Egocentric Speech in Children and Machines

Ben Gaskin

Man knows himself only inasmuch as he knows the world; he knows the world only within himself and he is aware of himself only within the world. Each new object truly recognised, opens up a new organ within ourselves.

—Goethe

An overview of contents

1. Egocentric speech in children
2. What does this have to do with LLMs?
3. Chain of thought prompting
4. Closing remarks, further considerations

Egocentric speech in children

What is egocentric speech?

- **Jean Piaget**
  - "[Egocentric speech refers to] remarks that are **not addressed to anyone** … and that … **evoke no reaction** adapted to them on the part of anyone to whom they may chance to be addressed"
- **Lev Vygotsky**
  - "Egocentric speech is **inner speech** in its functions; it is speech **on its way inward**, intimately tied up with the **ordering of the child's behaviour**."

How can these theories be distinguished?

- **Hewes and Evans, 1978**
  - The relevant metric is “the coefficient of egocentric speech, i.e., the **ratio of egocentric remarks to total remarks** in a given timespan.”
    - Piaget would predict no relation between task difficulty and this metric
    - “**Increasing task difficulty** did produce a **significant increase** in the coefficient of egocentric speech.”
- **Mitsuhashi et al., 2018**
  - The Luria hand test (LHT) requires participants to **reproduce an ordered sequence of movements** as made during the examiner’s demonstration
    - Conducted under conditions of articulatory suppression (repeat an irrelevant letter) and spatial suppression (visually-guided sequential tapping) during presentation of the sequence
    - “Performance on the LHT was **significantly lower** in the **articulatory suppression condition**, but not in the spatial suppression condition.”

Decision-making behaviour in children

- Vygotsky reports an initial study: “We requested four- and five-year-old children to **press one of five keys on a keyboard** as they identified each one of a series of **picture stimuli assigned to each key**.”
  - The act of choosing is **externally apparent**, evident from bodily behaviour
    - “… the child resolves her choice **not through a direct process of visual perception but through movement**, hesitating between two stimuli, her fingers hovering above and moving from one key to another, going half-way and then coming back.”
- This is followed by a variation: “Subsequent to the experiment described above we attempted to simplify the task of selection by **marking each key with a corresponding sign** to serve as an additional stimulus that could direct and organise the choice process.”
  - The act of choosing is **no longer manifest in external behaviour**
    - “There are **no uncertain groping movements** in the air …”

Language and the structuring of attention

- Vygotsky, *Mind in Society*, p. 35:
  - ”With the help of the indicative function of words, the child begins to **master his attention**, creating **new structural centres in the perceived situation**. As K. Koffka so aptly put it, the child is able to **determine for herself the “centre of gravity”** of her perceptual field; her behaviour is not regulated solely by the salience of individual elements within it. The child **evaluates the relative importance** of these elements, singling out new ‘figures’ from the background and thus **widening the possibilities for controlling her activities**.”
- Language, in other words, is an activity which **structures attention**
  - This goes beyond the perceptual present to create **logical space**
  - We can thus refer to things that are **not present** or **even impossible**
- The language user is thus **both speaker and recipient**
  - These are **combined in egocentric speech**, but why should this have any effect?
  - Piaget’s view is simple: there should be no effect, no difference for the individual

Internalisation of signs during development

- Vygotsky describes a study by Leontiev: “Children were asked to play a game in which they were to **answer a set of questions without using certain words** in their answers.”
  - These were colour words, that there might be **two colours prohibited**
  - Some children were given a **set of colour cards**, including the prohibited colours
- Leontiev investigated subjects from **five to twenty-seven years old**
  - First stage (preschool), **little difference** between subjects with and without cards
  - Second stage (school), **with cards performs much better** than subjects without
  - Third stage (adults), **little difference** between subjects with and without cards
- Vygotsky, p. 45: “What takes place is what we have called **internalization**; the external sign that school children require has been transformed into an **internal sign produced by the adult** as a means of remembering.”

What does this have to do with LLMs?

What does this have to do with LLMs?

- Not a matter of identity or even equivalence, instead **analogy**
  - Hence the ultimate criterion: **whether this is fruitful, valuable**
- **What must language be like that this is possible?**
  - The mode of **acquisition**: text data, unsupervised learning, backpropagation
    - Deep structure of language
  - The computational **architecture**: RNNs, LSTMs, Transformers
    - Long-term dependencies

Correlations between human and machine

- **Convolutional neural networks** (Yamins et al., 2014)
  - Categorisation of natural categories: animals, boats, cars, etc.
  - Model activity predictive of inferior temporal and V4 neural activity
  - Predictivity further **correlated with classification performance**
- **Language models** (Schrimpf et al., 2021)
  - High performing models are predictive of contrastive neural activity
  - The same models are also predictive of behaviour, reading times
  - **Untrained language models** further demonstrated above-chance predictivity
- **Self-attention** (Bensemann et al., 2022)
  - Layer one attention, averaged across attention heads, related to eye gaze
  - **Attention correlated with dwell time** during reading comprehension tasks

Chain of thought prompting

Large language models (LLMs)

- What does the word ‘large’ mean here?
  - Models begin to demonstrate emergent capacities with increasing scale
  - Chain of thought prompting is one of these, not present in smaller models
- What does a language model model?
  - Not a world model
    - See, e.g., the reversal curse (Berglund et al., 2023)
  - Not even a language user, rather **language use**
    - The user is inferred to the extent this aids prediction (Andreas, 2022)
    - Language as an activity, as a behaviour, in its relations to self and other

What sort of language does an LLM model?

- Trained on **written language**, which is not neutral
  - This differs from **spoken language**
  - This differs from **inner speech** (e.g., abbreviation)
    - Inner speech expanded to written equivalent: ~4,000 words per minute (Korba, 1990)
- Language as written **largely by adults**
  - This differs from the language use typical in human development
  - The data here likely includes less babbling, less “self-evident” statements, etc.
- **Training on code** seems to benefit reasoning broadly—why?
  - Programming is a **strictly explicit form of linguistic reasoning**
  - LLMs **generalise with increasing model size** (Grosse et al., 2023)

A brief history of language modelling

- McCulloch and Pitts, 1948
  - A logical calculus of the ideas immanent in nervous activity
- Bengio et al., 2003
  - A neural probabilistic language model
- Elman, 1980
  - Finding structure in time
- Hochreiter and Schmidhuber, 1997
  - Long short-term memory
- Vaswani et al., 2017
  - Attention is all you need
- Liu et al., 2018
  - Generating Wikipedia by summarising long sequences
- Brown et al., 2020
  - Language models are few-shot learners

Static and dynamic interpretability

- **Static interpretability**
  - Language models can be used to **label neuron behaviour** en masse (Bills et al., 2023)
  - **Polysemantic neurons** encode for multiple contradictory features (Elhage et al., 2023)
  - Polysemanticity is tractable, can be **decomposed via dictionary learning** (Bricken et al., 2023)
- **Dynamic interpretability**
  - Bricken et al., 2023:
    - "One of the most striking phenomena we've observed in our study of the features in one-layer models is the existence of "finite state automata"-like assemblies of features. These assemblies aren't circuits in the conventional sense—they're formed by **one feature increasing the probability of tokens, which in turn cause another feature to fire on the next step, and so on**."
  - Berglund et al., 2023:
    - "If a model is trained on a sentence of the form "<name> is <description>" (where a description follows the name) then **the model will not automatically predict the reverse direction** "<description> is <name>." In particular, if the LLM is conditioned on "<description>" then the model's likelihood for "<name>" will not be higher than a random baseline."

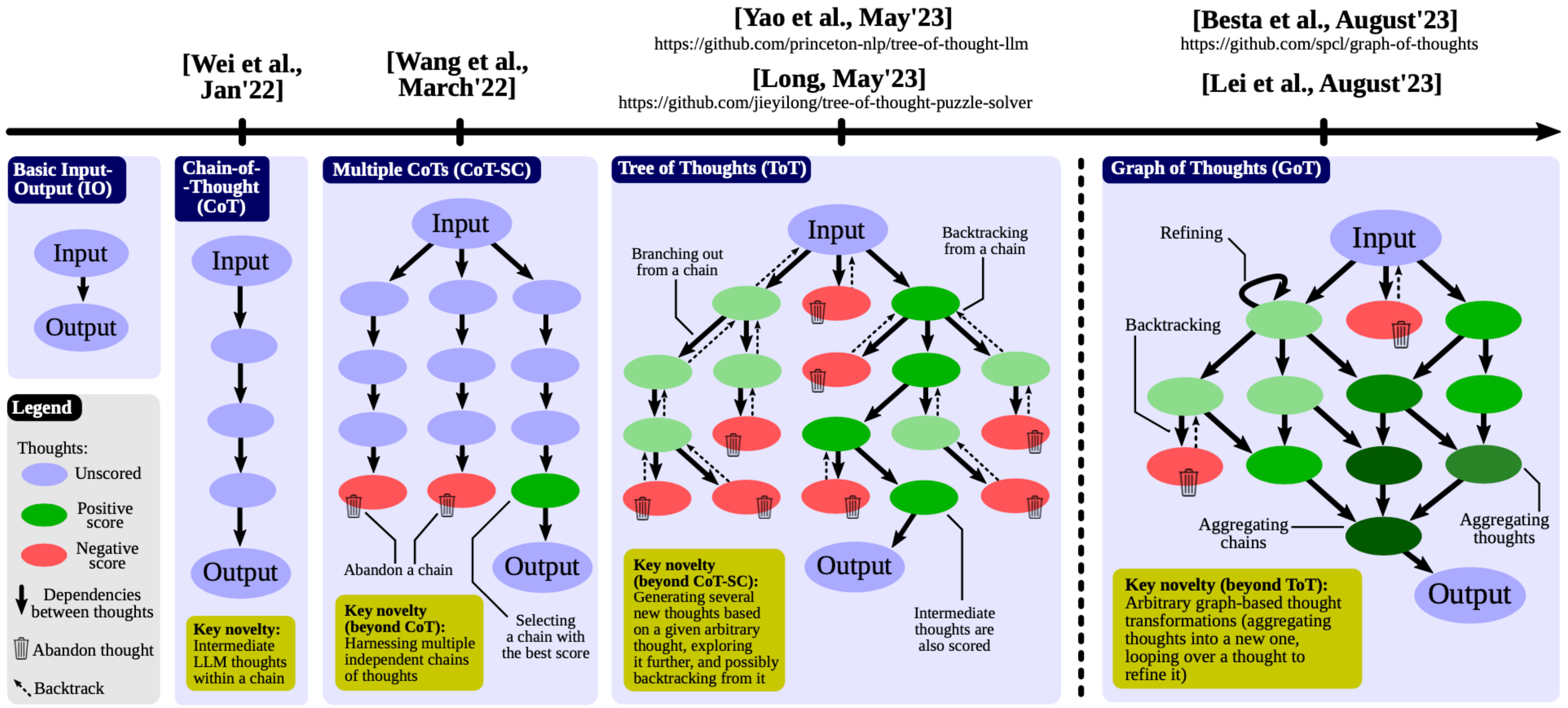
Zeno’s paradox of the arrow

- Per Aristotle, “If everything when it occupies an equal space is at rest, and if that which is in locomotion is always in a now, the flying arrow is therefore motionless.”
- For LLMs: **if at each step we find inference, then where is reasoning?**
  - It emerges from the **pattern of movement**, that this implies constraints
  - Take the reversal curse, here the **angle of approach** determines outcome
- Chain of thought prompting is an emergent phenomenon and a particularly clear case requiring dynamic interpretability
  - What matters here, as with the reversal curse, is the movement of inference

Chain of thought (CoT) prompting

- **Training** to use scratchpads in Nye et al. (2021)
  - "Our proposal is simple: Allow the model to produce an arbitrary sequence of intermediate tokens, which we call a scratchpad, before producing the final answer. For example, on addition problems, the scratchpad contains the intermediate results from a standard long addition algorithm. To train the model, we encode the intermediate steps of the algorithm as text and use standard supervised training."
- **Few-shot prompting** of chain of thought in Wei et al. (2023)
  - "The goal of this paper is to endow language models with the ability to generate a similar chain of thought—a coherent series of intermediate reasoning steps that lead to the final answer for a problem."
- **Zero-shot prompting** of chain of thought in Kojima et al. (2023)
  - "… our Zero-shot-CoT successfully generates a plausible reasoning path in a zero-shot manner and reaches the correct answer in a problem where the standard zero-shot approach fails. Importantly, our Zero-shot-CoT is versatile and task-agnostic …"

Variants of CoT

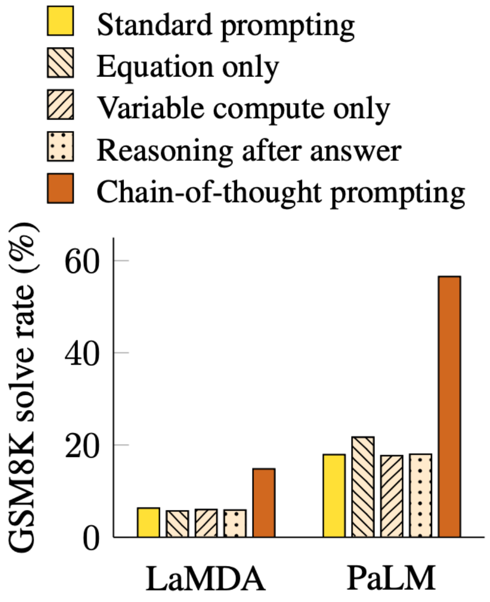


—from Besta et al. (2024), p. 2.

Characteristics of CoT

- **Significant performance increases** on a **variety of benchmark tasks**
  - Mathematical, logical, symbolic, commonsense (Wei et al., 2023)
- **Emergent feature**: arises with **model scale**, not trained for directly
  - Performance gains strongest from ~100B parameters (Wei et al., 2023)
- **Not always faithful**, may be systematically misleading
  - If told to answer A, for instance, will justify with fluent reasoning but not mention the overriding instruction to select this (Turpin et al., 2023)
  - Faithfulness **varies with model size and task difficulty**, with there being an area of task difficulty where reasoning is most faithful (Lanham et al., 2023)

Explanations for CoT



- Wei et al. (2023) perform three ablation studies
  - Equation only
  - Variable compute only
  - Chain of thought after answer
- Invalid reasoning achieves ~90% of performance (Wang et al., 2023)
  - Further consider chains in terms of **bridging objects** and **language templates**
  - Ablation studies: relevance (based on query) and coherence (proper ordering)
  - Relevance and coherence are key
    - Relevance matters more for bridging objects, entities correspond to the initial query
    - Coherence matters more for language templates, sequential ordering of structuring text

Conditional compute as circuit complexity

- Transformers expend the **same compute per forward pass**
  - So perhaps it is that CoT allows the model to **allocate more compute to a task**
- There have been various proposed augmentations—
  - For **more**, as in **PonderNet** (Banino et al., 2021)
  - For **less**, as in **Mixture-of-Depths** (Raposo et al., 2024)
- This perspective has been formalised by Li et al. (2024):
  - "Intuitively, without CoT, the **number of serial computations** conducted by the transformer is **bounded by the depth** (which is considered as a fixed constant for this work), whereas **with T intermediate steps**, the **number of serial computations possible is boosted to T**. Note that T can easily increase as the sequence length increases where the depth is a fixed number that depends on the architecture."
  - They prove this theoretically; then empirically show projected depth requirements for standard transformers, while CoT enables consistent success at minimal depth

Intermediate tokens as recurrent state

- Merrill & Sabarwhal (2024), is attention all you need?
  - "The intuition here is that the transformer **lacks recurrent connections**, and recurrence is **required to solve these sequential reasoning problems**. Empirically, … the **reasoning performance of GPT-4 negatively correlates with the depth of the problem's computation graph** (Dziri et al., 2023)."
  - "These methods [i.e., CoT] allow the transformer to output a sequence of intermediate tokens before answering … Intuitively, such methods could unlock greater expressive power on sequential reasoning problems because **the model can use each intermediate token as a kind of recurrent state**."

The transformer architecture

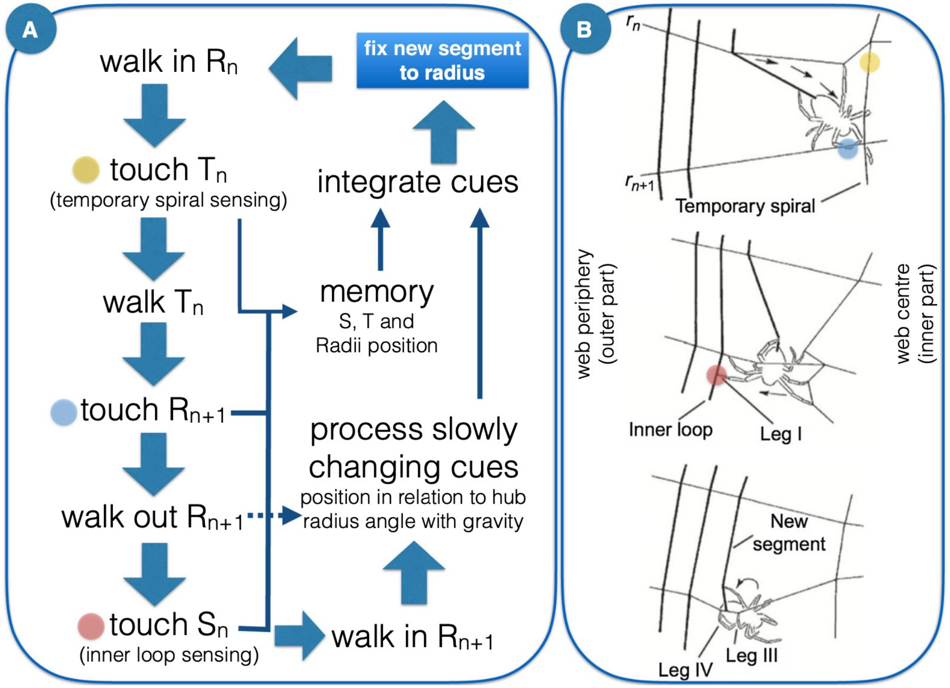
- Development took place in the context of RNNs, LSTM
  - From RNNs to LSTM, vanishing gradient problem
  - From LSTMs to transformers, **hidden state bottleneck**
- Bahdanau et al. (2015) propose **self-attention** to allow encoder–decoder translation models to deal with **longer sequences**
- Vaswani et al. (2017), all you need: **no recurrence**, **no hidden state**
  - Now performance does not decay, instead **cost scales quadratically**
- Liu et al. (2018) introduce the **decoder-only** architecture
  - **Unidirectional transformer**, left–right masked self-attention

What is self-attention?

- The meaning of a token for a model is an **n-dimensional embedding**
  - Word2Vec, for instance, per Mikolov et al. (2013)
- From left to right, the **model evaluates each input token in turn**
  - For token n, the **elements of self-attention** for the input sequence are:
    - **Query** vectors—what do I want?
    - **Key** vectors—what have you got?
    - **Value** vectors—what does it mean?
  - The **match between query and key** determines the **weighting** of prior tokens
  - The **values** of these attended tokens are taken to produce a **weighted sum**
  - This weighted sum **‘bends’** the embedding of the **current token**

Spider webs and the extended mind

- Japyassu and Laland, 2017: “Since web threads are **reliably out there** while the spider is building its trap, there is **no need to memorise** all the details of the emerging structure … because **at each new step** of the building process the spider **can reset the memory used in the previous step**.”
  - “… at each new fixation of one spiral segment, the spider can forget the distance memorised for the fixation of the previous spiral segment. Thus, the spider is able to trade long-term for short-term spatial memory, simply because the threads already fixed will remain in place, cueing the next steps.”

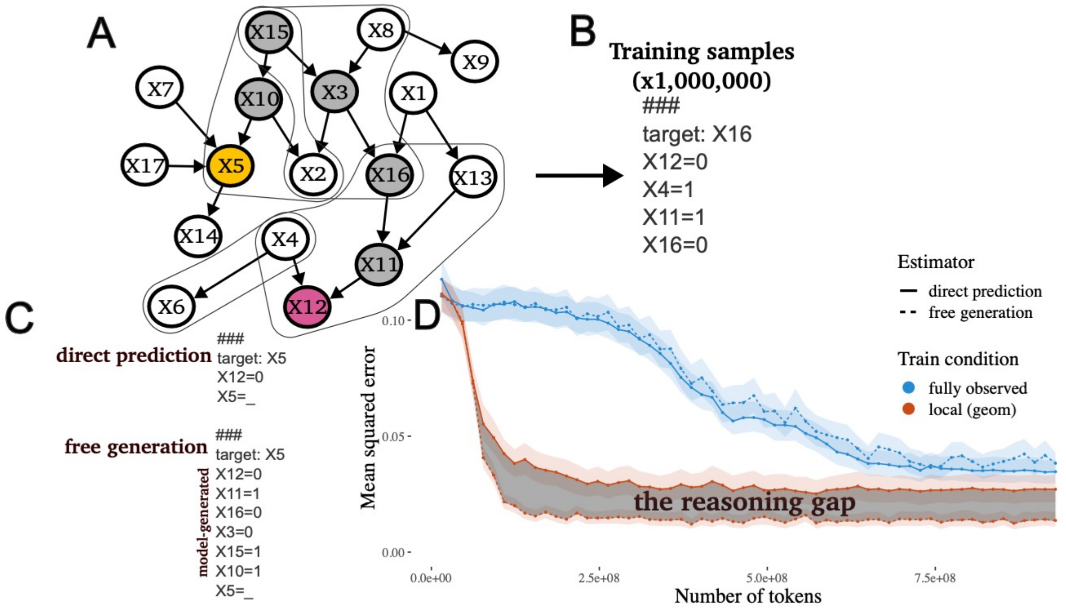


Depth and the role of language

- The **number of steps** is **not enough** on its own
  - Wei et al. (2023), tested with **ellipses** equivalent in length to the chain
  - Lanham et al. (2023) repeated this test, extend and confirm the finding
  - **Thinking "dot by dot"** (Pfau et al., 2024)—**exception** that **proves the rule**
- Comparing 'thought' tokens and language
  - **Pause tokens** (Goyal et al., 2023)
  - Quiet-STaR (Zelikman et al., 2024)
    - "Goyal et al. (2023) show that learning a single 'pause' token (**essentially representing each token as two tokens**) improves LM performance. However, unlike the thought tokens in our work, this **pause token does not initialize a thought**—instead, it can be seen as acting as the entirety of the thought. **We find that reasoning in language is significantly more helpful**."
- **Language itself** seems to play an **essential role** here
  - Explains CoT as emergent, leveraging more or less **task-agnostic patterns in the data**

Data structures and the locality of experience

- Prystawski et al. (2023) train a simple model to investigate reasoning
  - Models trained to predict **conditional probabilities** of generated Bayes nets
  - Chained reasoning over this topology, varying the **structure of training data**
    - Direct prediction
    - Scaffolded generation
    - Free generation
    - Negative scaffolded generation
  - "… when we need to infer the effect of one piece of information on another but **have not encountered them together**, we must make a **series of inferences** that **jump between pairs of concepts** to **connect what we know with what we want to infer**."



The autoregressive aspect of LLMs

- These models are trained on the task of **next-token** prediction
  - This token is then **appended to the input sequence**, iterative processing
  - The generation of tokens thus **iteratively alters the attention landscape**
- This allows the model to **steer its own attention** by generating tokens
  - Simplest case as in Prystawski et al. (2023), scaffolded generation of chain traversals
  - Serial reasoning and task decomposition, self-assembly on the scratchpad
- Leverages **reasoning patterns implicit in the deep structure of language**
  - Wang et al., 2023: "the LLM has already **gained a lot of such complex reasoning ability from pretraining** … and the provided reasoning steps serve more as the role of an **output format/space**, that regularizes the LLM to generate rationales that look step-by-step while being coherent and relevant to the query."

At last, back to Vygotsky

- “Beginning with Köhler, scholars have noted that **the ability or inability to direct one's attention is an essential determinant of the success or failure of any practical operation**. … children are capable of reconstructing their perception and thus freeing themselves from the given structure of the field. **With the help of the indicative function of words, the child begins to master his attention**, creating new structural centers in the perceived situation.”
- “**New motives, socially rooted** and intense, provide the child with direction. K. Lewin described these motives as Quasi-Beduerfnisse (quasi-needs) … Because he is able to form quasi-needs, the child is capable of **breaking the operation into its separate parts**, **each of which becomes an independent problem** that he **formulates for himself with the help of speech**.”

Closing remarks, further considerations

What does this mean for LLMs?

- Tokens are meaningful in three ways:
  - To the user, as **text**
  - To the model, as the material of **embeddings**
  - To the model, as the structuring of **attention**
- CoT then emerges from the **synergy** of two elements:
  - Self-attention
  - Autoregression
- This leverages the deep structure of training data
  - Most obvious in zero-shot, evident also in invalid few-shot (Wang et al., 2023)
  - Something resembling these dynamics must be latent in human language use
  - Egocentric speech in Vygotsky, but also global workspace theory (Baars, 1997)

What does this mean for humans?

- Taking language models seriously as models of language
  - Guest and Martin (2023): **multiple realisability**
    - Two clocks, one digital and one mechanical
    - Both tell the time, but it would be a mistake to consider them equivalent
      - True
  - Instead of the mechanisms, however, what if we want to understand time—
    - Then what sort of a thing must time be that this is possible?
  - Similarly, **what sort of a thing must language be that this is possible?**
- Meanwhile, reasoning in humans is itself not a settled matter
  - Not so much a question of whether, of stark contrasts between true and false
  - Instead in the spirit of Jain logic, **syāt eva**: “in some respect, certainly”

Agency and intentions in artificial intelligence

- Final section of the print-out, from Vygotsky's *Mind in Society:*
    - "… the **inclusion of signs in temporal perception** does not lead to a simple lengthening of the operation in time; rather, it creates the conditions for the development of a **single system that includes effective elements of the past, present, and future**. This emerging psychological system in the child now encompasses two new functions: **intentions and symbolic representations of purposeful action**."
- Further reading:
    - *Plans and the Structure of Behaviour* (Miller et al., 1960)
    - *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Jaynes, 1976)