn.upsaclay.fr

1. Introduction
2. Language Models
3. Fine-tuning of Neural Language Models
4. Declination of Neural Language Models
5. Agent systems
6. Link LLM & Agents
7. What's next?

Introduction

A bit of Hystory... wait, what?!



Language Models

Predictive models

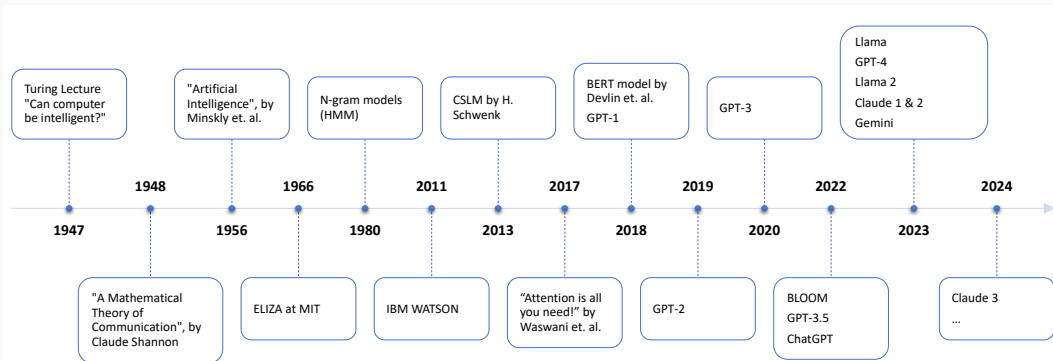
- Uses left context to guess the next word
input: [CLS] the man went to
input: [CLS] the man went to the
input: [CLS] the man went to the store
input: [CLS] the man went to the store [SEP]

Masked models

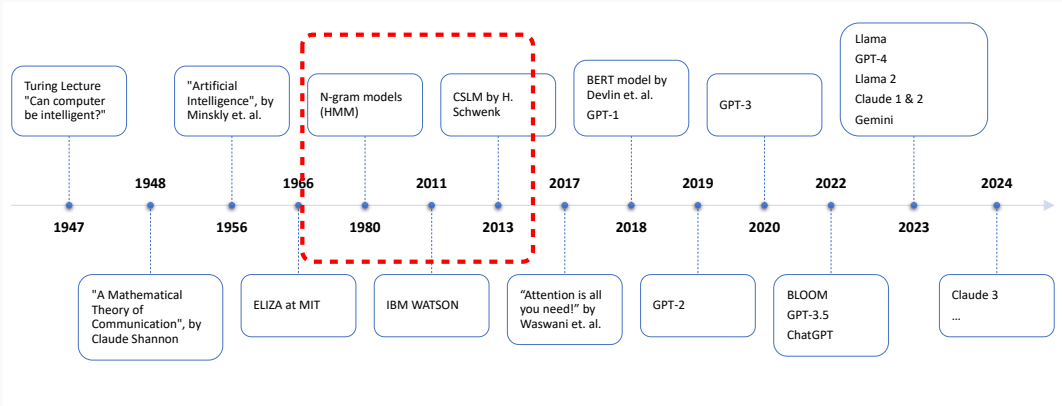
- Uses left and right contexts to guess the word
input: [CLS] the man [MASK] to the store [SEP]
input: [CLS] the man went [MASK] the store [SEP]
input: [CLS] the man went to [MASK] store [SEP]
input: [CLS] the man went to the [MASK] [SEP]

- Statistical Language Models since 80's (HMM)
- include in many tasks:
 - Machine Translation [Koehn et al., 2003]
 - Speech Recognition [Povey et al., 2011]
 - Spoken / Natural Language Understanding [Servan et al., 2006]
 - Any log-linear model

Back to the time-line



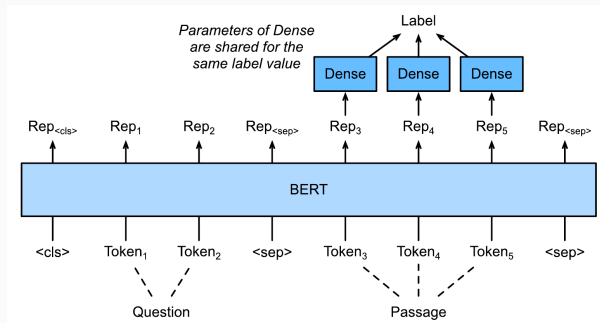
Back to the time-line



Fine-tuning of Neural Language Models

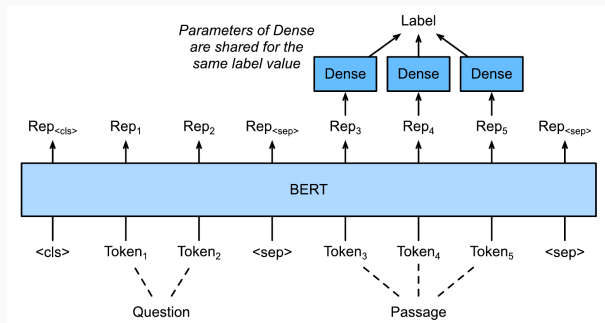
Language Models: transfert learning

- Transformer models [Vaswani et al., 2017]
- Transfert learning with BERT [Devlin et al., 2019]



Language Models: transfert learning

- Transformer models [Vaswani et al., 2017]
- Transfert learning with BERT [Devlin et al., 2019]



👉 Game changer, introducing the nowadays well-known fine-tuning phase

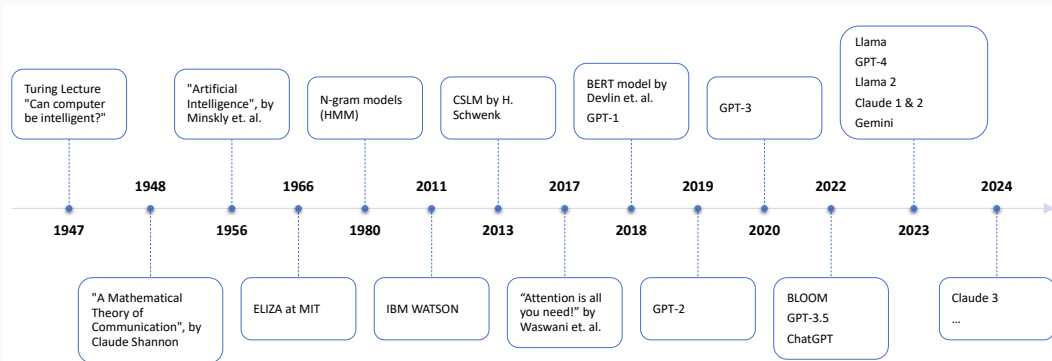
Many tasks:

- token classification: NER, POS-tagging, slot-filling...
- sentence classification: intention detection, sentiment analysis...
- span detection: Question-Answering...
- text generation: summarization, response generation...
- *etc*

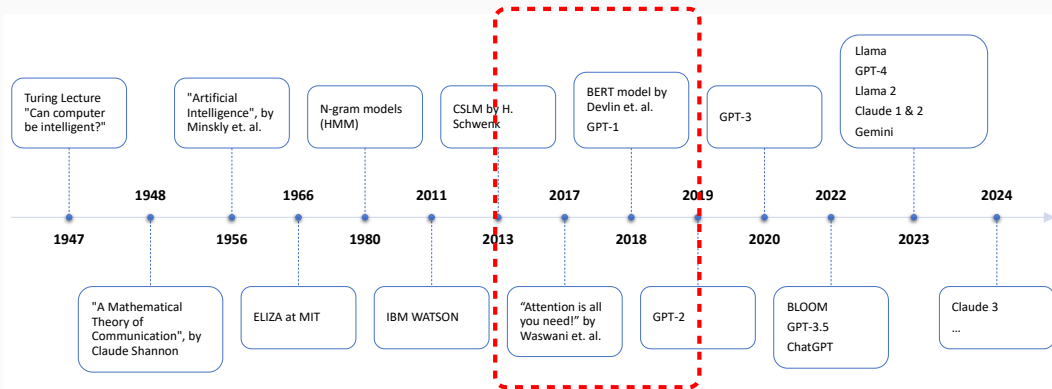
Many tasks:

- token classification: NER, POS-tagging, slot-filling...
- sentence classification: intention detection, sentiment analysis...
- span detection: Question-Answering...
- **text generation**: summarization, response generation...
- *etc*

Back to the time-line



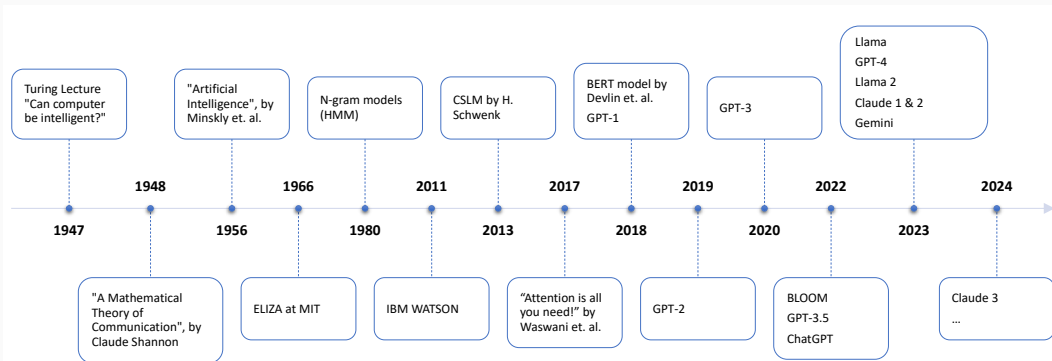
Back to the time-line



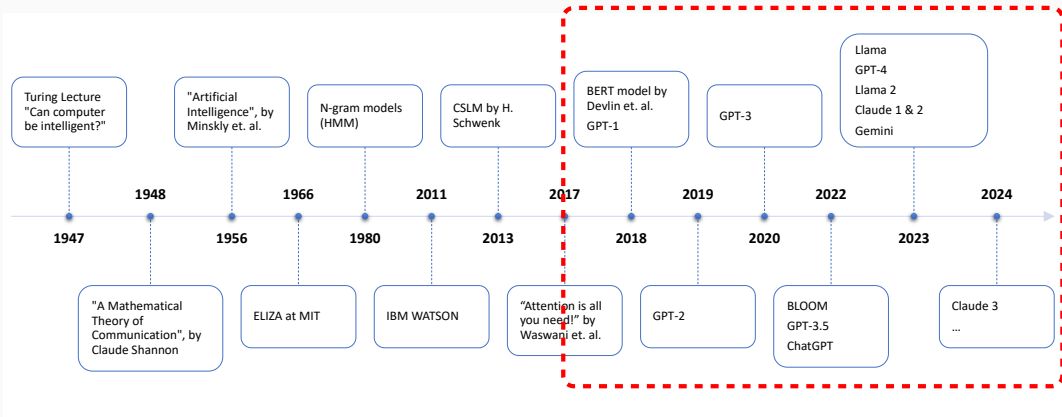
Many tasks:

- introduced by OpenAI in 2018
- enable to generate sequence of token according a query (called *prompt*)
- transfert learning to specify the generation (instructGPT)
- *etc*

Back to the time-line



Back to the time-line



Rebranding Pre-trained Transformer models as Large Language Models

- what does mean “Large”?

Rebranding Pre-trained Transformer models as Large Language Models

- what does mean “Large”?
 - ⇒ according [Shoeybi et al., 2019] from GPT-2 (225 Millions parameters)

Rebranding Pre-trained Transformer models as Large Language Models

- what does mean “Large” ?
 - ⇒ according [Shoeybi et al., 2019] from GPT-2 (225 Millions parameters)
 - ⇒ BERT-large [Devlin et al., 2019] (336 Millions parameters)

Rebranding Pre-trained Transformer models as Large Language Models

- what does mean “Large” ?
 - ⇒ according [Shoeybi et al., 2019] from GPT-2 (225 Millions parameters)
 - ⇒ BERT-large [Devlin et al., 2019] (336 Millions parameters)
- are “classical” BERT models with 110M parameters are small?

Rebranding Pre-trained Transformer models as Large Language Models

- what does mean “Large” ?
 - ⇒ according [Shoeybi et al., 2019] from GPT-2 (225 Millions parameters)
 - ⇒ BERT-large [Devlin et al., 2019] (336 Millions parameters)
- are “classical” BERT models with 110M parameters are small?
 - ⇒ actually... No.

Rebranding Pre-trained Transformer models as Large Language Models

- what does mean “Large” ?
 - ⇒ according [Shoeybi et al., 2019] from GPT-2 (225 Millions parameters)
 - ⇒ BERT-large [Devlin et al., 2019] (336 Millions parameters)
- are “classical” BERT models with 110M parameters are small?
 - ⇒ actually... No.
- maybe we should consider the definition according to the amount of computationnal resources...

Declination of Neural Language Models

Then came the instruct-LLM

Fine-tuning of LLM to generate sequence of tokens from sequence of tokens:

- instruct-GPT
- ChatGPT
- BloomZ
- *etc.*

Then came the instruct-LLM

Fine-tuning of LLM to generate sequence of tokens from sequence of tokens:

- instruct-GPT
- ChatGPT
- BloomZ
- *etc.*

☞ User's game changer \Rightarrow non-expert people can use and exploit LLM

Then came the instruct-LLM

Fine-tuning of LLM to generate sequence of tokens from sequence of tokens:

- instruct-GPT
 - ChatGPT
 - BloomZ
 - *etc.*
-
- ☞ User's game changer \Rightarrow non-expert people can use and exploit LLM
 - ☞ User's behavior drift...

Then came the instruct-LLM

Definition of new practices and methods:

- prompt-engineering
- prompt-tuning
- prompt-generation (by LLM)
- *etc.*

Definition of new practices and methods:

- prompt-engineering
- prompt-tuning
- prompt-generation (by LLM)
- *etc.*

☞ begining of the idea of human-computer interaction optimization

Agent systems

What is an “Agent?”

- persistence (code is not executed on demand but runs continuously and decides for itself when it should perform some activity)
- autonomy (agents have capabilities of task selection, prioritization, goal-directed behavior, decision-making without human intervention)
- social ability (agents are able to engage other components through some sort of communication and coordination, they may collaborate on a task)
- reactivity (agents perceive the context in which they operate and react to it appropriately).

What is an “Agent?”

- all agent are program but all programs are not agent [Franklin and Graesser, 1996]
- agents are not objects [Wooldridge, 2009]
- agents are not expert systems [Wooldridge, 2009]
- agents may be machines, human beings, communities of human beings or anything that is capable of goal-directed behavior [Russell and Norvig, 2003]



What is an “Agent?”

- all agent are program but all programs are not agent [Franklin and Graesser, 1996]
- agents are not objects [Wooldridge, 2009]
- agents are not expert systems [Wooldridge, 2009]
- agents may be machines, human beings, communities of human beings or anything that is capable of goal-directed behavior [Russell and Norvig, 2003]



- ☞ within an environment, with other agents: a multi-agent system

- an eco-system of several agents which interact with one-another
- agents collaborate to complete a task
- act on behalf of users with different goals and motivations
- they must cooperate, coordinate and negotiate to successfully interact



- an eco-system of several agents which interact with one-another
- agents collaborate to complete a task
- act on behalf of users with different goals and motivations
- they must cooperate, coordinate and negotiate to successfully interact

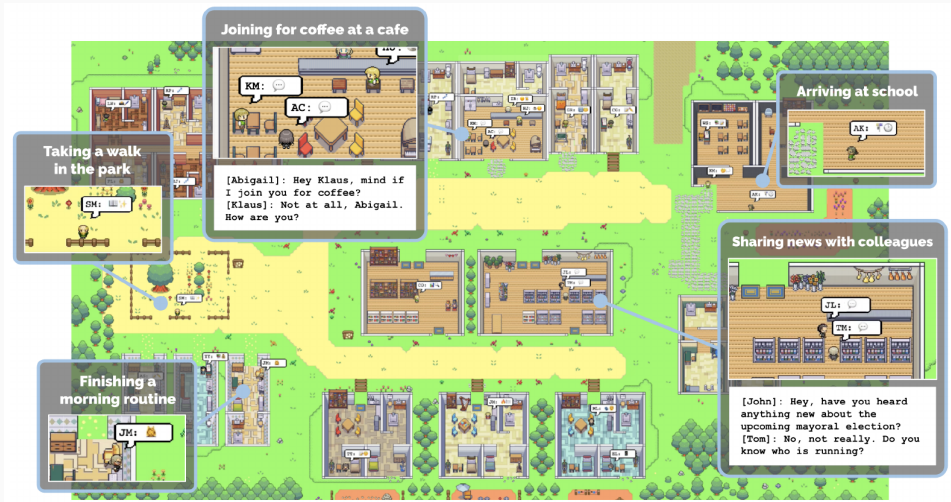


☞ specialized LLMs which communicate with one-another to achieve a goal...

Interaction in close environments: video games

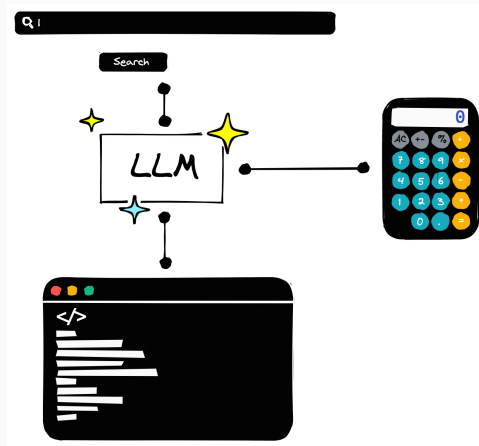
- an eco-system of several agents which interact with one-another
- agents collaborate to complete a task
- act on behalf of users with different goals and motivations
- they must cooperate, coordinate and negotiate to successfully interact



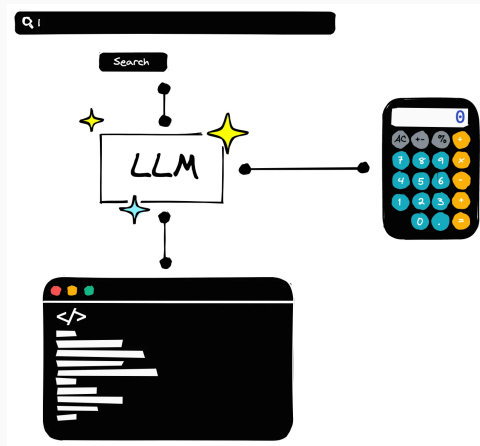


Link LLM & Agents

- Aim to create a high level of interaction
- Combination of several specilized LLM
- Chain of LLM
- Available tools:
LangChain, BabyAGI, and AutoGPT...



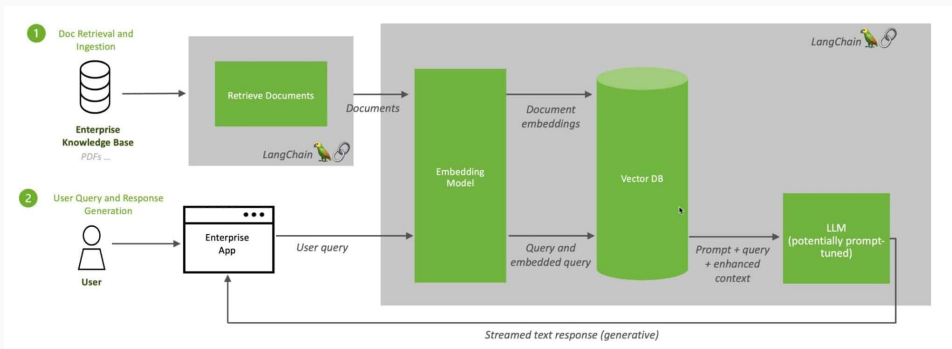
- Aim to create a high level of interaction
- Combination of several specilized LLM
- Chain of LLM
- Available tools:
LangChain, BabyAGI, and AutoGPT...



👉 why not computer-computer interaction?

Retrieval Augmented Generation

- Question-Answering
- Augmented search engine



What's next?

- *In 2045, let's imagine a future where coding is obsolete and replaced by a combination of agents...*
[Ferber, 1999]
- The future is now:
 - computer scientist behaviors has changed
 - no need to be an expert to use LLMs
 - one can assemble LLM as child assemble lego bricks



Conclusion

- Even if it's easier, research questions still open

- Even if it's easier, research questions still open
- LLMs are not allways better than “classical” models

- Even if it's easier, research questions still open
- LLMs are not allways better than “classical” models
- LLMs require **huge** amount of computationnal resources

- Even if it's easier, research questions still open
- LLMs are not allways better than “classical” models
- LLMs require **huge** amount of computationnal resources
- LLMs require **huge** amount of data



- Even if it's easier, research questions still open
 - LLMs are not allways better than “classical” models
 - LLMs require **huge** amount of computationnal resources
 - LLMs require **huge** amount of data
- ☞ Design of specialized models



- Even if it's easier, research questions still open
 - LLMs are not allways better than “classical” models
 - LLMs require **huge** amount of computationnal resources
 - LLMs require **huge** amount of data
-
- ☞ Design of specialized models
 - ☞ Models adapted to environment


- Even if it's easier, research questions still open
 - LLMs are not allways better than “classical” models
 - LLMs require **huge** amount of computationnal resources
 - LLMs require **huge** amount of data
-
- ☞ Design of specialized models
 - ☞ Models adapted to environment
 - ☞ Mix of Small (Language) Models and LLMs in Multi-Agent System

Many thanks for your attention!

`christophe.servan@lisn.upsaclay.fr`


-  Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of deep bidirectional transformers for language understanding.
In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
-  Ferber, J. (1999).
Multi-agent systems: an introduction to distributed artificial intelligence.
Addison Wesley Longman.

-  Franklin, S. and Graesser, A. (1996).
Is it an agent, or just a program?: A taxonomy for autonomous agents.
In International workshop on agent theories, architectures, and languages, pages 21–35. Springer.
-  Koehn, P., Och, F. J., and Marcu, D. (2003).
Statistical phrased-based machine translation.
In Joint Conference on Human Language Technology and of the North American Chapter of the Association for Computational Linguistics, pages 127–133.

 Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023).



Generative agents: Interactive simulacra of human behavior.

In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pages 1–22.

 Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011).

The kaldi speech recognition toolkit.


In IEEE 2011 workshop on automatic speech recognition and understanding, number CONF. IEEE Signal Processing Society.

-  Russell, S. J. and Norvig, P. (2003).
Probabilistic reasoning.
Artificial intelligence: a modern approach.
-  Servan, C., Raymond, C., Béchet, F., and Nocéra, P. (2006).
Conceptual decoding from word lattices: application to the spoken corpus media.
In INTERSPEECH - ICSLP, page 4, Pittsburgh, Pennsylvanie, États Unis d'Amérique.

 Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019).

Megatron-lm: Training multi-billion parameter language models using model parallelism.

arXiv preprint arXiv:1909.08053.

 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).

Attention is all you need.

Advances in neural information processing systems, 30.

 Wooldridge, M. (2009).

An introduction to multiagent systems.

John wiley & sons.