CLADES 2025
House of Connection, Groningen

CAN CHILD DIRECTED SPEECH HELP US BUILD MORE HUMAN-LIKE LANGUAGE MODELS?

PHD STUDENT
Francesca Padovani

Under the supervision of
Arianna Bisazza & Yevgen Matusevych

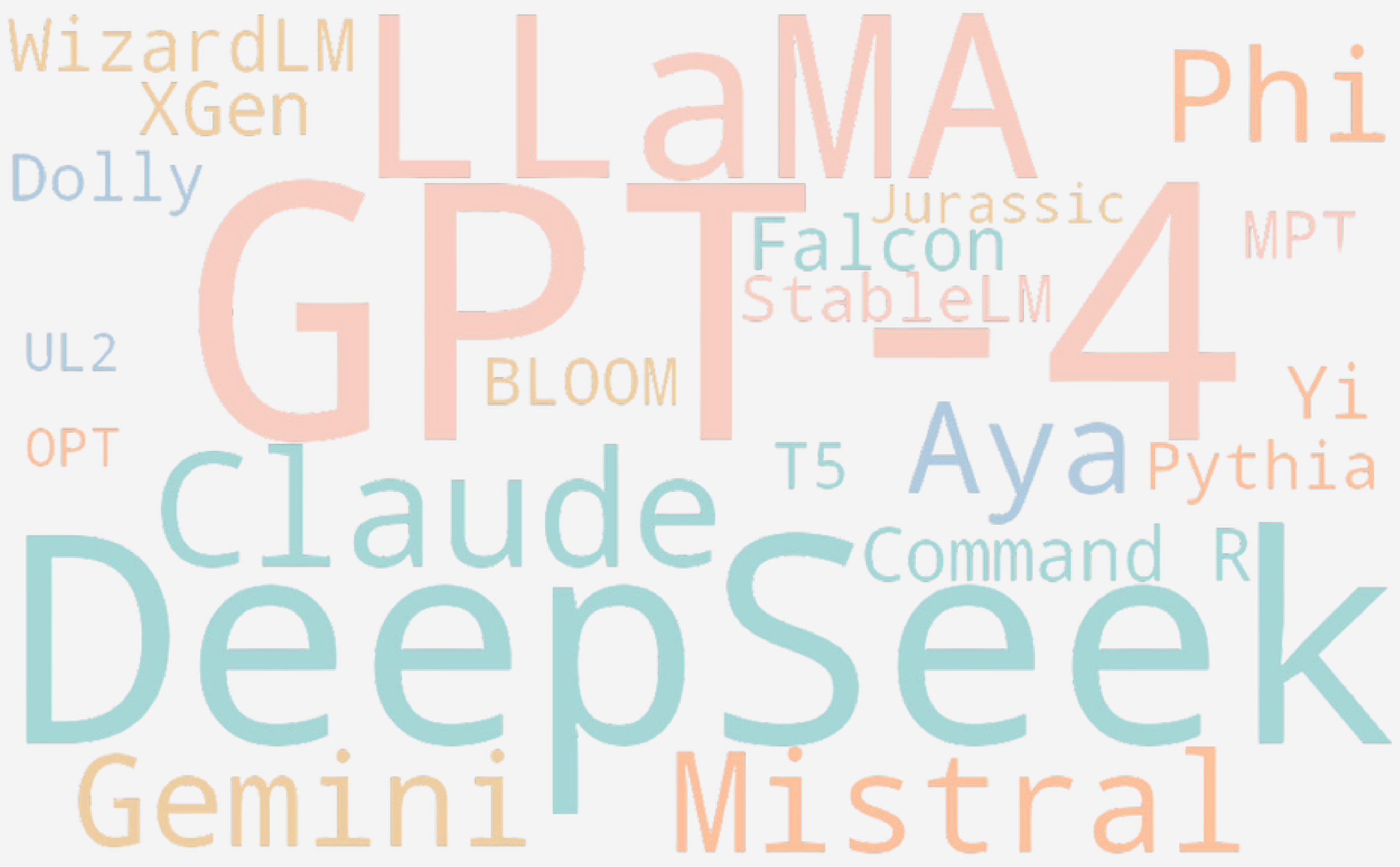
DATE
25/02/2025

rijksuniversiteit groningen

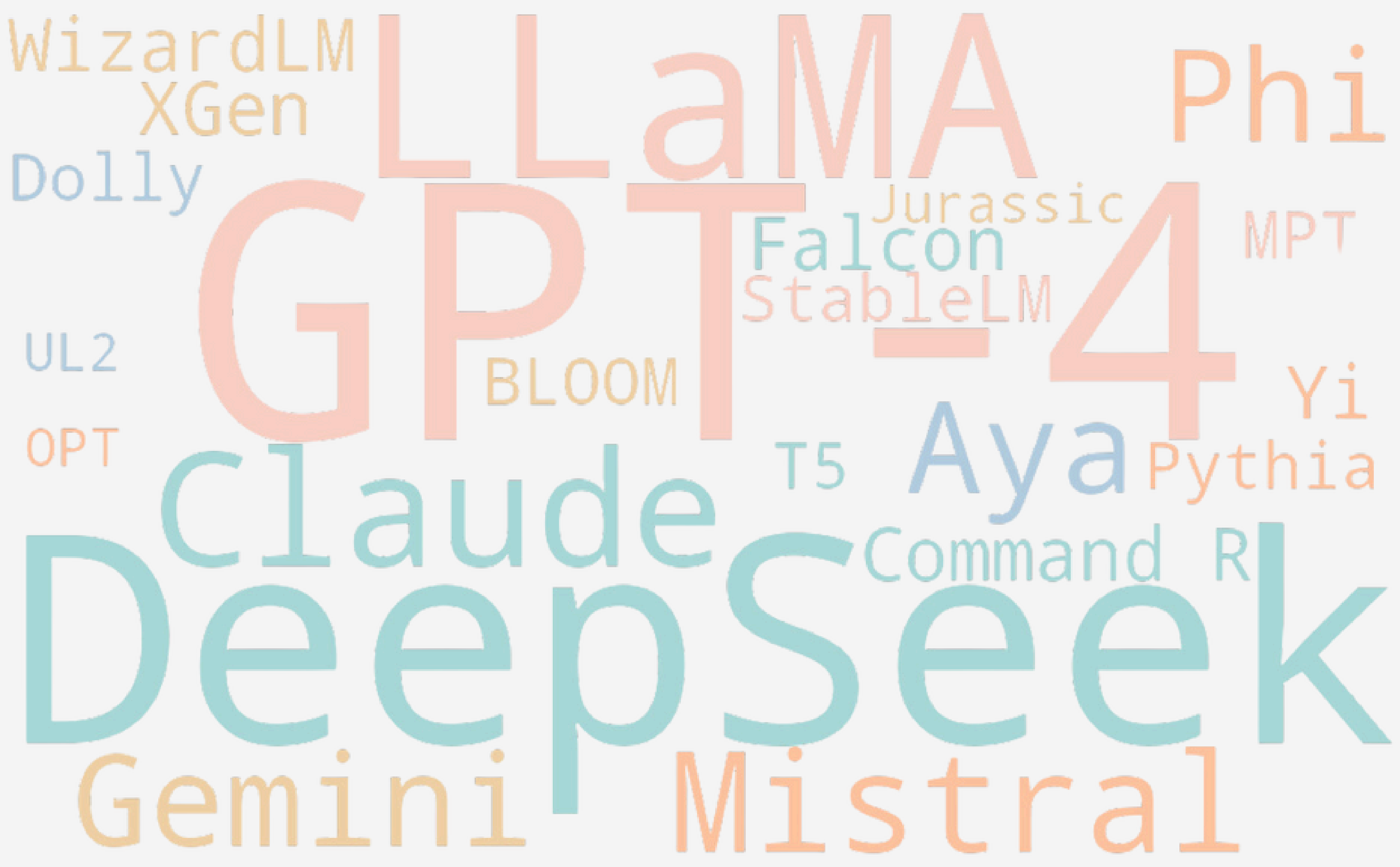
NAME OF PROJECT
Polyglot Machines

TABLE OF CONTENTS

01

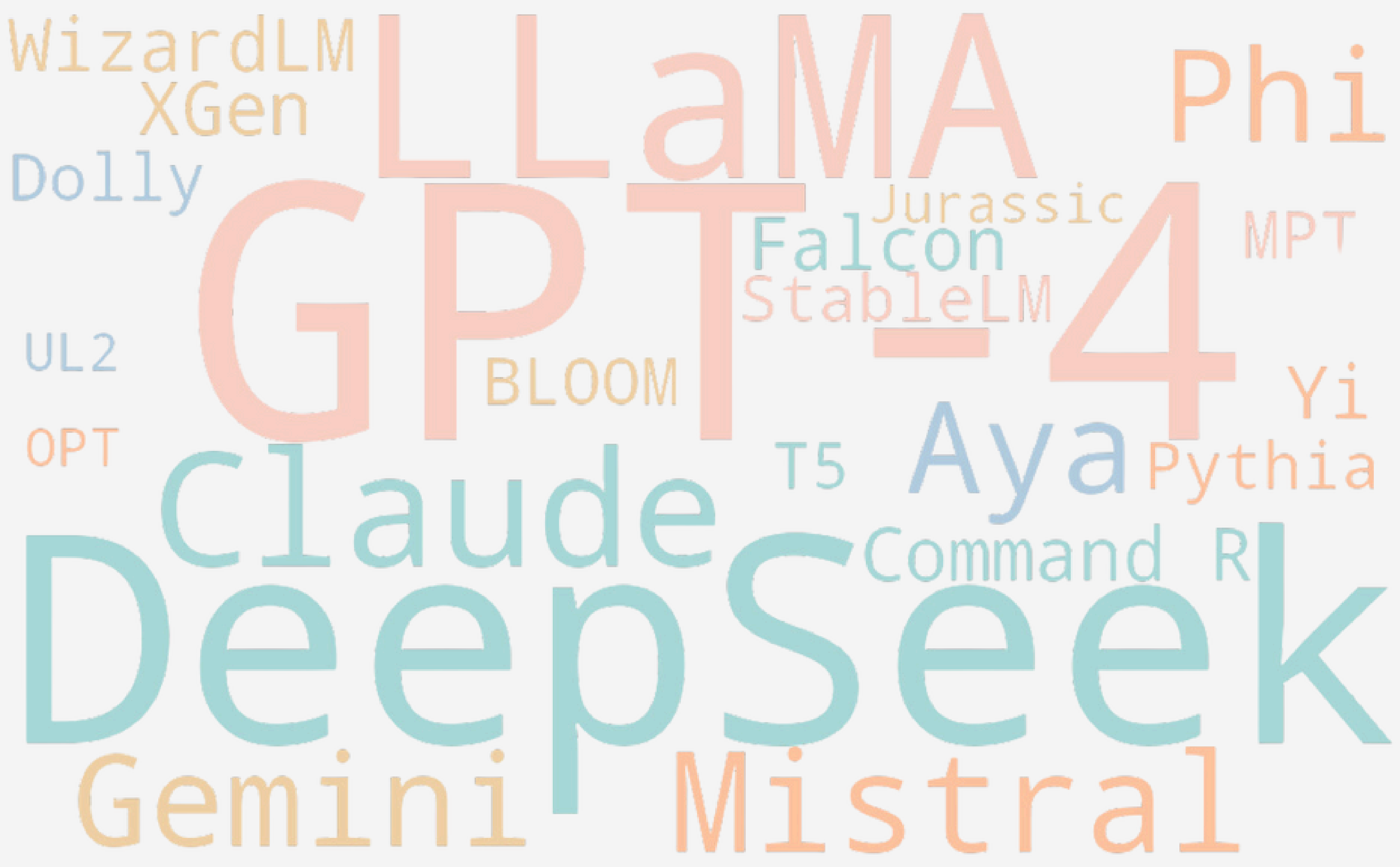

WizardLM
LLaMA
Phi
XGen
Dolly
GPT-4
Jurassic
MPT
Falcon
StableLM
UL2
BLOOM
Yi
OPT
Aya
T5
Pythia
Claude
Command R
DeepSeek
Gemini
Mistral

01



scaling as the only viable way to go

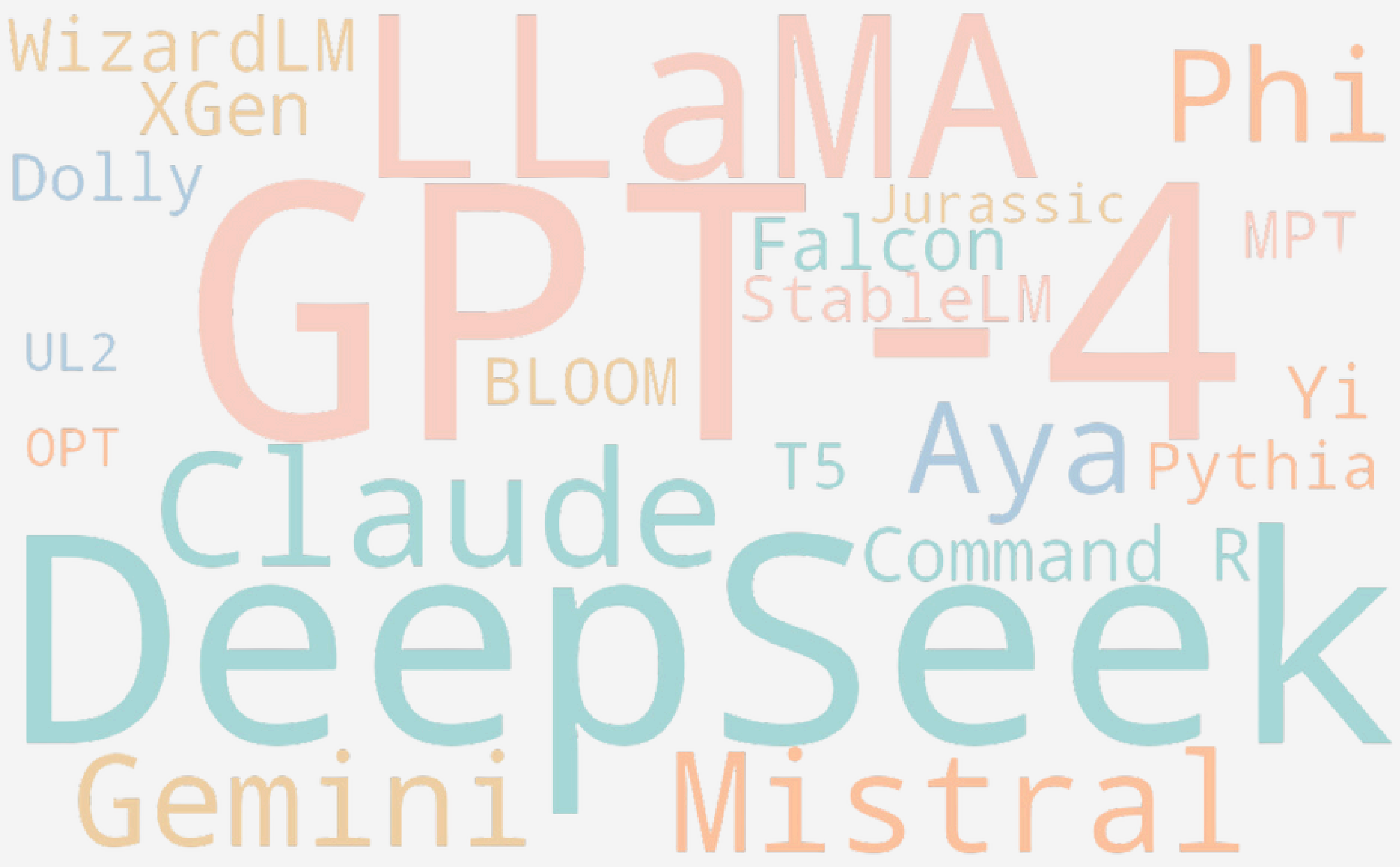
01



scaling as the only viable way to go

extremely high computational and infrastructure costs

01

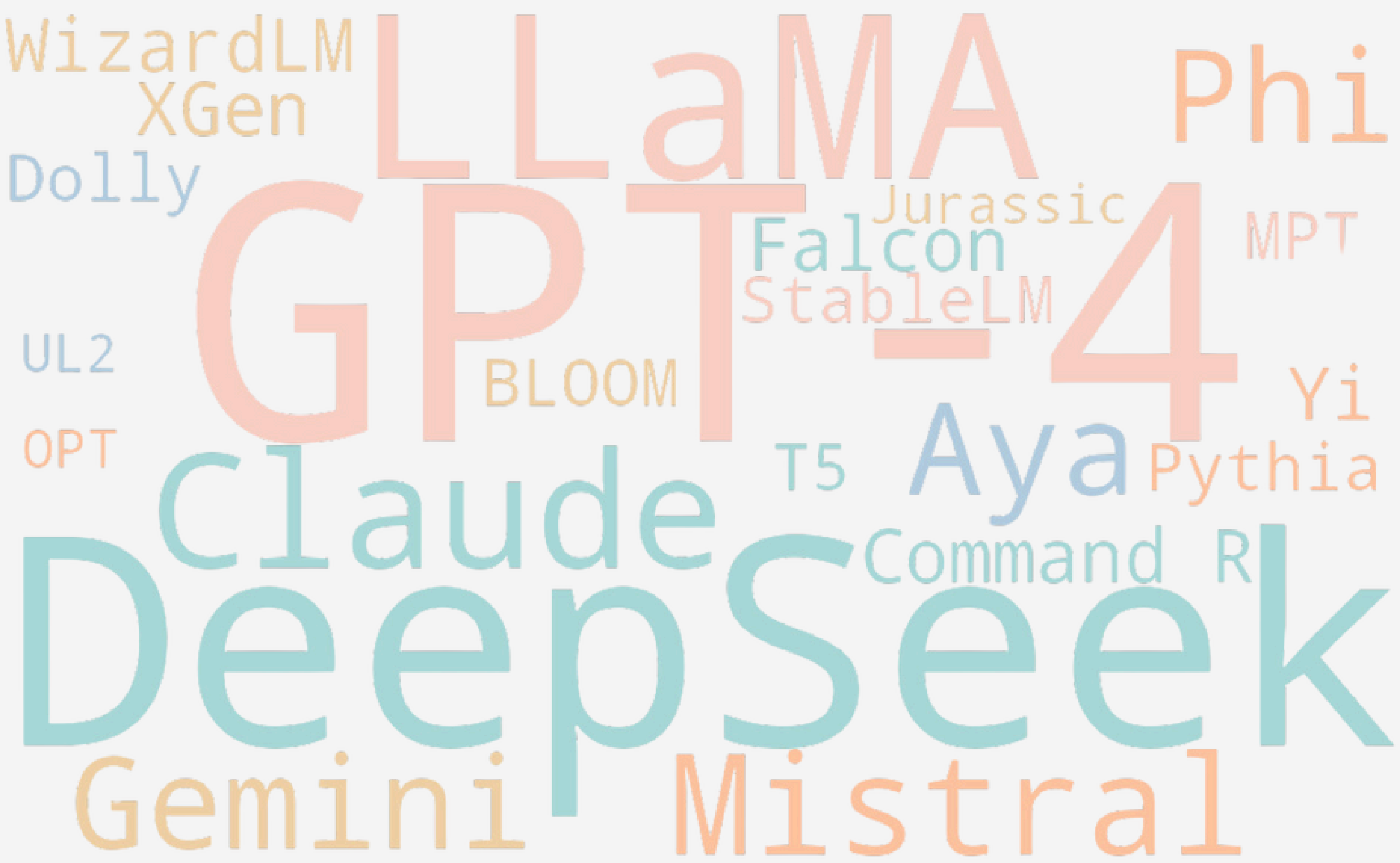


scaling as the only viable way to go

extremely high computational and infrastructure costs

near-total industrial monopoly

01



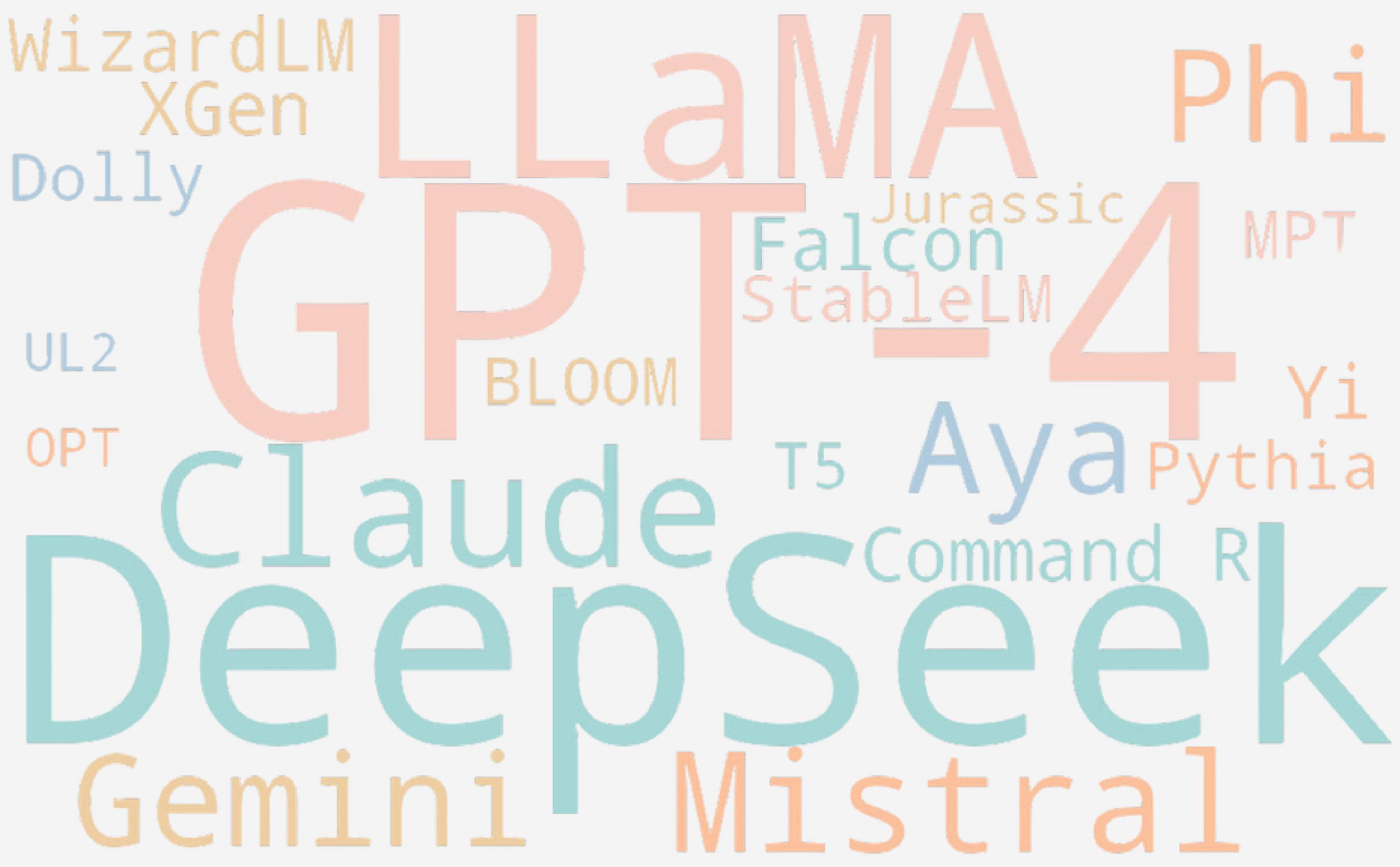
scaling as the only viable way to go

extremely high computational and infrastructure costs

near-total industrial monopoly

lack of transparency and reliability

01



scaling as the only viable way to go

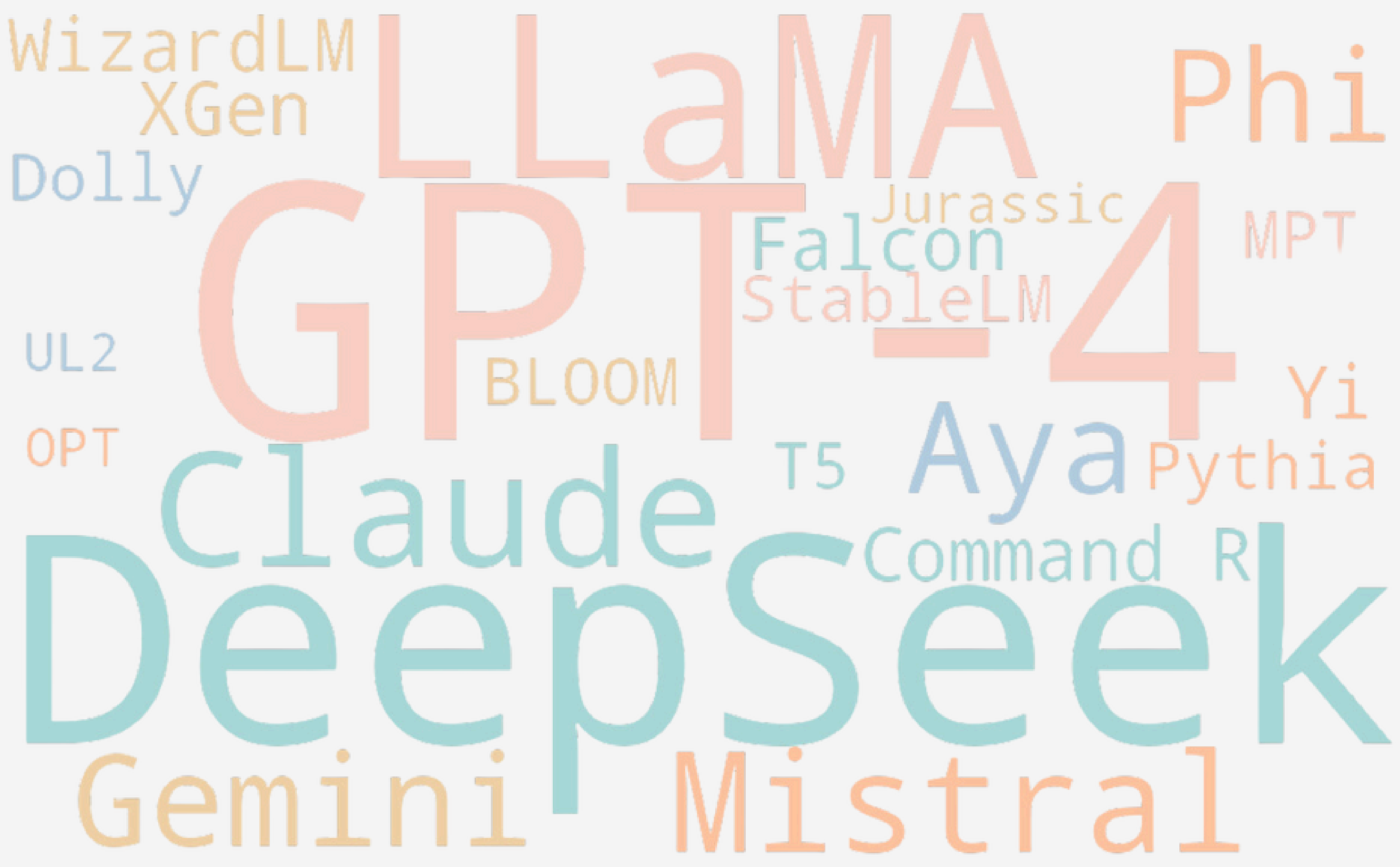
extremely high computational and infrastructure costs

near-total industrial monopoly

lack of transparency and reliability

high carbon footprint

01



scaling as the only viable way to go

extremely high computational and infrastructure costs

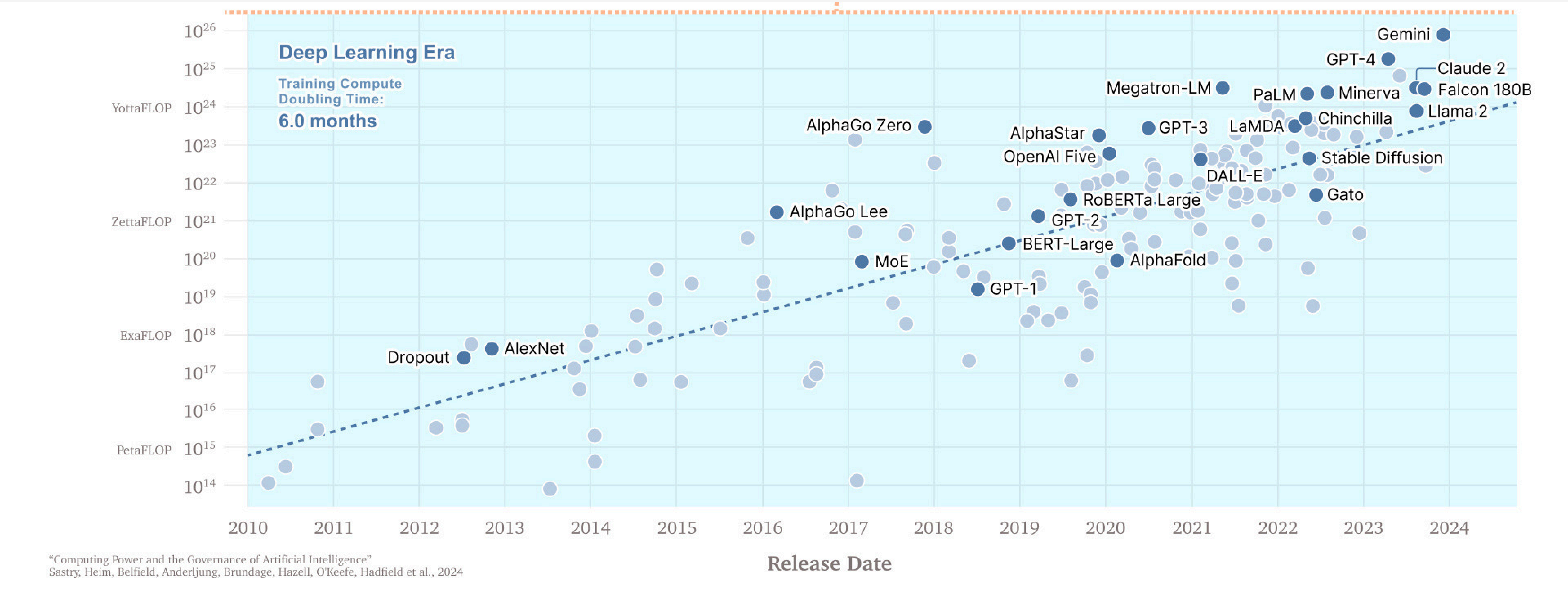
near-total industrial monopoly

lack of transparency and reliability

high carbon footprint

lack of cognitive plausibility, straying from human-like language learning

01



Source: Scaling: The State of Play in AI

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

02

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

ARE EXPOSED TO A WAY SMALLER AMOUNT OF TOKENS

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

ARE EXPOSED TO A WAY SMALLER AMOUNT OF TOKENS

USE SELECTIVE ATTENTION AND COGNITIVE CONSTRAINT

02

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

ARE EXPOSED TO A WAY SMALLER AMOUNT OF TOKENS

USE SELECTIVE ATTENTION AND COGNITIVE CONSTRAINT

RECEIVE FEEDBACK CORRECTIONS AND THEY LEARN IN INTERACTION

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

ARE EXPOSED TO A WAY SMALLER AMOUNT OF TOKENS

USE SELECTIVE ATTENTION AND COGNITIVE CONSTRAINT

RECEIVE FEEDBACK CORRECTIONS AND THEY LEARN IN INTERACTION

LEARN THROUGH MULTIMODAL STIMULI (IMPORTANCE OF VISUAL COMPLEMENTATION)

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

ARE EXPOSED TO A WAY SMALLER AMOUNT OF TOKENS

USE SELECTIVE ATTENTION AND COGNITIVE CONSTRAINT

RECEIVE FEEDBACK CORRECTIONS AND THEY LEARN IN INTERACTION

LEARN THROUGH MULTIMODAL STIMULI (IMPORTANCE OF VISUAL COMPLEMENTATION)

DETECT PATTERN AND SEGMENT SPEECH DIFFERENTLY AT DIFFERENT LEARNING STAGES

02

The way LMs acquire language competence is very different from how children around the world learn any language in their first years of life.

Children:

ARE EXPOSED TO A WAY SMALLER AMOUNT OF TOKENS

USE SELECTIVE ATTENTION AND COGNITIVE CONSTRAINT

RECEIVE FEEDBACK CORRECTIONS AND THEY LEARN IN INTERACTION

LEARN THROUGH MULTIMODAL STIMULI (IMPORTANCE OF VISUAL COMPLEMENTATION)

DETECT PATTERN AND SEGMENT SPEECH DIFFERENTLY AT DIFFERENT LEARNING STAGES

FOR THEM LANGUAGE TYPOLOGY AND COMPLEXITY DO NOT SLOW DOWN OR OSTACOLATE LEARNING

03

SOME LANGUAGES ARE INTRINSICALLY HARDER TO MODEL THAN OTHERS WITH MODERN MACHINE-LEARNING METHODS

03

SOME LANGUAGES ARE INTRINSICALLY HARDER TO MODEL THAN OTHERS WITH MODERN MACHINE-LEARNING METHODS

INFERIOR ABILITY OF SOTA NLP MODELS TO DEAL WITH MORPHOLOGICALLY RICH LANGUAGES (MRL)
such as Ukranian and Turkish, as opposed to Dutch and Italian

03

SOME LANGUAGES ARE INTRINSICALLY HARDER TO MODEL THAN OTHERS WITH MODERN MACHINE-LEARNING METHODS

INFERIOR ABILITY OF SOTA NLP MODELS TO DEAL WITH MORPHOLOGICALLY RICH LANGUAGES (MRL)
such as Ukranian and Turkish, as opposed to Dutch and Italian

REDUCED AVAILABILITY OF READILY AVAILABLE DIGITIZED TEXT FOR MOST OF LANGUAGE

SOME LANGUAGES ARE INTRINSICALLY HARDER TO MODEL THAN OTHERS WITH MODERN MACHINE-LEARNING METHODS

INFERIOR ABILITY OF SOTA NLP MODELS TO DEAL WITH MORPHOLOGICALLY RICH LANGUAGES (MRL)
such as Ukranian and Turkish, as opposed to Dutch and Italian

REDUCED AVAILABILITY OF READILY AVAILABLE DIGITIZED TEXT FOR MOST OF LANGUAGE

MANY UNDERREPRESENTED LANGUAGES, REPRODUCING AND REINFORCING ALREADY EXISTING INEQUALITIES

	Type	|*vocab*|10k	Example sentence: *"I don't want you to play with my lego"*
Dutch	fusional	2690	Ik wil niet dat je met mijn lego speel-t I want NEG COMP you with my lego play.PRES
Italian	fusional	2780	Non voglio che giochi con il mio lego NEG want.1SG.PRES COMP play.2G.PRES with DET.PL.M POSS.SG.M lego
Turkish	agglutinative	3970	Lego-m-la oyna-ma-n-ı iste-mi-yor-um lego-1SG.POSS-INS play-VNOUN-2SG.POSS-ACC want-NEG-PRES-1SG

SOME LANGUAGES ARE INTRINSICALLY HARDER TO MODEL THAN OTHERS WITH MODERN MACHINE-LEARNING METHODS

INFERIOR ABILITY OF SOTA NLP MODELS TO DEAL WITH MORPHOLOGICALLY RICH LANGUAGES (MRL)
such as Ukranian and Turkish, as opposed to Dutch and Italian

REDUCED AVAILABILITY OF READILY AVAILABLE DIGITIZED TEXT FOR MOST OF LANGUAGE

MANY UNDERREPRESENTED LANGUAGES, REPRODUCING AND REINFORCING ALREADY EXISTING INEQUALITIES

	Type	|*vocab*|10k	Example sentence: *"I don't want you to play with my lego"*
Dutch	fusional	2690	Ik wil niet dat je met mijn lego speel-t I want NEG COMP you with my lego play.PRES
Italian	fusional	2780	Non voglio che giochi con il mio lego NEG want.1SG.PRES COMP play.2G.PRES with DET.PL.M POSS.SG.M lego
Turkish	agglutinative	3970	Lego-m-la oyna-ma-n-ı iste-mi-yor-um lego-1SG.POSS-INS play-VNOUN-2SG.POSS-ACC want-NEG-PRES-1SG

Great deal of work has to be done to allow for a fair provision of technology for all the languages spoken around the world and to make LMs inclusive with this respect (Blasi et al., 2022).

CHILD DIRECTED SPEECH / LANGUAGE

“motherese”

CDL holds within itself morphological, syntactic and semantic features that differentiate it from the way adults speak to each other (Rüst et al., 2022; Lester et al., 2022; Onnis & Christiansen, 2008; Stumper et al. 2011; You et al., 2021).

SALIENT PROPERTIES

simplicity (shorter sentences, reduced vocabulary)

redundancy (repetition of a word in successive sentences, or rephrasing of the same intent)

increasingly growing vocabulary

many vocative expressions to address or call a child directly and gain the child's attention

extensive use of diminutives

They may facilitate identifying basic syntactic structures (Onnis et al., 2008; Lester et al., 2022)

BABYBERTA

Huebner et al., 2020

- Trained a Roberta-base model of reduced size
- On a portion of the English CHILDES corpus
- Obtaining astonishing results while comparing the model trained on CHILDES to a model trained on
- Using a minimal pair benchmark called Zorro

	RoBERTa-base	**BabyBERTa**
parameters	125M	5M
data size	160GB	0.02GB
words in data	30B	5M
batch size	8K	16
max sequence	512	128
epochs	>40	10
max step	500	260
hardware[1]	1024x V100	1x GTX1080
training time	24hours	2hours
accuracy[2]	81.0	80.5

04

ZORRO

Minimal Pairs Benchmark

Huebner et al., 2020

example paradigm: anaphor_agreement_pronoun_gender

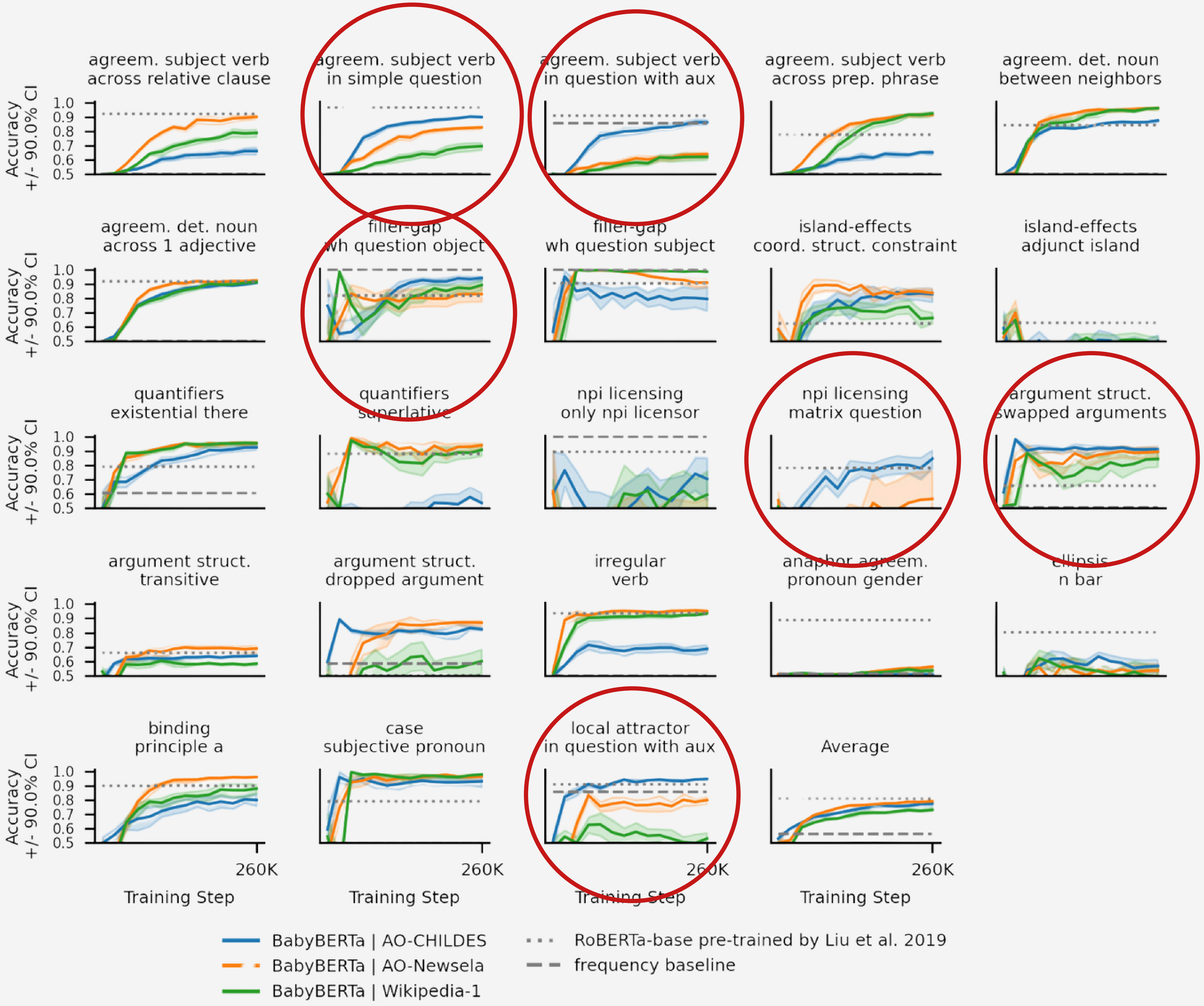
should richard tell himself about the cast ?	GRAMMATICAL
should richard tell herself about the cast ?	UNGRAMMATICAL
should she tell herself about the cast ?	GRAMMATICAL
should she tell himself about the cast ?	UNGRAMMATICAL
richard should not tell himself about the cast .	GRAMMATICAL
richard should not tell herself about the cast .	UNGRAMMATICAL
she should not tell herself about the cast .	GRAMMATICAL
she should not tell himself about the cast .	UNGRAMMATICAL

04

Overall accuracy

Wikipedia	81.0
CHILDES	80.5

But better to have a look at what happens at paradigm level...



05

CURRENT FOCUS

datasets

CHILDES in three languages
English - 5M tokens (Huebner et al., 2020)
French - 2M tokens
German - 4M tokens

VS

WIKIPEDIA (comparable size)
English - 5M tokens
French - 2M tokens
German - 4M tokens

models

GPT-2 base model
&
Roberta base model (like BabyBERTA)

probing benchmark

CLAMS: which evaluates agreements in syntactic constructions of different complexities

05

CLAMS

SIMPLE AGREEMENT

the author laughs
the author laugh
the author swims
the author swim

AGREEMENT IN SUBJECT RELATIVE CLAUSES

the pilots that like the guard laugh
the pilots that like the guard laughs
the pilots that like the guard swim
the pilots that like the guard swims

AGREEMENT IN OBJECT RELATIVE CLAUSES

the farmers that the skater hate are young
the farmers that the skater hates are short
the farmers that the skater hate are short
the farmers that the skater loves laugh

AGREEMENT IN PREPOSITIONAL PHRASES

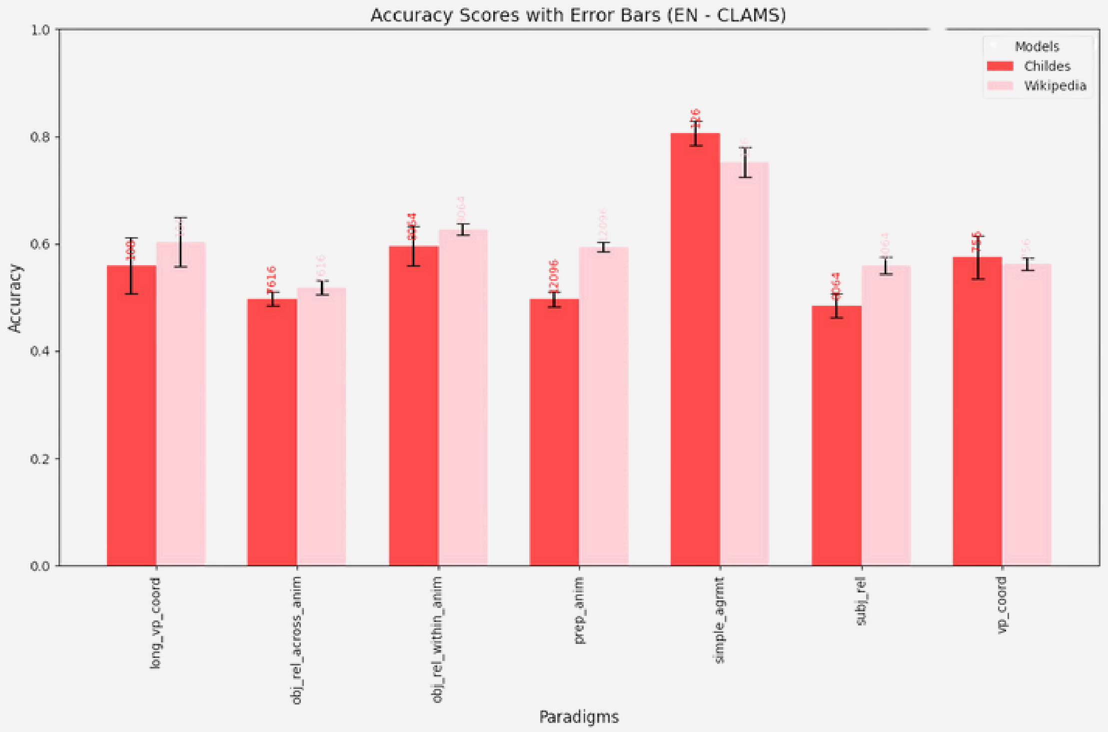
the managers to the side of the dancer swims
the managers to the side of the dancer smile
the customers next to the executive are young
the customers next to the executive is young

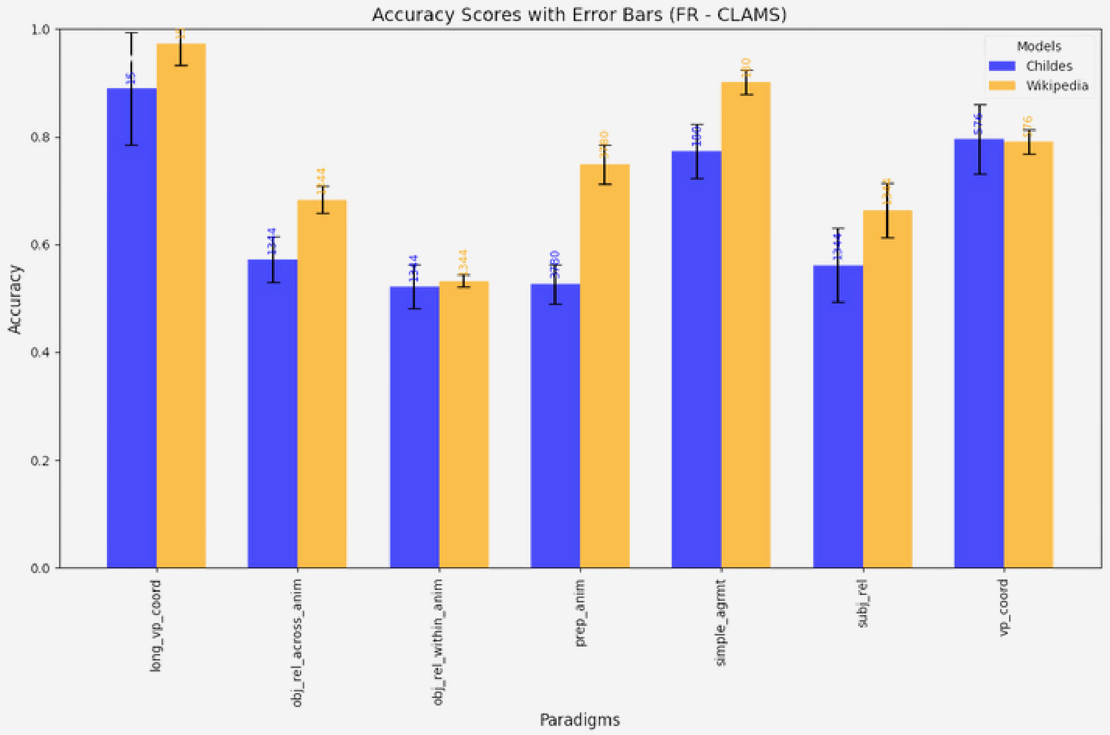
AGREEMENT IN COORDINATES

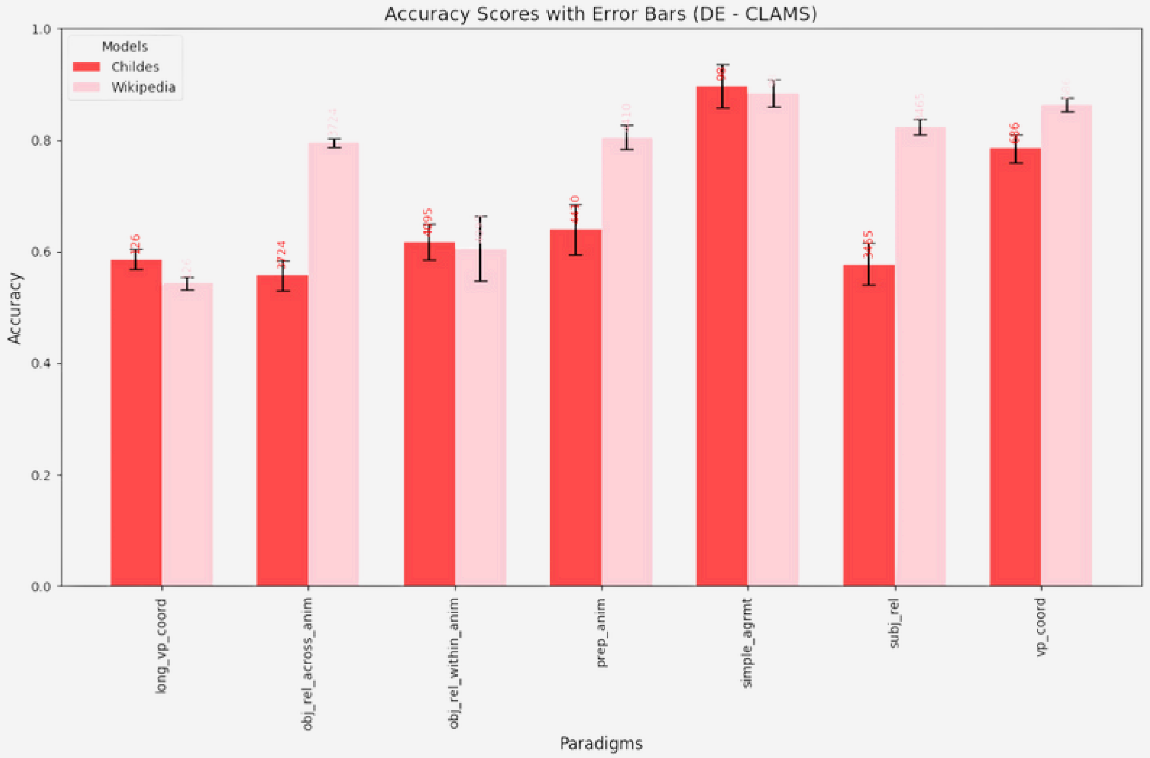
the senator swims and laughs
the senator swims and laugh
the senator swims and smiles
the senator swims and smile

05 ACCURACY SCORES

- The reported results refer to Roberta-base models scores
- Averaged on three seeds







Results from GPT-2 base models are pretty much comparable

CHILDES -trained models outperform WIKIPEDIA-trained ones only on simple agreement for English and German

For French both tests on Agreement in coordinates and Agreement in long coordinates seem to have an advantage when the model is trained on CHILDES

05

REGRESSIONS

Linear Regression and Linear Mixed Effects Models

Modeling the model's log probability scores as a function of training dataset features such as:

- frequency of nsubj (in the two training domains)
- frequency of root (in the two training domains)
- bigram frequency of dependency parsing token-nsubj and token-root (in the two training domains)

minimal_pair	dataset	paradigm	nsubj	verb1	verb2	# subj_log_freq	# verb1_log_freq	# verb2_log_freq
the dreamer assumes, the dreamer assume	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer approves, the dreamer approve	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer insists, the dreamer insist	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer learns, the dreamer learn	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer speaks, the dreamer speak	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer expects, the dreamer expect	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer crashes, the dreamer crash	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer slips, the dreamer slip	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer stands, the dreamer stand	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156
the dreamer knocks, the dreamer knock	wiki	simple_agrmt_new	teachers	write are	write is	4.762173934797756	5.886104031450156	5.886104031450156

05

NEXT STEPS

Linear Regression and Linear Mixed Effects Models

Extensively explore features in regressions

Consider structural variability (making use of dependency parsing tags)

Extend regressions for the three languages

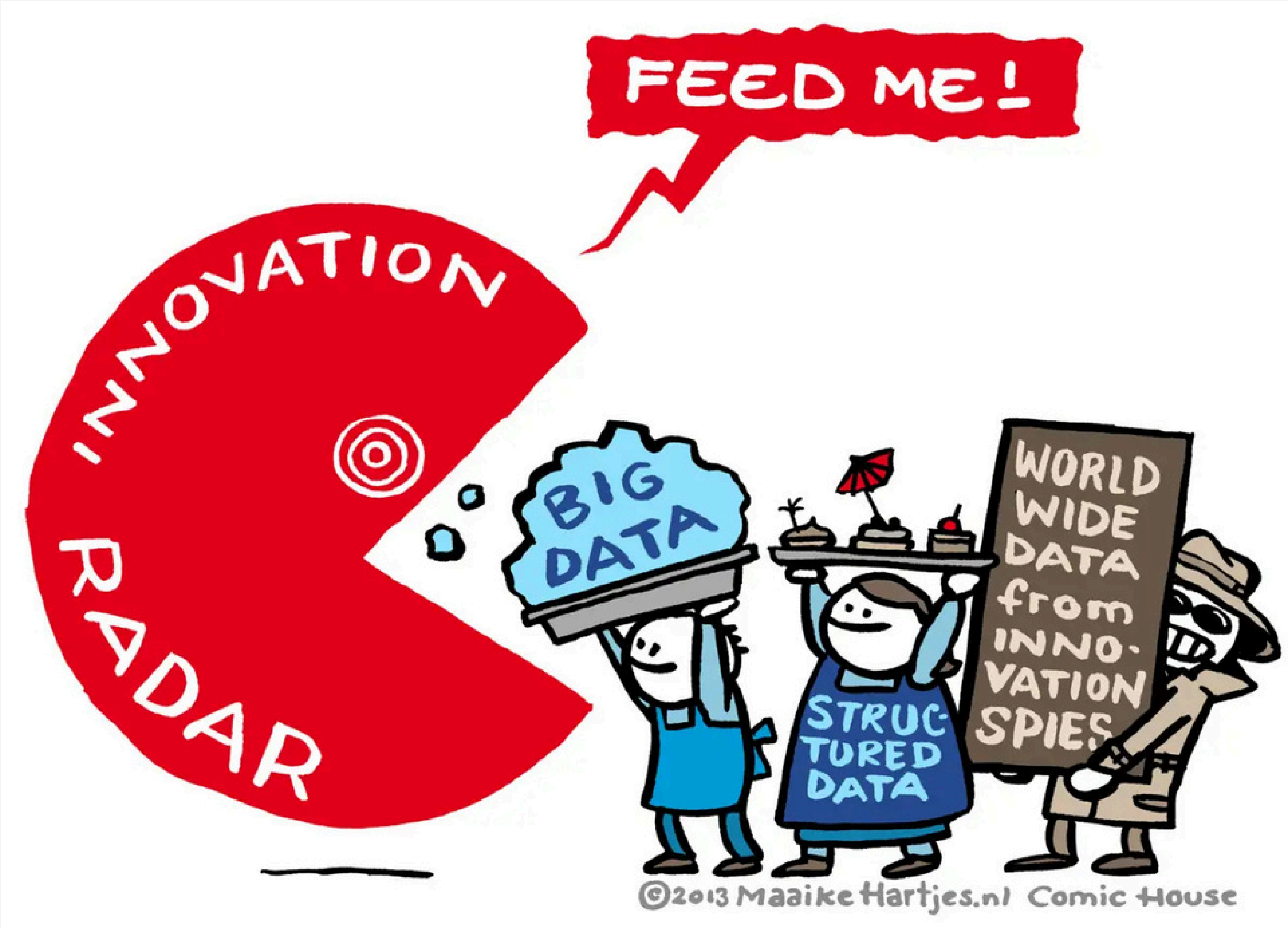
Hopefully get interesting results :)))

06

Thank you ! : -))

CONTACT US

Email Address	fpadovani@rug.nl, a.bisazza@rug.nl
Website	CLCG Group
Bluesky Profile	https://bsky.app/profile/gronlp.bsky.social



NAME OF PROJECT
Polyglot Machines

07 REFERENCES

Blasi, Dami.n, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.

Bunzeck, Bastian, and Sina Zarrieß. "Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly." Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning. 2024.

Huebner, Philip A., et al. "BabyBERTa: Learning more grammar with small-scale child-directed language." Proceedings of the 25th conference on computational natural language learning. 2021.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282-6293.

Lester, Nicholas A., Steven Moran, Aylin C. Küntay, Shanley E.M. Allen, Barbara Pfeiler, Sabine Stoll. 2022. Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages. Cognition, Volume 221.

Mueller, Aaron, et al. "Cross-linguistic syntactic evaluation of word prediction models." *arXiv preprint arXiv:2005.00187* (2020).

Onnis, Luca, Heidi R. Waterfall, and Shimon Edelman. "Learn locally, act globally: Learning language from variation set cues." *Cognition* 109.3 (2008): 423-430

Rüst, Olivier, et al. "The Acquisition of Case Systems in Typologically Diverse Languages: Children Gradually Generalize Grammatical Rules." (2021): 672-685

Stumper, Barbara, et al. "“Frequent frames” in German child-directed speech: A limited cue to grammatical categories." Cognitive science 35.6 (2011): 1190-1205.

You, Guanghao, et al. "Child-directed speech is optimized for syntax-free semantic inference." *Scientific Reports* 11.1 (2021): 16527.

Warstadt, Alex, et al. "BLiMP: The benchmark of linguistic minimal pairs for English." Transactions of the Association for Computational Linguistics 8 (2020): 377-392.