Construction Grammar Evidence for How LLMs Use Context-Directed Extrapolation to Solve Tasks

Harish Tayyar Madabushi
University of Bath, U.K.
htm43@bath.ac.uk

Claire Bonial
DEVCOM Army Research Laboratory, U.S.A.
claire.n.bonial.civ@army.mil

Abstract

In this paper, we apply the lens of Construction Grammar to provide linguistically-grounded evidence for the recently introduced view of LLMs that moves beyond the 'stochastic parrot' and 'emergent Artificial General Intelligence' extremes. We provide further evidence, this time rooted in linguistic theory, that the capabilities of LLMs are best explained by a process of *context-directed extrapolation from their training priors.* This mechanism, guided by in-context examples in base models or the prompt in instruction-tuned models, clarifies how LLM performance can exceed stochastic parroting without achieving the scalable, general-purpose reasoning seen in humans. *Construction Grammar is uniquely suited to this investigation, as it provides a precise framework for testing the boundary between true generalization and sophisticated pattern-matching on novel linguistic tasks.* The ramifications of this framework explaining LLM performance are three-fold: first, there is explanatory power providing insights into seemingly idiosyncratic LLM weaknesses and strengths; second, there are empowering methods for LLM users to improve performance of smaller models in post-training; third, there is a need to shift LLM evaluation paradigms so that LLMs are assessed relative to the prevalence of relevant priors in training data, and Construction Grammar provides a framework to create such evaluation data.

1 Introduction

Understanding how Large Language Models (LLMs) solve complex tasks is a critical yet unsettled question, and the field remains divided between two primary viewpoints. One perspective characterizes LLMs as 'stochastic parrots,' which do little more than generate statistically probable outputs based on their training (Bender et al., 2021; Bender and Koller, 2020; Mitchell and Krakauer, 2023). The opposing view contends that with sufficient scale in parameters and data, LLMs exhibit 'emergent reasoning' (Brown et al., 2020a; Wei et al., 2022b; Srivastava et al., 2023a), a phenomenon claimed to be 'sparks of Artificial General Intelligence' (AGI) (Bubeck et al., 2023).

Our recent work (Tayyar Madabushi et al., 2025b) has sought to bridge this divide with an alternative framework.[1] Rather than viewing LLMs as either 'stochastic parrots' or as possessing advanced, human-like reasoning, we contend that the capabilities and limitations of these models are best explained by **context-directed extrapolation from their training priors.** In our framework, the necessary context is supplied by in-context learning examples for base models, or directly by the prompt for instruction-tuned models.

This position paper first summarizes the framework proposed in Tayyar Madabushi et al. (2025b) (Section 2). We then present our working definition of reasoning and generalization while providing linguistic examples of the generalization of constructions (Section 3). We discuss the two prevalent views of LLM capabilities along with evidence from CxG rsearch for and against each view. First, we explore stochastic parroting and present evidence of LLM success in solving difficult, non-memorizable problems that require more than next-token prediction (Section 4). Second, we explore the possibility of AGI, where we present research demonstrating that models are incapable of completing certain tasks that are trivial for humans (Section 5). This pattern, we will argue, suggests a specific shortcoming in what is termed 'advanced reasoning.' We then present new evidence from Construction Grammar (CxG) that substantiates this view (Section 6) and provides insights into the limitations of the more extreme, alternative views.

[1] Mentions of our past research have been de-anonymized after double-blind review and paper acceptance.

Proceedings of the Second International Workshop on Construction Grammars and NLP, pages 190–201

From this foundation, we turn to the problem of evaluation. We argue that even though LLMs have mastered many superficial linguistic elements, sound linguistic theory provides the necessary tools to test their deeper reasoning (Sections 7, 8). *Specifically, we demonstrate how the principles of CxG can be used to design precise tests that probe the inherent capabilities of these models, and we suggest extensions informed by usage-based theories.*

2 Context-Directed Extrapolation from Training Priors

In the framework of context-directed extrapolation, an LLM makes use of the entire prompt context to generate its output. This process is straightforward in base models, which are trained exclusively on the next-token prediction objective. For base models, the input prompt provides the sequence context from which the most probable subsequent token is generated. However, dealing with the more common models, which are additionally trained to follow instructions (instruction-tuned models), the instructions in the prompt establish a *semantic context*. This context is then used to extrapolate from relevant priors acquired during pre-training, as opposed to treating the prompt merely as a token sequence.

Specifically, for base models, while there is wide debate over how LLMs function, their capabilities and their ability to truly generalize, their capacity for *in-context learning (ICL)* is an indisputable fact (Brown et al., 2020b; Olsson et al., 2022). ICL is an ability of LLMs to learn a new task on the fly, simply by being given a few examples within the prompt. To illustrate this, Tayyar Madabushi et al. (2025b) use the example of a modified addition task. In this task, when provided with the input prompt:

$$1 + 3 = 5; 7 + 12 = 20; 8 + 3 =$$

LLMs, trained only on the next token prediction task, can infer the novel pattern $(a + b + 1)$ from the examples and produce the correct, non-obvious answer of 12.

In Tayyar Madabushi et al. (2025b), we derive the notion that ICL is a method of solving tasks by extrapolating from pre-training priors from a convergence of several distinct theories. We note that research consistently supports this view, whether by directly linking ICL to the distributions in pre-training data (Chan et al., 2022; Hahn and Goyal, 2023), or by explaining it through frameworks like Bayesian inference (Zhang et al., 2023; Xie et al., 2021) and Probably Approximately Correct (PAC) learning (Li et al., 2023b). This conclusion is reinforced by other studies that liken ICL to fine-tuning (Dai et al., 2023) or show that it can implicitly perform gradient descent, a process linked to meta-learning (Akyürek et al., 2023; Li et al., 2023a; Zhang et al., 2024; Von Oswald et al., 2023). Ultimately, we argue that regardless of the specific mechanism, all existing research indicates that ICL fundamentally relies on priors from pre-training data, with the in-context examples serving to guide the model toward the relevant priors needed for the task at hand.

2.1 Context-Directed Extrapolation in Base vs Instruction-Tuned Models

A critical observation is that LLMs trained solely on next-token prediction (i.e. base models) are by construction nothing more than sequence completion engines. However, *these base models cannot solve tasks that require abstract reasoning without being provided with examples through in-context learning (ICL)* (Lu et al., 2023). Consider, for instance, the following logical deduction problem from the Big-Bench benchmark:

> Question: On a shelf, there are five books: a gray book, a red book, a purple book, a blue book, and a black book. The red book is to the right of the gray book. The black book is to the left of the blue book. The blue book is to the left of the gray book. The purple book is the second from the right.
>
> Targets: 'The gray book is the leftmost.': 0; 'The red book is the leftmost.': 0; 'The purple book is the leftmost.': 0; 'The blue book is the leftmost.': 0; 'The black book is the leftmost.': 1

Base models fail on such reasoning tasks when presented without examples, however, they can solve this task when presented with a prompt that includes examples. Central to our augment is the fact that, instruction-tuned models can solve this task without examples based purely on a description of the task (Lu et al., 2023).

Context-directed extrapolation from training data priors offers a unifying framework to explain both the capabilities and, importantly, the limitations of LLMs: In base models, the in-context

examples provide the context direction to allow the model to infer and solve the relevant task at hand. In instruction-tuned models, however, the process of instruction tuning allows the models to interpret the semantic context of the prompt without explicit examples, and similarly direct extrapolation. We contrast our framework with that of stochastic parroting in Section 4.1.

2.2 Extrapolation and Grounding

An important implication of context-directed extrapolation is that *it allows for a limited form of grounding.* By this we do not mean that models achieve grounding in the human sense of connecting language to embodied experience. Rather, because the mechanism involves extrapolating from priors activated by the prompt, information that is not explicitly present in surface form can nevertheless become available to the model. For example, when confronted with a nonce verb whose definition is provided in the prompt, the model can project that meaning into novel contexts and apply it productively. Indeed, this same process allows models to respond effectively in tasks such as the Sally–Anne test (Wimmer and Perner, 1983), enabling models to succeed on certain Theory of Mind evaluations that would be inaccessible to 'stochastic parroting' (Kosinski, 2024).

This is categorically different from stochastic parroting. A purely parroting mechanism cannot accommodate genuinely novel input that falls outside its memorized distribution. The fact that LLMs can extend prompt-based definitions, apply abstract patterns, and generate context-appropriate interpretations indicates that extrapolation yields access to extrapolatable information that is not reducible to surface statistics. In this sense, context-directed extrapolation provides a pathway to limited grounding, albeit one constrained by the priors in training data and the context supplied at inference time.

2.3 A Mechanistic Basis for Context-Directed Extrapolation

To understand the underlying mechanics of this capability in LLMs, in Tayyar Madabushi et al. (2025b), we first point to the foundational work of Olsson et al. (2022), who systematically showed that LLMs could complete abstract patterns with random tokens (e.g., given a sequence $[A][B]...[A]$, LLMs correctly respond with $[B]$). While this compellingly refutes the 'stochastic parrot' notion by suggesting an algorithmic capability, we introduce a crucial caveat from recent research (Niu et al., 2025): this pattern-matching ability degrades significantly as the tokens become less frequent in the pre-training data. This finding demonstrates that even this seemingly abstract skill is fundamentally tethered to the model's training priors.

We then argue that this powerful, data-dependent pattern-matching ability is the same core mechanism that allows LLMs to solve more complex tasks via ICL. This view is substantiated by evidence showing that ICL remains effective even when the labels in the examples are manipulated, such as being flipped between positive and negative or replaced with entirely unrelated words like 'Foo' and 'Bar' for a sentiment classification task (Wei et al., 2023). Therefore, in Tayyar Madabushi et al. (2025b), we conclude that ICL, while impressive, is a sophisticated but ultimately constrained process. We argue that because its operation is always guided by the user-provided examples and bound by the limits of its training data, it fails to meet the requirements for advanced, generalizable reasoning. In this setting, the model never gains true 'agency,' as its performance is always a function of the input, preventing it from making the leap from guided pattern-matching to unguided, human-like cognition.

3 Construction Grammars and Generalization

In this section, we outline our definition of human-like reasoning and provide insights into such reasoning in linguistic settings from CxG. Following our work in Tayyar Madabushi et al. (2025b), we embrace a definition of advanced reasoning that requires mastery and understanding of knowledge taken from one set of members instantiating a class, and then generalization and application of that knowledge to a novel set of items. In terms of CxG, constructions (defined as pairings of meaning and form at any level—morphological, lexical, phrasal (Goldberg, 2003; Hoffmann and Trousdale, 2013)) should be thought of as classes, and members are certain instantiations or realizations of that construction/class. Psycholinguistic evidence from child language acquisition demonstrates that children acquire frequently-heard constructions first and initially only use the member instantiation that they have heard (Tomasello, 2009). For example, a child's first Resultative construction will likely involve the high-frequency verb "make" along with

other lexical items the child is frequently exposed to: "Mommy made me mad." An 'understanding' of this construction is achieved when a speaker can recognize the similarity of other instantiations of this construction, which generally involve some kind of verb of change-of-state semantics within the structure (e.g., "Berries turned me blue!"). True generalization of the construction requires abstracting and applying knowledge of the construction from heard instantiations to novel items—in this example, novel instantiations of the phrasal constructions where the individual lexical items have likely not been experienced within that construction before: e.g., "The dog barked me awake."

Over the next sections, we will discuss the more extreme viewpoints of LLM performance as either "stochastic parrots" or advanced general intelligence. In each section, we will close with relevant research from CxG. Our review of work on CxG will reveal a mixed picture: models can make the required generalization in some instances, but fail in others. However, based on our framework of context-directed extrapolation, these seemingly contradictory performances become explainable.

4 LLMs are NOT Stochastic Parroting

While the 'Stochastic Parrots' paper from Bender et al. (2021) rightly identifies the risks of bias propagation in large-scale models, its claim that these models merely generate the next most likely token is demonstrably false, as we will show. We define stochastic parroting as the mechanism of generating the *precise* statistically most likely next token given the immediate input sequence. In this view, an instruction is merely more text to be completed. In the following sections we contrast this view, with our view that LLMs solve tasks using context-directed extrapolation from training priors.

4.1 Stochastic Parroting vs. Context-Directed Extrapolation

Functional Commonalities. *From the perspective of the performance of base models, there is no functional difference between context-directed extrapolation and stochastic parroting.* Base models consistently fail tasks such as the one described in Section 2.1 when presented without examples. One can argue that the examples simply form a long context, where the correct answer is the most probable sequence completion. This makes both theories appear to describe the same mechanism: the model completes a given sequence based on statistical patterns. Consequently, the two views are indistinguishable when analyzing this alone.

Most LLMs in wide use, such as public chat models, undergo instruction fine-tuning after their initial pre-training so they can 'understand' and follow instructions presented within their prompts (Wei et al., 2022a). This additional training, however, complicates their evaluation. It becomes difficult to tell whether a model's success on a new task is a sign of genuine emergent reasoning or simply a consequence of its training on similar tasks.

This issue was explored in a systematic study by Bigoulaeva et al. (2025), who fine-tuned over 90 models and demonstrated that the performance of instruction-tuned models is strictly correlated with that of base models. This suggests a single underlying mechanism is at play in both. Building on this, in Tayyar Madabushi et al. (2025b), we argue that this mechanism is context-directed extrapolation from pre-training data. We propose that *instruction-tuning simply allows the model to perform the same kind of extrapolation from a natural language prompt, rather than needing the explicit in-context examples that base models require.*

Functional Difference. The functional difference between these two views becomes apparent with instruction-tuned models. A base model, provided with enough examples, generates the correct output because it becomes the most probable completion of that long sequence. In contrast, the context-directed extrapolation view posits that instruction-tuning enables a different mechanism. It allows the model to interpret an instruction not as a literal sequence to be continued, but as a directive to construct an implicit context for a task. This allows the model to activate relevant priors (just as examples do for base models) from its pre-training data to perform the task specified by the prompt, rather than simply completing the text of the prompt itself. *Critically, the evidence for this distinction is that instruction-tuned models can solve the logical deduction (and similar) problems presented in Section 2.1 without any examples (Lu et al., 2023). This phenomenon cannot be explained by stochastic parroting, but is directly accounted for by context-directed extrapolation.*

This distinction becomes even more stark in tasks involving novel words, as this eliminates the model's ability to rely on pre-existing statistical associations. The Winodict benchmark (Eisensch-

los et al., 2023), for instance, modifies Winograd schemas by replacing a critical verb with a nonce word defined within the prompt. Consider:

> The verb 'to plest' means to be scared of... The city councilmen refused the demonstrators a permit because they **plested** violence."

To correctly resolve the pronoun "they," the model cannot use any stored knowledge about the word "plest." It must parse the definition provided in the prompt and apply that meaning to the sentence. The success of models on this task provides compelling evidence that the model is not merely predicting a statistically likely token, but is using the in-prompt definition to build a context and reason accordingly. The ability of LLMs to successfully solve this task is directly explained by context-directed extrapolation as it allows models to extropolate meaning from context. In contrast, a pure stochastic parroting mechanism based on predicting the next likely token along *cannot* account for this ability. As discussed previously, unlike base models (Section 2), instruction-tuned models succeed on tasks such as logical deduction without explicit examples (Section 4.1), a result that cannot be explained by stochastic parroting. The Winodict benchmark illustrates this distinction especially clearly. By replacing a key verb with a nonce word defined only within the prompt, the task prevents the model from relying on stored associations. Yet models are still able to resolve the pronoun correctly by projecting the definition into novel contexts (Section 6), a behavior that cannot be accounted for by a purely stochastic parroting mechanism. Indeed, mechanistic studies exploring 'induction heads' further support this view (Section 2.3). In what follows, we turn to CxG research relating to the notion of stochastic parroting.

4.2 CxG & Stochastic Parroting

There is relevant research demonstrating first that information on certain constructions is present in pre-training data, such that models may rely on stochastic parroting to provide the impression of proficiency with the constructions of the language. Tayyar Madabushi et al. (2020) probe a variety of BERT-based models for access to knowledge of several constructions proposed in Dunn (2017). In this work, Tayyar Madabushi et al. (2020) test BERT models on their ability to distinguish sentences that are instances of a given construction from those that are not. Alongside the base model, the authors trained several BERT "clones" with additional exposure to constructional information, varying the frequency of constructions during pre-training so that some clones saw high-frequency items and others saw low-frequency ones. The expectation was that clones trained on rarer constructions would benefit most, since such items were unlikely to appear often in the original pre-training data. However, the results showed little improvement over the base BERT model, leading the authors to conclude that constructional knowledge was already accessible to BERT. It is worth noting, though, that the constructions targeted were identified in a data-driven way using the methods proposed by Dunn (2017), and typically involved fixed lexical items. More schematic phrasal patterns, such as argument structure constructions (Goldberg, 1992), were not included. As a result, it is plausible that the constructions tested were already present in the base model's pre-training corpus at sufficient frequencies to allow strong performance through context-directed extrapolation rather than deeper generalization.

In Bonial et al. (2025), we provide evidence from Multi-Word Expression (MWE) constructional templates, which demonstrate that LLM abilities must go beyond stochastic parroting. While it is expected that the conventional meanings of common MWEs such as "spill the beans" (meaning *to tell a secret*) would be encoded in LLMs, we show that they can use in-context learning and a single example to teach models entirely novel MWEs, such as "winking at pringles" (meaning *to indulge in frivolity*). Not only can the models explain the meaning of a new (non-verbatim) usage of the novel MWE, GPT-4o is able to answer targeted questions about the new usage, for example that "a meeting where everyone is winking at a bunch of chips" is probably not a productive use of time. Thus, this research demonstrates that ICL in the form of a definition of the novel MWE and single usage example facilitates performance that cannot be dismissed as stochastic parroting, as these novel MWEs are guaranteed not to be found in pre-training data. Table 1 presents further examples from this work.

5 LLMs are also NOT AGI

On the other end of the spectrum, some research argues that models, when scaled up, gain abilities

MWE Definition & Example	Novel Usage	GPT4o Interpretation	Annotator Interpretation
Expression: **to send one's LLM (instead)** Meaning: Not worth one's time. Usage: There's no way I'm attending that meeting, I'll probably send my LLM	The bank teller seems to have sent his LLM to work today.	The statement means that the bank teller seems disengaged or inattentive, implying they are not putting in effort or paying attention, as if they had metaphorically 'sent their LLM' to do their job instead of being present themselves.	Annotator 1: The bank teller is not putting effort into their work, relying on rote memorization of their tasks, thereby making many mistakes and potentially misunderstanding some of their interactions.
Expression: **dance with the deer** Meaning: to have a deep and meaningful conversation, especially about something one is passionate about. Usage: When they were planning their research, they often danced with the deer.	I don't think they were dancing with the same herd of deer, possibly not even the same species.	The statement implies that the individuals involved were having deep and meaningful conversations, but their passions, ideas, or perspectives were significantly different—perhaps even fundamentally incompatible. It suggests a lack of alignment in their discussions or goals.	Annotator 1: Although all parties are interested and passionate about their topic of conversation, they have not reached common ground as to what the underlying meaning and implications of what the others are saying.

Table 1: Sample evaluation results from Bonial et al. (2025), in which we tested the ability of LLMs to generalize to novel MWEs, given and defined in the first column. Note that models could not have been exposed to these specific MWEs during pretraining, yet the interpretation of the novel usage (second column) is quite similar to that of human annotators.

akin to high-level human reasoning (Brown et al., 2020b; Wei et al., 2022b; Srivastava et al., 2023b; Lu et al., 2024; Wei et al., 2024).

In Tayyar Madabushi et al. (2025b), we argue that high-level reasoning is demonstrated only when a model solves tasks it was not explicitly trained for, distinguishing genuine cognitive application from simpler forms of understanding (Krathwohl, 2002). In line with Chollet (2019), we note that a model trained solely to master a single task such as chess, even to a superhuman level, does not exhibit the kind of reasoning that matters here, since it is not generalizing knowledge to a truly new domain. To make this distinction precise, here and in Tayyar Madabushi et al. (2025b), we adopt the framework of Krathwohl (2002), a revision of Bloom's original taxonomy of educational objectives (Bloom et al., 1956), which defines advanced reasoning as the ability to apply and extend knowledge beyond familiar instances to novel contexts.

To argue that LLMs are not performing advanced reasoning, we point to two key shortcomings: models' tendency for hallucination and their failure on seemingly simple tasks. First, LLM hallucinations—outputs that are not aligned with reality—are cited as a major piece of evidence against advanced reasoning (Huang et al., 2025). We argue this phenomenon should not be confused with human confabulation, as there is no evidence for LLM agency (Lu et al., 2024), and these errors can be traced to the model defaulting to statistical patterns from its training data when the prompt's context is insufficient (Hanneke et al., 2018). Second, we highlight that LLMs often fail at tasks that are trivial for humans (Nezhurina et al., 2025). For instance, even top models perform poorly on clinical psychology *faux-pas* tests compared to children (Shapira et al., 2023), and they are significantly outperformed by non-expert humans in simple AI planning domains like Blocksworld (Valmeekam et al., 2023).

5.1 CxG & Advanced Reasoning

From a constructional perspective, Li et al. (2022) probe models of varying sizes for access to knowledge of purely schematic argument structure constructions, including DITRANSITIVE, RESULTATIVE, CAUSED-MOTION, and REMOVAL constructions. In their design, the authors adopt a sorting task where both human participants and models are asked to judge sentence similarity. The dataset is deliberately constructed so that the constructions under investigation are expressed through a range of lexical verbs. Importantly, the verbs chosen to instantiate different constructions belong to overlapping semantic classes—for instance, verbs such as *cut* and *slice*. This setup allows them to test if participants and models cluster sentences on the basis of verb meaning, as traditional generative grammar would suggest, or if they recognize the broader constructional pattern. The findings reveal a sharp divergence depending on model scale. MiniBERTas

(Warstadt et al., 2020), a model with only one million parameters, aligns sentences primarily by verb-level semantics, whereas the much larger RoBERTa model (30B parameters; (Liu et al., 2021)) instead groups them in line with constructional semantics. While the authors do not point to this as evidence of advanced reasoning *per se*, they do conclude that larger models perform like native speakers while smaller models perform more like second language learners. However, we emphasize that these results can also be interpreted as larger models successfully extrapolating from pre-training priors that the smaller models do not have.

Additional studies using CxG highlight similar limits to the reasoning abilities of LLMs. Weissweiler et al. (2022a) examine the Comparative-correlative construction (e.g., The higher you fly, the harder you fall) as a test case for whether models can capture both its syntactic properties and its associated semantic meaning. Their methodology first targets the syntax by evaluating whether models can reliably recognize instances of the construction in natural corpus data and in controlled, synthetic examples. On this task, several BERT-based models perform well, successfully identifying and discriminating the construction. Such results are not unexpected given that the Comparative-correlative includes fixed lexical items in key structural positions. The crucial question, however, is whether models can also handle the semantics of the construction. To probe this, the authors evaluate performance on a downstream task that requires reasoning about the correlational meaning encoded by the construction. Here the models perform poorly, especially on nonce words, with accuracy barely above chance, indicating that while BERT-based models can recognize the formal template of the Comparative-correlative, they fail to grasp its interpretive content. We highlight that this failure on nonce words is, yet again, indicative of context-directed extrapolation. Similar research evaluating both formal recognition and semantic interpretation of the Causal-excess construction underscores this finding—models can pick out the construction but perform poorly on semantic understanding tests in the form of downstream questions (Zhou et al., 2024).

6 CxG & Context-Directed Extrapolation

In Bonial and Tayyar Madabushi (2024a), we find that even the largest models available at the time (GPT-3.5 and GPT-4) are restricted to recognizing substantive constructions (with fixed words), whereas schematic constructions (without fixed words) elude recognition of either form or meaning. In that research, we collect and leverage the CoGS dataset (Bonial and Tayyar Madabushi, 2024b), which includes approximately 500 corpus instances of 10 different phrasal constructions of varying schematicity (i.e. some constructions are fully fixed words, while others are argument structure constructions with no fixed words). The corpus includes relatively frequent constructions, but is limited to instantiations of those constructions that are not the most frequent, entrenched instantiations. For example, the Ditransitive construction instances do not include usages with the verb "give," which is the most frequent verb to instantiate this construction: "He gave me a book." Instead, CoGS Ditransitives include only cases where the lexical semantics of the instantiating verb do not inherently include transfer semantics: "He poured her a martini." In other words, the constructions in CoGS have high type frequency, but these particular instantiations have relatively low token frequency. Nonetheless, the fixed words of the substantive constructions facilitate tapping into the appropriate pre-training data in order to recognize the construction (but not necessarily a deeper understanding, as suggested by (Weissweiler et al., 2022b)). In contrast, although the schematic argument structure constructions are the most fundamental constructions of the English language with very high type frequency (Goldberg, 1992), the models are not able to apply generalized formal and semantic properties of the construction to novel instantiations. This suggests that models can extrapolate to a point to account for relatively infrequent, creative instantiations of constructions, but the level of generalization required for recognizing the structural slots and associated semantics of argument structure constructions is beyond model abilities.

Similarly, Scivetti et al. (2025a) find that the "human-scale" BabyLM demonstrates strong formal knowledge of the Let-alone construction, but no understanding of the associated scalar semantics. Further experiments on the templated evaluation dataset first remove all Let-alone constructions from pre-training data, as well as filtering all related constructions (e.g., Much-less). The authors find that this does not change BabyLM performance on formal recognition of the construction.

The authors then remove all individual "let" and "alone" tokens from pre-training, and this significantly degrades performance on formal recognition, leading us to conclude that the model is drawing on compositional, lexical information of the individual words as opposed to the form of the phrasal construction as a whole. Thus, this research underscores the notion that generalizing the semantics associated with syntactic slots of a construction eludes models, and casts further doubt on whether or not even the formal features learned by models are generalized at the constructional level or limited to lexical, collocational features.

In Scivetti et al. (2025b), we provide further evidence of both the extrapolation abilities of LLMs when it comes to constructions, as well as the limits of their generalization abilities. We leverage a subset of the CoGS dataset described previously, specifically using corpus constructional usages as the premises for Natural Language Inference (NLI) triples in which templates are leveraged to semi-automatically generate entailed, neutral, and contradicted hypotheses. In leveraging an NLI task, we test downstream, functional understanding of the CoGS constructions, which again are of relatively high type frequency (e.g., Ditransitive: "he gave me a book") but the instantiations of those constructions are relatively low token frequency (e.g., Ditransitive: "he poured her a martini"). Interestingly, although we found that models failed to recognize more schematic constructions in Bonial and Tayyar Madabushi (2024a), in our subsequent NLI research (Scivetti et al., 2025b), we find that the largest models available (GPT-4 and 4o) perform comparably on the constructional NLI and Stanford NLI, ostensibly demonstrating that the models are able to draw inferences correctly over the constructional premises.

In Scivetti et al. (2025b), we then conduct follow-on experiments where the models are evaluated on a new set of NLI triples involving schematic constructions that are not the high type-frequency constructions of CoGS but are formally indistinguishable. For example, the Depictive construction (e.g., "She bought the apples fresh") has the same syntactic slots as the Resultative (e.g., "She hammered the metal flat"), but distinct semantic roles associated with the slots. We then test the same models on NLI triples involving the formally identical but semantically distinct premises, and find that model performance drops substantially. We posit that this research therefore demonstrates the limits of extrapolation as opposed to true generalization of the meaning of constructions. The strong performance on the original CoGS premises shows that models can effectively extrapolate from pre-training data, which is ample for these high type-frequency constructions. However, the degradation in performance on the formally identical but semantically distinct premises shows that because models are extrapolating from the higher-frequency constructions, they will perform the task (incorrectly) according to those priors when faced with lower-frequency constructions that the model seems unable to distinguish.

Finally, in the second set of results from Bonial et al. (2025), we extend this line of evidence. We show that while LLMs can learn and use entirely novel MWEs when definitions are provided in the prompt (as discussed in Section 4.2, see also Table 1), performance degrades when models are asked to reason across multiple MWEs at once. For example, given novel MWE definitions for "drown the cables" (an invented MWE defined as *to sever or overwhelm communication*) and "dance with the deer" (*to have a deep, meaningful conversation*), the models were evaluated for their ability to reason about the semantic interaction of the two MWEs in a novel usage involving both MWEs. Human annotators were able to do this consistently, but even advanced models like GPT-o1 and GPT-4o faltered. This demonstrates the limits of context-directed extrapolation, which enables models to extend clear, explicit definitions to new usages (as shown in this work for single MWE), but that the mechanism struggles once the links between constructions become less direct.

7 Discussion and Implications

Context-directed extrapolation explains LLM behavior as the use of priors activated by prompt context. Because of this, the very same capability, such as apparent Theory of Mind, will be observed when the relevant priors are strong, but absent or much weaker when priors are sparse. The same holds for grounding: it will appear when relevant information is easily extrapolatable from context and fail when it is not. This means that evaluation must carefully distinguish between cases where models are simply drawing on rich priors and cases where success would require true human-like generalization. Counterfactuals are ideal for making this

distinction, since they force the model to reason beyond memorized or extropable priors, and LLMs consistently fail on such tests despite succeeding on superficially similar ones (Wu et al., 2024).

For decades NLP research sought to build pipelines around symbolic templates and formal reasoning systems. Over time the pipeline itself became an end goal. LLMs now shift this landscape by allowing us to fill templates more easily and then use established resources, such as AMR (Banarescu et al., 2013; Bonial et al., 2018) or frame semantics (Fillmore et al., 2012), to support reasoning processes in systematic, verifiable ways (e.g., Tayyar Madabushi et al. (2025a)). *Given that models continue to struggle with more advanced reasoning tasks, it is increasingly important to see them as an interface between the complexity of language and downstream formal reasoning rather than as reasoning systems themselves.*

CxG is a particularly strong testbed for this view. It allows us to probe the line between semantics and syntax and to see where models succeed because of exposure to canonical patterns of language and where they fail to generalize. Because there is already extensive evidence of how humans learn and extend constructions (e.g., Tomasello (2009)), CxG provides the right framework to compare human generalization against model extrapolation and to identify the precise gaps that remain. Usage-based theories of learning, such as Frame Semantics (Fillmore et al., 2012), can also be incorporated into the design of systems. We need an interface between the lexical, surface form of text and the higher-level structures of meaning, and LLMs get us part of the way there by exploiting priors in context. Usage-based theories can then provide the conceptual tools to take us the rest of the way, enabling a more systematic connection between linguistic form, meaning, and true human-like generalization.

In sum, LLMs offer a powerful but incomplete bridge between raw text and meaning. Their strengths lie in exploiting priors through context, but their limitations highlight the need for theoretical frameworks that go further. Usage-based approaches such as CxG provide exactly this. By combining the empirical reach of LLMs with the conceptual depth of usage-based theory, we can move toward a more systematic account of how form and meaning connect, and build systems that move towards human-like generalization.

8 Conclusions & Recommendations

The 'stochastic parrots' versus 'sparks of AGI' debate has become a roadblock to clarity in LLM performance and avenues to advance performance. This paper offers a more productive, middle-ground theory, providing a theoretically-grounded argument for context-directed extrapolation from training priors. The implications of this are significant: it provides a coherent explanation for the seemingly idiosyncratic and unpredictable strengths and weaknesses of LLMs, demystifying phenomena like hallucinations and, as we have detailed, clarifying their contradictory performance on CxG tasks.

Second, it suggests that meaningful improvements can be achieved not just through scale, but through better methods of directing this extrapolation via prompting and fine-tuning. This understanding demands that we re-evaluate how we improve language models. The prevailing paradigm, which chases unpredictable 'emergent' abilities by scaling up models and data, is not the only way forward. Our work suggests a more principled approach: focusing on the 'context' and 'priors' of the reasoning equation to achieve significant performance gains. This shift opens exciting new avenues for research beyond a simple reliance on scale. It points toward a more sustainable path to innovation, focused on augmenting models in novel ways, such as by equipping them with external memory.

Finally, and most urgently, our work demands a paradigm shift in how we evaluate LLMs. To genuinely measure a model's reasoning, we must move past benchmarks that might be tainted by training data or that only test for simple extrapolation. The goal should be to assess a model's ability to generalize and apply knowledge, not just to understand or remember it (in terms of Bloom's taxonomy (Bloom et al., 1956)). We therefore recommend a new focus on out-of-distribution evaluation, using grounded linguistic theory like CxG for language tasks. By testing models on examples that are grammatically valid but highly unlikely to be in the training data, such as formally identical but semantically distinct constructions, we can clearly distinguish between true generalization and mere pattern-matching.

Taken together, these recommendations call for a shift from chasing scale to building a linguistically principled science of evaluation and improvement, where CxG and related usage-based theories play a central role.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Irina Bigoulaeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. The inherent limits of pretrained llms: The unexpected convergence of instruction tuning and in-context learning capabilities.

Benjamin S Bloom et al. 1956. Taxonomy of. *Educational Objectives*.

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 2018. Abstract meaning representation of constructions: The more we include, the better the representation. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Claire Bonial, Julia Bonn, and Harish Tayyar Madabushi. 2025. Dancing with deer: A constructional perspective on mwes in the era of llms.

Claire Bonial and Harish Tayyar Madabushi. 2024a. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*, pages 1–40.

Claire Bonial and Harish Tayyar Madabushi. 2024b. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Jonathan Dunn. 2017. Computational learning of construction grammars. *Language and cognition*, 9(2):254–292.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Charles J. Fillmore, Russell Lee-Goldman, and Russell Rhomieux. 2012. The framenet constructicon. In Hans C. Boas and Ivan A. Sag, editors, *Sign-Based Construction Grammar*, number 193 in CSLI Lecture Notes, pages 309–372. CSLI Publications, Stanford, CA.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Adele Eva Goldberg. 1992. *Argument structure constructions*. University of California, Berkeley.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Steve Hanneke, Adam Tauman Kalai, Gautam Kamath, and Christos Tzamos. 2018. Actively avoiding nonsense in generative models. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 209–227. PMLR.

Thomas Hoffmann and Graeme Trousdale. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023a. Transformers as algorithms: generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023b. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning?

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2025. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models.

Jingcheng Niu, Subhabrata Dutta, Ahmed Elshabrawy, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. Illusion or algorithm? investigating memorization, emergence, and symbolic processing in in-context learning.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads.

Wesley Scivetti, Tatsuya Aoyama, Ethan Wilcox, and Nathan Schneider. 2025a. Unpacking let alone: Human-scale models generalize to a rare construction in form but not meaning. *arXiv preprint arXiv:2506.04408*.

Wesley Scivetti, Melissa Torgbi, Austin Blodgett, Mollie Shichman, Taylor Hudson, Claire Bonial, and Harish Tayyar Madabushi. 2025b. Assessing language comprehension in large language models using construction grammar.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023a. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, et al. 2023b. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs].

Harish Tayyar Madabushi, Taylor Pellegrin, and Claire Bonial. 2025a. Generative framenet: Scalable and adaptive frames for interpretable knowledge storage and retrieval for llms powered by llms. *Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning@ COLING 2025*, page 107.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. Cxgbert: Bert meets construction grammar. *arXiv preprint arXiv:2011.04134*.

Harish Tayyar Madabushi, Melissa Torgbi, and Claire Bonial. 2025b. Neither stochastic parroting nor agi: Llms solve tasks through context-directed extrapolation from training data priors.

Michael Tomasello. 2009. The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. 2020. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2010.05358*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2024. Larger language models do in-context learning differently.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022a. The better your syntax, the better your semantics? probing pretrained language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022b. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. *arXiv preprint arXiv:2210.13181*.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2024. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R. Mortensen, and Lori Levin. 2024. Constructions Are So Difficult That Even Large Language Models Get Them Right for the Wrong Reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 3804–3811, Torino, Italia. ELRA and ICCL.