Analogical Structure, Minimal Contextual Cues and Contrastive Distractors: Input Design for Sample-Efficient Linguistic Rule Induction

Chunyang Jiang[1,2], **Paola Merlo**[1,2]

[1]Idiap Research Institute, Switzerland [2]University of Geneva, Switzerland
Firstname.Lastname@unige.ch

Abstract

Large language models achieve strong performance on many tasks, but their training makes it hard to see which properties of the input support efficient linguistic rule learning. We ask how three cognitively-inspired principles of input design support sample-efficient linguistic rule induction: analogical structure, contrastive learning, and minimal contextual cue. We also ask how their effects compare to those of LLMs on the same controlled tasks. We implement these principles in structured sentence completion tasks that test English verb alternations. Lightweight models trained on hundreds to one-thousand such examples learn the alternation rules with high F1 on these tasks. Ablation studies show that analogical organisation is the main driver of sample efficiency, and contrastive distractors and minimal context help further gains. We also evaluate zero- and few-shot LLMs on the same tasks. In this controlled setting, the lightweight models reach higher F1 with far fewer task-specific data. We treat this contrast as a comparison between learning regimes rather than a general verdict on LLMs. Our results show that careful input organisation supports sample-efficient learning of linguistic rules and reveals distinct learning signatures for trained lightweight models and prompted LLMs.

1 Introduction

Analogical reasoning, recognising parallelism of structural relationships across different contexts, enables efficient learning by abstracting patterns from minimal examples and transferring knowledge to novel situations (Lake et al., 2017). However, this principle has faced persistent scalability challenges across computational approaches: early symbolic systems required extensive knowledge engineering (Forbus et al., 1995), neural word embeddings showed inconsistent performance (Mikolov et al., 2013), and even transformer-based models like BERT show mixed capabilities (Ushio et al., 2021; Thrush et al., 2020). While recent systems achieve strong performance (Yasunaga et al., 2024; Jiayang et al., 2023; Wijesiriwardene et al., 2023), they still require substantial computational resources or show inconsistent results across complexity levels.

Rather than scaling analogical reasoning systems, we operationalise analogical principles through strategic input organisation. Research shows that processing constraints can optimise attention allocation (Christiansen and Chater, 2016), while contrastive learning frameworks show how systematic positive-negative comparisons improve discriminative learning (Chen et al., 2020; He et al., 2020). This suggests that structural organisation, not architectural scaling, may unlock analogical efficiency.

In this paper, we ask how input organisation can support sample-efficient linguistic rule induction. We also ask how the behaviour of lightweight models that use such organisation compares to that of large language models on the same controlled tasks. We develop a computational approach that organises input data into analogical paradigms through three cognitively-inspired strategies: (i) *Analogical Structure*, the systematic paradigmatic relationships that support pattern recognition; (ii) *Contrastive Learning*, the systematically designed distractors that enable discriminative learning; and (iii) *Minimal Contextual Cues*, the implicit annotations providing essential semantic information without explicit labelling.

We evaluate this setup using English verb alternations (Levin, 1993) through structured completion tasks based on Blackbird Language Matrices (Merlo, 2023). These tasks require identifying correct sentence completions from systematically organised contexts with contrastive alternatives. Figure 1 illustrates our approach: contexts encode analogical mappings across parallel

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 485–500
March 24-29, 2026 ©2026 Association for Computational Linguistics

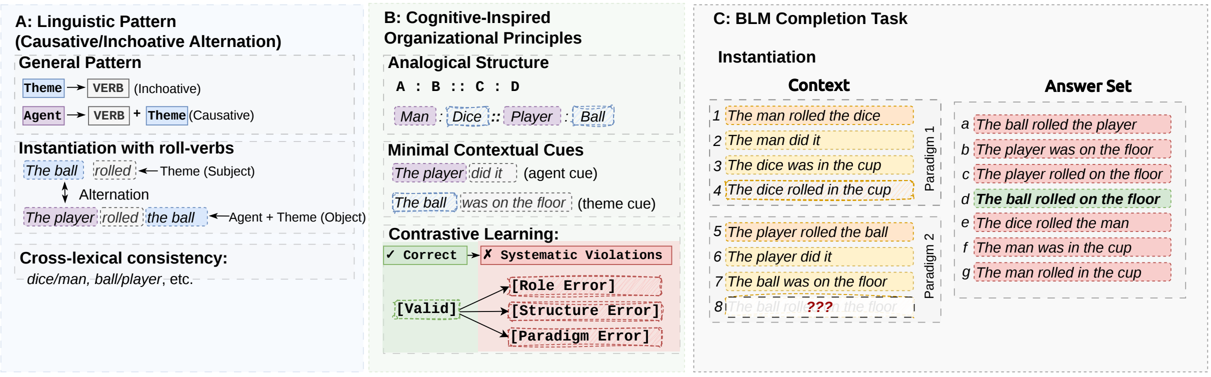

A: Linguistic Pattern (Causative/Inchoative Alternation)
General Pattern
Theme → VERB (Inchoative)
Agent → VERB + Theme (Causative)
Instantiation with roll-verbs
The ball rolled ← Theme (Subject)
Alternation
The player rolled the ball ← Agent + Theme (Object)
Cross-lexical consistency:
dice/man, ball/player, etc.
B: Cognitive-Inspired Organizational Principles
Analogical Structure
A : B :: C : D
Man : Dice :: Player : Ball
Minimal Contextual Cues
The player did it (agent cue)
The ball was on the floor (theme cue)
Contrastive Learning:
✓ Correct
✗ Systematic Violations
[Valid]
[Role Error]
[Structure Error]
[Paradigm Error]
C: BLM Completion Task
Instantiation
Context
Answer Set
1 The man rolled the dice
2 The man did it
3 The dice was in the cup
4 The dice rolled in the cup
Paradigm 1
5 The player rolled the ball
6 The player did it
7 The ball was on the floor
8 The ball rolled on the floor ???
Paradigm 2
a The ball rolled the player
b The player was on the floor
c The player rolled on the floor
d The ball rolled on the floor
e The dice rolled the man
f The man was in the cup
g The man rolled in the cup

Figure 1: Analogical paradigm organisation for sample-efficient linguistic rule learning. (A) Causative/inchoative alternation pattern showing systematic Agent↔Theme mapping in English *roll*-class verbs, with cross-lexical consistency across multiple verb instances. (B) Three cognitive-inspired organisational principles: analogical structure enables cross-paradigm pattern recognition (*Man:Dice :: Explorer:Mat*), contrastive learning provides discriminative boundaries through systematic constraint violations, and minimal contextual cues offer semantic scaffolding without explicit labelling. (C) Implementation through structured completion tasks where models must integrate all three principles to identify correct answer **d** from systematically designed alternatives, each testing specific aspects of analogical reasoning capability. (Distractor taxonomy details in Table 1).

paradigms (e.g., *Man:Dice :: Exploring:Mat*), requiring models to abstract structural and semantic role relationships rather than memorise surface forms.

Results demonstrate sample efficiency: lightweight models (BERT+CNN, $\sim 0.5M$ parameters) trained on only 100 structured examples achieve $F1 = 0.95$, and, in this setting, is higher than zero-shot GPT-o3 ($F1 = 0.87$) on the same tasks. Cross-phenomenon validation confirms robustness beyond causative constructions.

This paper makes three contributions. First, we introduce a computational approach that organises input into analogical paradigms for sample-efficient linguistic rule induction. Second, we provide empirical evidence that this approach enables competitive performance for lightweight models with minimal training examples. Third, we show that, in this controlled setting, lightweight models reach higher F1 than zero-and few-shot LLMs on the same tasks, under a different learning regime.

In all experiments, lightweight models are finetuned on $100-1000$ task-specific examples. LLMs are evaluated in zero-shot and few-shot prompt-based settings without parameter updates. We therefore treat this as a comparison between learning regimes rather than a general verdict on LLM capabilities.

2 Data and Task Design

Our data consist of structured sentence completion tasks based on Blackbird Language Matrices (Merlo, 2023). We extend this framework to instantiate three cognitively inspired principles of input design. This section describes our task framework, linguistic domain, and distractor design methodology.

2.1 Task Framework

Our completion tasks organise linguistic data into analogical paradigm structures. Each instance contains a structured *Context* comprising two parallel paradigms encoding analogical relationships (Pattern A ↔ Pattern B), supplemented with minimal contextual cues that provide semantic scaffolding without explicit labelling. It also contains a systematic *Answer Set*: One correct completion that maintains established patterns, and six systematically designed distractors that violate specific structural, semantic, or paradigmatic constraints.

Success requires integrating our three cognitive principles. Models must recognise analogical mappings across paradigms (analogical structure), discriminate between valid and invalid transformations (contrastive learning), and use semantic scaffolding from minimal soft annotations (minimal contextual cues).

2.2 Linguistic Test Domain: English Verb Alternations

Causative/Inchoative Alternations We use English verbs in the *roll*-class (Levin, 1993), such as *roll*, *drop* and *slide*. These verbs participate in the causative/inchoative alternation: they ap-

Type	Example	P	S	R	Tests
Correct	*The mat rolled into a pillow*	×	×	×	All principles (multi-layered analogical reasoning)
RR	*The explorer rolled into a pillow*	×	×	✓	Semantic roles (minimal contextual cues)
SC-RR	*The explorer was into a pillow*	×	✓	✓	Syntactic structure (contrastive learning)
SCRS	*The mat rolled the explorer*	×	✓	✓	Argument structure (syntax + semantics)
PC-RR	*The man rolled in the cup*	✓	×	✓	Cross-paradigm consistency (analogical structure)
PSC-RR	*The man was in the cup*	✓	✓	✓	Multiple constraints (analogical + contrastive)
PSC-RS	*The dice rolled the man*	✓	✓	✓	Complex analogical mapping (all principles)

P=Paradigm, S=Structure, R=Role violations

Table 1: Systematic distractor taxonomy testing cognitive principles through hierarchical constraint violations. Each error type violates specific linguistic dimensions (✓) to isolate analogical reasoning components.

pear both in an intransitive frame (*The ball rolled*) and a transitive causative frame (*The player rolled the ball*). This alternation requires systematic Agent↔Theme mappings, where the same entity can take different grammatical roles depending on causation.

This linguistic phenomenon provides an ideal test case because it requires analogical reasoning across different lexical items sharing the same structural pattern. It also needs to discriminate between valid and invalid argument structure transformations, and integrate semantic role information with syntactic patterns.

Cross-Phenomenon Validation We validate the generalisability of the approach with *bake*-class verbs, a class of verbs exhibiting the unspecified object alternations (*The chef baked a cake* vs. *The chef baked*). In this class, the subject of both the transitive and the intransitive is an *Agent*. While the two classes share syntactic structure, they instantiate semantically distinct processes, testing whether our organisational principles generalise beyond specific semantic classes.

2.3 Analogical Paradigm Organisation

Figure 1 illustrates our systematic organisation approach. Each context follows a 2×4 paradigmatic structure.

Multi-Layered Analogical Structure Contexts encode analogical relationships at multiple levels: *Man:Dice :: Explorer:Mat* in terms of Agent-Theme relationships, with identical structural transformations applied across different lexical content. This tests whether models can abstract relational patterns beyond surface similarity.

Minimal Contextual Cues We provide semantic scaffolding through subtle annotations within each paradigm. For example, action descriptors indicate agentivity (*The player did it*) while state descriptions suggest thematic roles (*The ball was on the floor*). These cues clarify semantic relationships without explicit grammatical labelling.

Contrastive Learning Through Systematic Distractors Each answer set implements contrastive learning by providing systematic negative examples alongside each correct completion. Table 1 defines our hierarchical error taxonomy designed to test specific cognitive principles through constraint violations.

Each distractor violates exactly one constraint dimension while preserving others, enabling precise evaluation of rule component acquisition. Role errors test semantic understanding, structural errors test syntactic knowledge, and paradigm errors test analogical consistency across contexts.

Notice that distractors are contextually inappropriate rather than inherently ungrammatical. For example, *The explorer rolled into a pillow* is grammatical but violates the analogical mapping established in the context.

2.4 Data Generation

Our systematic generation process creates sentences based on templates with controlled lexical variation and seed alternation pairs, following established verb classifications (Levin, 1993).[1] We follow a hybrid, expert-guided process. Linguistic experts select and validate seed sentences and design a small set of templates with soft annotations for each alternation class. We then automatically instantiate full paradigms from these seeds and templates, controlling lexical variation and error types. This procedure creates a dataset of structured completion tasks over *roll*- and *bake*-class verbs, which we refer to as BLM-Roll-BakeE. [2]

[1] Verb class and seed sentences are detailed in Appendix A.

[2] Data is available at https://www.idiap.ch/en/scientific-research/data/blm-roll-bakee/

3 Experiments

We systematically evaluate our analogical paradigm organisation approach through controlled experiments testing three core claims: (i) our organisational principles enable sample efficiency compared to unstructured baselines; (ii) individual components contribute systematically to performance; (iii) lightweight models trained with structured data can match zero-shot and few-shot LLMs on the same controlled linguistic rule learning tasks, under a different learning regime. The performance on BLM puzzle completion is used as a proxy for a model's level of successful linguistic rule learning.

3.1 Experimental Design

Ablation Framework: We test five systematic conditions that isolate organisational components while controlling information content.

BASE is the complete paradigmatic organisation with analogical structure and contextual cues. SHUFFLED has identical content in random order, (tests organisational vs. content effects). NOANALOGY removes the first paradigm (tests analogical structure contribution). NOSOFTCUE removes contextual annotations (tests scaffolding contribution). TRANSPOSED preserves analogical patterns but changes spatial arrangement ($C_{\text{Trans}} = C_{\text{Base}}^{\top}$). [3]

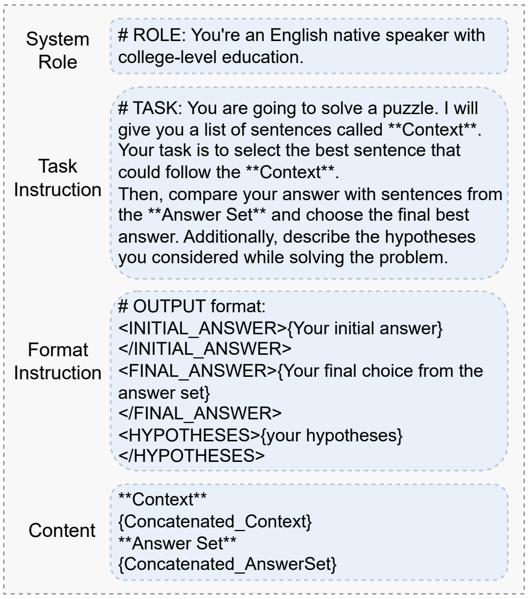
Data Configurations: *Type I* uses the same verbs across paradigms (3000 examples, 80:10:10 splits, direct analogical mapping). *Type II* uses different verbs across paradigms (15000 examples, same splits, tests abstraction). *Cross-phenomenon* uses unspecified object alternations with *bake*-class verbs (validates generalisability).

Scale Testing Training sizes range from 10 to 2700 examples for lightweight models; and 300-example test sets (same as for *Type I* test data, and 300 random examples from *Type II* test set) for LLMs. This setup balances computational feasibility with statistical reliability for cross-model comparisons.

3.2 Models and Training

Lightweight Models We use several models of representation. We embed each sentence using BERT[4](Devlin et al., 2019) as our primary model, RoBERTa[5] (Liu et al., 2019) and ELECTRA[6] (Clark et al., 2020) for encoder comparison. We reason that using a weaker base clarifies contributions attributable to data organisation rather than encoder capabilities.

As for architecture types, we use previously tested parameter-efficient neural models (CNN and FFNN in Nastase and Merlo (2023) for comparison). The CNN captures localised patterns and positional relationships, while the FFNN processes concatenated embeddings through fully-connected layers, integrating distributed information across the entire context. Both architectures maintain minimal parameter counts ($\sim 0.5M$ parameters) compared to the encoder ($\sim 110M$), isolating the influence of data organisation from architectural capacity. Both models output an answer embedding given the concatenated embeddings of the sentences from the context, evaluated against answer candidates using a max-margin objective with cosine similarity. Training runs for 120 epochs, with learning rate 0.001, Adam Optimizer, batch size 100, and early stopping (patience 10). All experiments use random seed 42 with results averaged over 3 independent runs.[7]



System Role

# ROLE: You're an English native speaker with college-level education.

Task Instruction

# TASK: You are going to solve a puzzle. I will give you a list of sentences called **Context**. Your task is to select the best sentence that could follow the **Context**.
Then, compare your answer with sentences from the **Answer Set** and choose the final best answer. Additionally, describe the hypotheses you considered while solving the problem.

Format Instruction

```
# OUTPUT format:
<INITIAL_ANSWER>{Your initial answer}
</INITIAL_ANSWER>
<FINAL_ANSWER>{Your final choice from the answer set}
</FINAL_ANSWER>
<HYPOTHESES>{your hypotheses}
</HYPOTHESES>
```

Content

```
**Context**
{Concatenated_Context}
**Answer Set**
{Concatenated_AnswerSet}
```

Figure 2: Zero-shot prompt.

Large Language Models We evaluate eight advanced models including reasoning-capable sys-

[3] See the formal definitions in Appendix B.1.

[4] bert-base-multilingual-cased

[5] xlm-roberta-base

[6] electra-base-discriminator

[7] Full encoder comparison and replication details are provided in Appendix C.

tems.

Reasoning models: `deepseek-R1` (DeepSeek-AI, 2025), `gpt-o3`, `gpt-o3-mini`,`qwq-32B`.

Standard models: `deepseek-V3` (DeepSeek-AI, 2024), `llama-3.3-70B-Instruct`, `llama-3.2-3B-Instruct`, `qwen3-32B`.

The $temperature$ is 0.1, where applicable (standard models), and the max_tokens is 2046. [8]

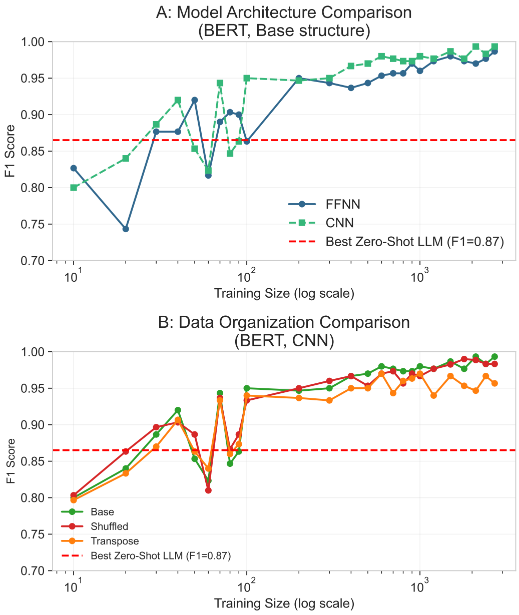


Figure 3: F1 Performance as a function of training size. (A) Comparison of model architectures with BERT embeddings on Base structure organisation. (B) Impact of data organisation using the best architecture.

Prompting and Evaluation We test zero, one and five-shot prompting with (w-CoT) and without (wo-CoT) chain-of-thought reasoning (Wei et al., 2021; Kojima et al., 2022). We use a puzzle-solving prompt that parallels our lightweight model setting. Figure 2 shows a zero-shot example. The LLM first sees the context, generates a provisional answer, then compares it to the answer set before generating the best final sentence (instead of answer key label), and it also generates the hypotheses considered. Any final choice not in the answer set is flagged as a system error (`ERR`). This prompt mimics our learning objective using pretrained models and presents a more challenging task than standard multiple-choice questions. Randomising the options in the answer set and asking to generate sentences instead of option labels help debias LLMs' token bias (e.g., A/B/C/D) in multiple-choice-format questions (Zheng et al., 2024). This uniform puzzle-solving format enables direct comparisons across model scales and training paradigms.

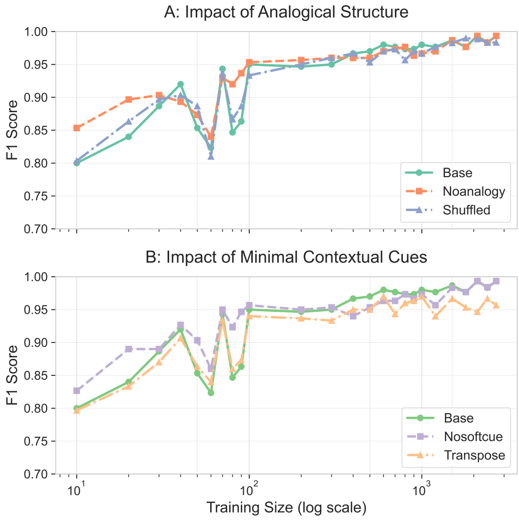


Figure 4: Isolated contributions of organisational components to best model performance. (A) Impact of analogical organisation, comparing BASE against NOANALOGY and SHUFFLED. (B) Impact of implicit soft annotations, comparing BASE with NOSOFTCUE and TRANSPOSED variants.

Evaluation F1 scores balance precision and recall across 7 answer choices. Responses not matching the answer set options are flagged as system errors for LLMs. A multi-factor ANOVA tests Structure × Model × Shots × CoT interactions ($p < 0.05$).

3.3 Results

We present results from our best-performing architecture (CNN) across all experimental conditions with lightweight models. While both CNN and FFNN demonstrate similar qualitative trends, CNN consistently outperforms FFNN across training regimes, likely due to its capacity to capture local sequential patterns critical for analogical mapping.

3.3.1 Sample Efficiency Validation

Figure 3 confirms our core finding: lightweight models with BASE organisation achieve $F1 = 0.95$ with only 100 examples, and, in this setting, are higher than our best zero-shot reasoning model `GPT-o3` ($F1 = 0.87$) on the same tasks. Performance stabilises at $1000 - 1200$ examples across architectures. Panel B shows the impact of data organisation, where BASE structure outperforms SHUFFLED and TRANSPOSED arrangements. Iden-

[8] Details on LLM configurations are in Appendix D.1.

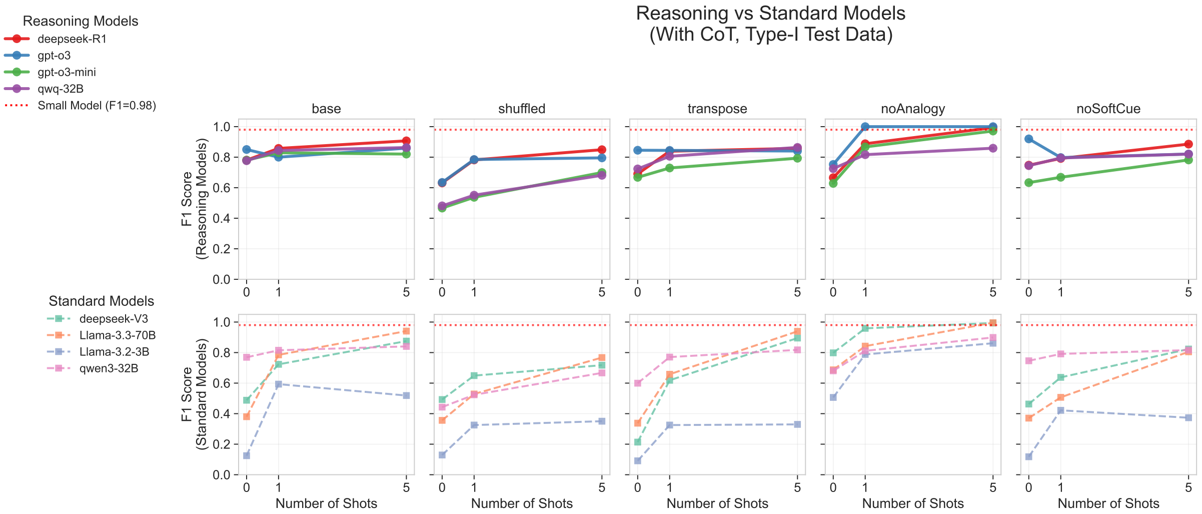


Figure 5: Reasoning models (top row: deepseek-R1, gpt-o3, gpt-o3-mini, qwq-32B) vs. standard models (bottom row: llama-3.3-70B-Instruct, llama-3.2-3B-Instruct, qwen3-32B, deepseek-V3) across data structures. Red dotted line shows small model baseline ($F1 = 0.98$). Reasoning models consistently outperform standard models, but struggle to match structured lightweight model performance in zero-shot settings.

tical content produces different learning trajectories, confirming that organisational structure, not information quantity, drives efficiency gains.

3.3.2 Component Analysis

Figure 4 validates the systematic contributions from individual organisational components. While all structures eventually converge to similar performance levels at larger training sizes, analogical structure (Base vs. NoAnalogy vs.Shuffled) shows slightly more consistent advantages in the $500 - 1500$ example range where data efficiency matters most. Contextual cues (Base vs. NoSoftCue vs. Transposed) provide additional but smaller benefits.

These subtle differences in learning curves suggest that while both components contribute to learning, modern neural architectures can compensate for deficiencies in either component as training sizes increases.

3.3.3 LLM Comparison

Figure 5 confirms our third contribution: lightweight models trained on 1000 structured examples reach higher $F1$ than all tested LLMs in zero-shot conditions, including state-of-the-art reasoning models, on these tasks and under our training regime. Multi-factor ANOVA reveals significant main effects for structure, model, and number of shots, with minimal CoT impact[9]. This result suggests that this task engages pattern recognition and rule application rather than multi-step reasoning.[10]

Reasoning vs. Standard Models Reasoning models start with high zero-shot performance ($F1 = 0.8+$) but plateau quickly, while standard models begin lower ($F1 = 0.1 - 0.5$) but show steeper improvement curves, making them more sensitive to organisational structure.

Organisational Structure Patterns Shuffled consistently performs worst across all models, confirming that random organisation disrupts learning. Base and NoAnalogy emerge as most effective, often performing comparably. Notably, NoAnalogy sometimes outperforms Base in zero-shot settings for reasoning models, suggesting analogical structure becomes less crucial when models already possess strong reasoning capabilities. However, standard models benefit most from organisational effects that have greater room for improvement.

3.4 Error Analysis and Patterns

Figure 6 shows organisational structure affects error patterns across different model types. Lightweight models are robust ($errors < 3.3\%$) across all organisational conditions, demonstrating consistent learning regardless of structural variations. LLMs show structure-dependent error pat-

[9] Complete results on Multi-factor ANOVA appear in Appendix D.2 (Table 3).

[10] Complete results on LLMs performance are in Appendix D.3 (Table 4).

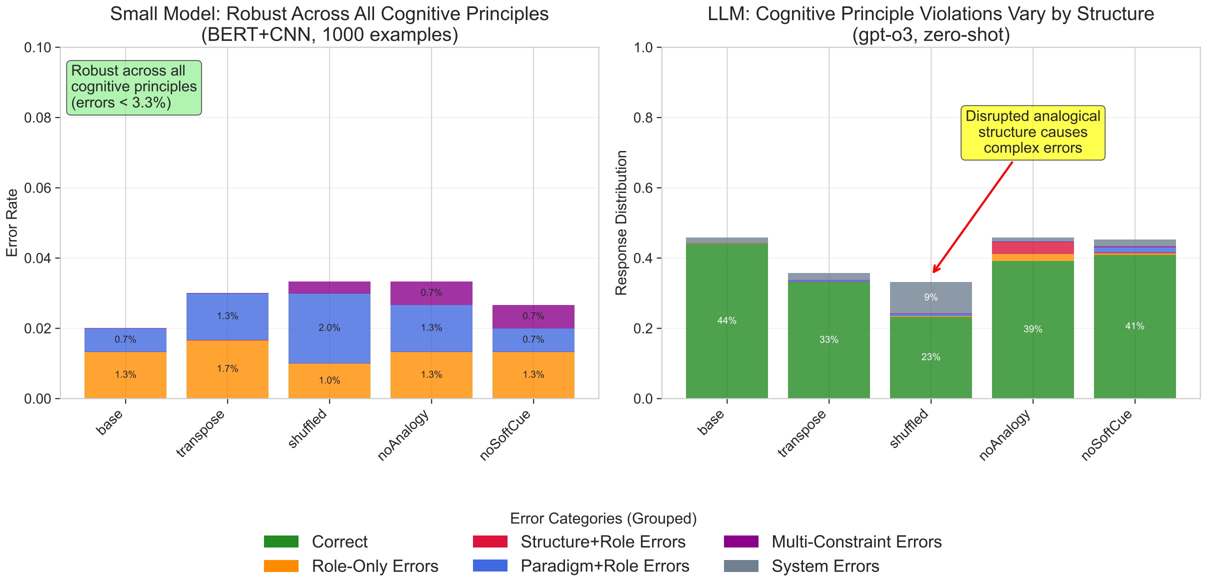


Figure 6: Error analysis across data organisations. Lightweight models (left) show robust performance with minimal errors ($< 3.3\%$) across all structures. LLMs (right, `gpt-o3` zero-shot) show structure-dependent error patterns following our cognitive-inspired error taxonomy (Table 1). SHUFFLED structures particularly disrupt LLMs' analogical reasoning, causing increased multi-constraint violations.

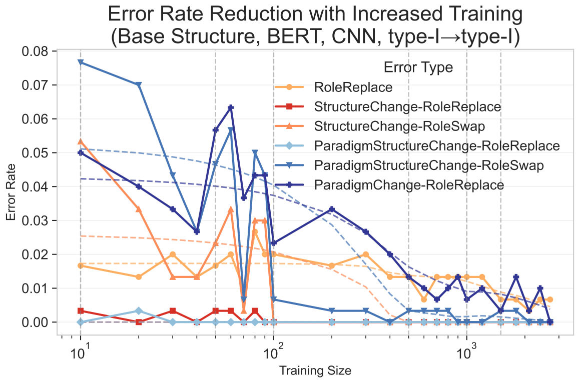


Figure 7: Reduction in error rates as training size increases for the BASE structure (train and test on *Type I* data). Dashed lines represent smoothed trends using LOWESS smoothing.

terns following our error taxonomy. Disrupted analogical structure (SHUFFLED condition) causes increased multi-constraint violations, with correct responses dropping from 55% (BASE) to 24% (SHUFFLED) for `gpt-o3`. This shows that even reasoning-capable models rely heavily on structural organisation for linguistic rule learning tasks.

Figure 7 tracks error reduction as training progresses with the BASE structure. ParadigmChange-RoleReplace errors show the most dramatic decrease, which suggests that paradigm-level information is learnt earlier than fine-grained role distinctions. The BASE structure facilitates more efficient error reduction across all types, which provides better implicit feedback about linguistic rules, consistent with error-driven learning theories.

3.5 Additional Analysis

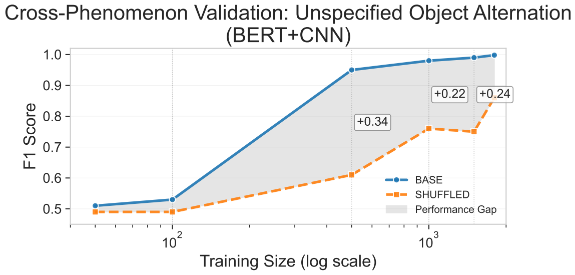


Figure 8: Cross-phenomenon validation using unspecified object alternations (*bake*-class verbs). Results show $F1$ performance using BERT+CNN architecture on *Type II* data, averaged over 3 runs.

Cross-Phenomenon Generalisation Figure 8 shows that BASE results are higher than SHUFFLED also with *bake*-class verbs. This suggests that our organisational benefits (BASE $>$ SHUFFLED) generalise beyond causative alternations and indicates that our approach captures general organisational principles rather than construction-specific optimisations.

Cross-Type Generalisation Figure 9 shows successful rule abstraction: models achieve robust

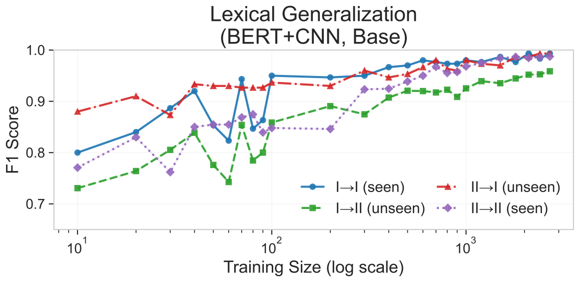


Figure 9: Seen type vs. Unseen Type performance across training sizes.

cross-type performance by $100 - 200$ examples, though all conditions eventually converge to high performance with sufficient training data. This confirms that structural relationships rather than surface patterns are being learnt.

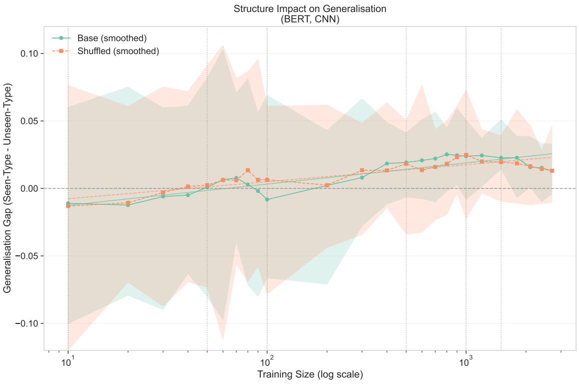


Figure 10: Generalisation gap between BASE and SHUFFLED structures. Smaller gaps indicate better generalisation. The BASE structure facilitates better transfer of linguistic rules across training sizes.

Figure 10 shows that the BASE maintains more stable and efficient generalisation throughout the learning process, while the SHUFFLED shows erratic performance in mid-range training sizes. At small training sizes, both structures show negative gaps, which suggests that the models avoid overfitting to surface lexical patterns.[11]

Discussion Our systematic evaluation validates all our three core claims. The finding that lightweight structured models ($F1 = 0.95$) exceed even gpt-o3 in zero-shot settings on our tasks ($F1 = 0.87$) demonstrates that systematic input organisation can achieve strong linguistic rule learning on controlled tasks with reduced task-specific data and computational requirements. The consistent organisational effects across both reasoning and standard LLMs indicate that analogical paradigm organisation addresses fundamental challenges in pattern recognition that persist even in the most advanced current systems. The contrasting error patterns between model types suggest that organisational benefits operate through different mechanisms, with accelerated learning for lightweight models and improved error discrimination for LLMs. The cross-phenomenon and cross-type evaluation confirms that our approach captures general principles rather than construction-specific optimisations.

These findings indicate that structured input data organisation is complementary to performance gains for linguistic rule learning beyond architectural advances or reasoning capabilities.

4 Related Work

Structured Data for Sample Efficiency Achieving competitive performance with reduced training needs remains a central challenge in NLP. Recent work emphasises how reorganising raw text into structured or hierarchically arranged contexts can significantly enhance model learning (Liu et al., 2024; Sourati et al., 2024). "Structured data" often refers to knowledge graphs, table data, databases (Jiang et al., 2023; Li et al., 2023). Liu et al. (2024) instead reorganise text into hierarchical formats (e.g., Scope-Aspect-Description) to mimic human knowledge consolidation, thereby improving task performance in question-answering. Existing sample efficiency approaches in NLP include data selection (Albalak et al., 2024), data augmentation (Kumar et al., 2020; Dai et al., 2023), curriculum learning (Bengio et al., 2009; Xu et al., 2020), among others. However, they typically focus on scaling or sequencing existing data rather than fundamental reorganisation principles. Our work instead targets linguistic rule learning through structure in a structured completion tasks inspired by cognitive principles.

Analogical Learning in Neural Language Models Analogical reasoning is the ability to perceive and use similarities on relational patterns rather than surface features (Gentner and Smith, 2012). Neural models have shown strong capabilities of analogical reasoning to generalise rules, such as word-level analogies (Mikolov et al., 2013; Brown et al., 2020; Ushio et al., 2021; Webb et al., 2023), concept analogies through analogical structure abduction (Yuan et al., 2023), narrative-based or story-

[11] Details for the generalisation gap metric and smoothing method appear in Appendix B.

level analogies (Nagarajah et al., 2022; Jiayang et al., 2023; Sourati et al., 2024), or analogical prompting where LLMs generate relevant exemplars or knowledge before solving the problem (Yasunaga et al., 2024). These approaches, like ours, suggest that explicit structural cues improve analogical processing.

In linguistics and cognitive science, work on argument structure and alternations has argued that children track families of related verbs and constructions when they learn causative alternations, rather than memorising isolated patterns. Bowerman (1974) studies how children acquire English causative verbs and the relation between cognitive, semantic, and syntactic development. Our approach is computationally inspired by this line of research, but it does not aim to model child learning directly. We instead study how analogical organisation of input, contrastive distractors, and minimal contextual cues affect sample efficiency and generalisation behaviour of neural models on controlled alternation tasks.

Contrastive Learning in NLP Contrastive learning has emerged as a powerful principle for neural representation learning. Systematic comparison between positive and negative examples enhances discriminative learning (Chen et al., 2020; He et al., 2020). However, existing contrastive approaches typically require large-scale training data and focus on representation learning rather than rule acquisition. Our systematic distractor design similarly implements a form of contrastive learning by organising examples that emphasise critical distinctions between alternations. This enables models to learn from the structured contrasts between valid and invalid linguistic transformations.

Linguistic Rule Learning in Neural Networks Neural approaches to linguistic rule learning are challenging despite recent advances. Neural language models demonstrate emergent syntactic capabilities (Linzen et al., 2016; Gulordava et al., 2018; Hewitt and Manning, 2019; Mueller et al., 2024) but require large-scale training. Kann et al. (2019) examine alternations in transformer embeddings with limited success. Thrush et al. (2020) find that systematic selectional preferences require careful training design. LLMs also struggle on metalinguistic tasks (Thrush et al., 2024). Merlo (2023) introduces Blackbird Language Matrices (BLM) as structured evaluation paradigms for systematic rule assessment, enabling controlled testing of linguistic competence. However, they show limited efficiency in data-lean settings. Our work addresses these efficiency challenges through analogical paradigm organisation, building on the BLM framework to achieve competitive argument structure learning with minimal training examples while demonstrating robustness across alternation types.

5 Conclusion

We investigated whether computational approaches inspired by cognitive principles can enable sample-efficient linguistic rule learning. Our analogical paradigm organisation approach operationalises three organisational strategies, analogical structure, contrastive learning, and minimal contextual cues, through systematic input structuring rather than architectural scaling.

Our results demonstrate substantial sample efficiency gains across multiple evaluation dimensions. Lightweight models (BERT+CNN, $\sim 0.5M$ parameters) trained on only 100 structured examples achieved $F1 = 0.95$, and, in this setting, reached higher $F1$ than zero-shot `GPT-o3` ($F1 = 0.87$) on the same structured tasks, and stabilising performance with just $1000-1200$ training instances. Component analysis confirmed that each organisational principle contributes systematically, with analogical structure showing the largest effect. Cross-phenomenon validation with *bake*-class verbs and cross-type generalisation experiments confirmed that benefits extend beyond construction-specific optimisations.

These findings have practical and theoretical implications. Practically, our approach enables competitive linguistic rule learning with much fewer task-specific examples than conventional LLM-based setups in this controlled setting. Theoretically, we demonstrate that systematic input organisation can achieve efficiency gains that complement rather than require architectural scaling. This suggests that cognitive-inspired data structuring is a distinct optimisation dimension for linguistic rule induction.

Future work should explore cross-linguistic validation, integration with pretraining objectives, and automated structure discovery methods. Overall, our study shows that computational approaches inspired by cognitive principles can support sample-efficient linguistic rule learning and offer a complementary perspective on the behaviour of lightweight models and prompted LLMs.

6 Limitations

While our approach demonstrates efficiency gains, it is not without limitations.

Expert Dependency Our structured paradigms require expert linguistic knowledge to design appropriate analogical mappings and systematic distractors. In practice, we follow a hybrid process where experts select and validate seed sentences and design a small set of templates, and full paradigms are then generated automatically. Although we demonstrate cross-phenomenon generalisability with *bake*-class verbs, scaling this approach to diverse linguistic phenomena would require either linguistic expertise or automated methods for discovering optimal organisational structures.

Pre-training Attribution Our lightweight models rely on pre-trained encoders (BERT, RoBERTa, ELECTRA) that already encode linguistic knowledge from unstructured data. This creates attribution challenges in isolating whether performance gains stem purely from our organisational principles or from interactions with pre-existing linguistic representations. While our BASE vs. SHUFFLED comparisons control for this confound, the fundamental attribution question remains.

Language and Phenomenon Scope Our evaluation was exclusively on English verb alternations, a specific subset of linguistic phenomena at the syntax-semantics interface. Cross-linguistic validation and extension to other grammatical constructions (e.g., morphological patterns, syntactic transformations) are needed to establish broader applicability of our organisational principles.

Computational Comparison Framework Our comparison between trained lightweight models and zero-shot and few-shot LLMs places the models in different learning regimes. The lightweight models are fine-tuned on $100 - 1000$ task-specific examples. The LLMs are evaluated in zero-shot and few-shot prompt-based settings without parameter updates. These settings are not matched in computational cost or data exposure. Our claims about lightweight models and LLMs are therefore restricted to this contrast and do not extend to general conclusions about LLM capabilities under full fine-tuning or matched training budgets.

Task Specificity Our approach targets pattern completion tasks that may favour analogical reasoning. This controlled design allows us to isolate the effects of input organisation on linguistic rule induction, but it does not cover broad downstream applications. Extension to diverse NLP applications (e.g., generation, open-ended reasoning) would provide stronger evidence for the general usability of our organisational principles.

Ethics

Our research presents minimal ethical concerns. Our synthetically generated data focuses on grammatical verb alternations rather than semantic content, minimising bias propagation risks from LLM training data.

Acknowledgements

We thank the anonymous reviewers of ACL Rolling Review and EACL for their constructive feedback. We also thank colleagues, Vivi Nastase and Giuseppe Samo at Idiap Research Institute for valuable discussions. We gratefully acknowledge the partial support of this work by the Swiss National Science Foundation, through grant TMAG-1_209426 to PM.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*. Survey Certification.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Melissa Bowerman. 1974. Learning the structure of causative verbs: A study in the relationship of cognitive, semantic, and syntactic development. *Papers and reports on child language development*, 8:142–178.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Morten H Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, 39:e62.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth D Forbus, Dedre Gentner, and Keith Law. 1995. Mac/fac: A model of similarity-based retrieval. *Cognitive science*, 19(2):141–205.

Dedre Gentner and L. Smith. 2012. *Analogical Reasoning*, pages 130–136. Elsevier Inc., United States.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537, Singapore. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Beth Levin. 1993. *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Haoyang Li, Jing Zhang, Cuiping Li, and Hong Chen. 2023. Resdsql: Decoupling schema linking and skeleton parsing for text-to-sql. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13067–13075.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kai Liu, Zhihang Fu, Chao Chen, Wei Zhang, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, and Jieping Ye. 2024. Enhancing LLM's cognition via structurization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Paola Merlo. 2023. Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Can large language models pass the test? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8119–8152, Singapore. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. 2024. In-context learning generalizes, but not always robustly: The case of syntax. 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 4761–4779, Mexico City, Mexico. Association for Computational Linguistics.

Thiloshon Nagarajah, Filip Ilievski, and Jay Pujara. 2022. Understanding narratives through dimensions of analogy.

Vivi Nastase and Paola Merlo. 2023. Grammatical information in BERT sentence embeddings as two-dimensional arrays. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 22–39, Toronto, Canada. Association for Computational Linguistics.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. ARN: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.

Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, and Douwe Kiela. 2024. I am a strange dataset: Metalinguistic tests for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8888–8907, Bangkok, Thailand. Association for Computational Linguistics.

Tristan Thrush, Ethan Wilcox, and Roger Levy. 2020. Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 265–275, Online. Association for Computational Linguistics.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. 2023. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2446–2460, Singapore. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Linguistic Phenomena

A.1 Causative/Inchoative Alternation (*roll*-class verbs)

The class of *roll* verbs, as categorised by Levin (1993), comprises verbs expressing dynamic actions with inherent motion characteristics. This class shows systematic alternation between:

- **Intransitive**: Theme as subject (e.g., "*The ball rolled*")

- **Transitive**: Agent causes Theme motion (e.g.,"*The player rolled the ball*")

This syntactic-semantic interface represents a complex mapping challenge for computational models.

Verb Inventory The verbs within this class include: *bounce*, *coil*, *drift*, *drop*, *float*, *glide*, *move*, *revolve*, *roll*, *rotate*, *slide*, *spin*, *swing*, *turn*, *twirl*, *twist*, *whirl*, and *wind*.

A.2 Cross-Phenomenon Validation (*bake*-class verbs)

Unspecified object alternation with *bake*-class verbs provides syntactically similar but semantically distinct validation:

- **Transitive**: "*The chef baked a cake*"
- **Intransitive**:"*The chef baked*" (object understood)

Verb Inventory The verbs within this class include: *bake*, *carve*, *chop*, *clean*, *cook*, *crochet*, *draw*, *drink*, *dust*, *eat*, *embroider*, *hum*, *hunt*, *fish*, *iron*, *knead*, *knit*, *mend*, *milk*, *mow*, *nurse*, *pack*, *paint*, *play*, *plow*, *polish*, *read*, *recite*, *sew*, *sculpt*, *sing*, *sketch*, *sow*, *study*, *sweep*, *teach*, *type*, *vacuum*, *wash*, *weave*, *whittle* and *write*.

This tests whether organisational benefits generalise across alternation types rather than being specific to a single alternation type.

B Experimental Setup Details

B.1 Ablation Design Formalisation

We systematically manipulate context structures to isolate organisational components. Given a full context matrix $C_{\text{full}} \in \mathbb{R}^{m \times n}$ and transformation matrix $T \in \{0,1\}^{m \times n}$, transformed contexts are computed as:

$$C_{\text{trans}} = C_{\text{full}} \odot T$$

Transformation Matrices

- NoAnalogy Context

$$T_{\text{af}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

- NoSoftCue Context

$$T_{\text{scf}} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

- For Transposed Context, it simply involves taking the transpose of the full context matrix:

$$C_{\text{trans}} = C_{\text{full}}^{T}$$

B.2 Statistical Analysis Framework

Multi-factor ANOVA Design We use a five-way ANOVA with factors: Structure (5 levels), Model (8 levels), Shots (3 levels), CoT (2 levels), and Data Type (2 levels).

Generalisation Gap Metric We define the generalisation gap as

$$\text{GenGap}(t,s) = \frac{1}{|M|} \sum_{m \in M} \Big[F1_{m,s}^{\text{seen}}(t) - F1_{m,s}^{\text{unseen}}(t)\Big] \tag{1}$$

where t is the training size, s the structure type, M the model set, and smaller gaps indicate better generalisation. Smoothing uses a centred rolling window ($size = 3$) with confidence bands representing standard error propagation.

C Model Implementation

C.1 Encoder Comparison

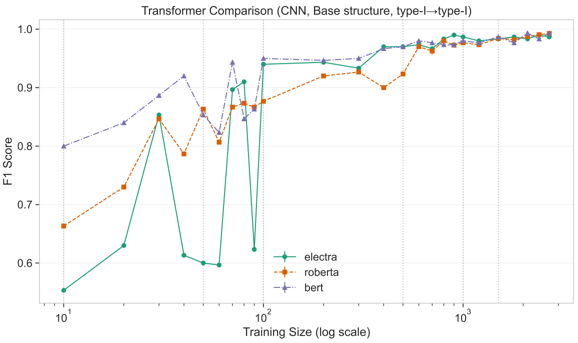


Figure 11: F1 performance as a function of training size, comparing encoder architectures (BERT, RoBERTa, ELECTRA) with the CNN architecture on Base structure. Train and test on *Type I* data.

We employ BERT (`bert-base-multilingual-cased`) as our primary encoder to emphasise contributions attributable to data organisation rather than encoder capabilities, as evidenced by Figure 11.

C.2 Architecture Specifications

Convolutional Neural Network (CNN)

- Input: array of embeddings, size 7 × 768

- Convolution layers: three 2D convolutional layers
  - kernel size: 3×3, Stride: 1, No dilation
- Output: fully connected layer compressing to size 768
- Function: localises sequential patterns for analogical mapping

Feed-Forward Neural Network (FFNN)

- Input: concatenated sentence embeddings, size 7×768
- Architecture: three fully connected layers
- Compression: $7 \times 768 \rightarrow 3.5 \times 768 \rightarrow 3.5 \times 768 \rightarrow 768$
- Function: distributed pattern integration across the entire context

C.3 Training Configuration

Models are trained over 120 epochs with learning rate 0.001 using the Adam optimiser and batch size 100. Early stopping with patience 10 prevents overfitting. All models employ a max-margin objective using cosine similarity:

$$\mathcal{L} = \sum_{\mathbf{e}_i \in \mathcal{A}} \max\Big(0,\ 1 + \cos(\mathbf{e}_i, \mathbf{e}_{\text{pred}}) - \cos(\mathbf{e}_c, \mathbf{e}_{\text{pred}})\Big) \tag{2}$$

where $\mathbf{e}_c$ is the correct answer embedding, $\mathbf{e}_i$ are incorrect distractor embeddings, and $\mathbf{e}_{\text{pred}}$ is the predicted completion. Models are run across three independent trials with seed 42 to ensure statistical reliability.

C.4 Computational Environment

Hardware Configuration

- **Workstation:** HP PAIR Workstation Z4 G4 MIT
- **Processor:** Intel Xeon W-2255
- **RAM:** 64 GB
- **GPU:** MSI GeForce RTX 3090 VENTUS 3X OC, 24 GB GDDR6X

Software Environment

- Ubuntu: 22.04, Python: 3.11.5, CUDA: 11.8.0
- PyTorch: 2.2.2, Transformers: 4.39.1, scikit-learn: 1.5.0
- NumPy: 1.25.0, pandas: 1.4.3

D Complementary Experimental Results

D.1 Large Language Model Specifications

We evaluate eight advanced models across four families with varying parameter scales. This systematic comparison allows us to assess how organisational structure affects different model capacities and architectural approaches.

D.2 Statistical Analysis Results

Complete multi-factor ANOVA results demonstrating significant main effects for structure, model, and number of shots across all experimental conditions.

D.3 Comprehensive Performance Results

Complete results across all models, structures, data types, and number of shot configurations with and without Chain-of-Thought reasoning.

Model Family	**Variant**	**Parameters**	**Context**	**Type**	**Release**
		Reasoning Models			
DeepSeek	deepseek-R1-0528	670B	64K	Reasoning	2025-05
OpenAI	gpt-o3	Undisclosed	128K	Reasoning	2025-04
OpenAI	gpt-o3-mini	Undisclosed	128K	Reasoning	2025-01
Alibaba	qwq-32B	32B	32K	Reasoning	2025-03
		Standard Models			
DeepSeek	deepseek-V3-0324	671B	64K	Standard	2025-03
Meta	llama-3.3-70B-Instruct	70B	128K	Standard	2024-12
Meta	llama-3.2-3B-Instruct	3B	128K	Standard	2024-09
Alibaba	qwen3-32B	32B	128K	Standard	2025-04

Table 2: Large language models evaluated in experiments. Context window indicates maximum input sequence length in tokens.

Factors	F	p
C(Structure)	9284.541	$p < 0.05$
C(Model)	9581.462	$p < 0.05$
C(Num_Shots)	14125.142	$p < 0.05$
C(CoT)	9.562	$p < 0.05$
C(Data_Type)	191.786	$p < 0.05$
C(Structure):C(Model)	313.487	$p < 0.05$
C(Structure):C(Num_Shots)	170.116	$p < 0.05$
C(Model):C(Num_Shots)	664.883	$p < 0.05$
C(Structure):C(CoT)	2.125	0.0601
C(Model):C(CoT)	3.625	$p < 0.05$
C(Num_Shots):C(CoT)	5.399	$p < 0.05$
C(Structure):C(Data_Type)	115.123	$p < 0.05$
C(Model):C(Data_Type)	37.116	$p < 0.05$
C(Num_Shots):C(Data_Type)	34.928	$p < 0.05$
C(CoT):C(Data_Type)	0.078	0.7795
C(Structure):C(Model):C(Num_Shots)	55.286	$p < 0.05$
C(Structure):C(Model):C(CoT)	1.644	$p < 0.05$
C(Structure):C(Num_Shots):C(CoT)	2.265	$p < 0.05$
C(Model):C(Num_Shots):C(CoT)	1.889	$p < 0.05$
C(Structure):C(Model):C(Data_Type)	13.123	$p < 0.05$
C(Structure):C(Num_Shots):C(Data_Type)	17.556	$p < 0.05$
C(Model):C(Num_Shots):C(Data_Type)	11.177	$p < 0.05$
C(Structure):C(CoT):C(Data_Type)	0.368	0.8709
C(Model):C(CoT):C(Data_Type)	0.214	0.9824
C(Num_Shots):C(CoT):C(Data_Type)	6.551	$p < 0.05$
C(Structure):C(Model):C(Num_Shots):C(CoT)	0.899	0.7079
C(Structure):C(Model):C(Num_Shots):C(Data_Type)	7.100	$p < 0.05$
C(Structure):C(Model):C(CoT):C(Data_Type)	0.919	0.6045
C(Structure):C(Num_Shots):C(CoT):C(Data_Type)	1.748	0.0657
C(Model):C(Num_Shots):C(CoT):C(Data_Type)	1.221	0.2531
C(Structure):C(Model):C(Num_Shots):C(CoT):C(Da...	0.944	0.6086

Table 3: Multi-ANOVA tests on data (Structure, Data_Type) and experimental (Model, Num_Shots, CoT) factors.

Model Config	Base	Noanalogy	Nosoftcue	Shuffled	Transpose
deepseek-R1 (0-shot, w-CoT)	0.743 ± 0.048	0.692 ± 0.037	0.711 ± 0.049	0.614 ± 0.043	0.655 ± 0.050
deepseek-R1 (0-shot, wo-CoT)	0.719 ± 0.063	0.707 ± 0.018	0.720 ± 0.020	0.641 ± 0.030	0.681 ± 0.035
deepseek-R1 (1-shot, w-CoT)	0.857 ± 0.004	0.876 ± 0.048	0.802 ± 0.017	0.788 ± 0.012	0.851 ± 0.028
deepseek-R1 (1-shot, wo-CoT)	0.848 ± 0.009	0.884 ± 0.054	0.802 ± 0.005	0.765 ± 0.021	0.873 ± 0.071
deepseek-R1 (5-shot, w-CoT)	0.884 ± 0.052	0.994 ± 0.004	0.875 ± 0.072	0.871 ± 0.073	0.927 ± 0.076
deepseek-R1 (5-shot, wo-CoT)	0.883 ± 0.058	0.911 ± 0.065	0.866 ± 0.069	0.871 ± 0.073	0.859 ± 0.007
deepseek-V3 (0-shot, w-CoT)	0.537 ± 0.193	0.859 ± 0.089	0.397 ± 0.163	0.445 ± 0.107	0.196 ± 0.130
deepseek-V3 (0-shot, wo-CoT)	0.487 ± 0.409	0.912 ± 0.102	0.630 ± 0.320	0.483 ± 0.410	0.139 ± 0.177
deepseek-V3 (1-shot, w-CoT)	0.729 ± 0.015	0.961 ± 0.008	0.634 ± 0.019	0.620 ± 0.056	0.544 ± 0.082
deepseek-V3 (1-shot, wo-CoT)	0.725 ± 0.011	0.963 ± 0.008	0.632 ± 0.019	0.604 ± 0.033	0.509 ± 0.092
deepseek-V3 (5-shot, w-CoT)	0.895 ± 0.063	0.994 ± 0.002	0.853 ± 0.058	0.703 ± 0.021	0.828 ± 0.100
deepseek-V3 (5-shot, wo-CoT)	0.963 ± 0.007	0.995 ± 0.003	0.899 ± 0.013	0.716 ± 0.036	0.867 ± 0.076
gpt-o3 (0-shot, w-CoT)	0.852 ± 0.008	0.754 ± 0.031	0.844 ± 0.053	0.688 ± 0.041	0.838 ± 0.012
gpt-o3 (0-shot, wo-CoT)	0.884 ± 0.057	0.754 ± 0.022	0.840 ± 0.064	0.700 ± 0.030	0.834 ± 0.015
gpt-o3 (1-shot, w-CoT)	0.891 ± 0.087	0.995 ± 0.012	0.858 ± 0.076	0.783 ± 0.053	0.869 ± 0.066
gpt-o3 (1-shot, wo-CoT)	0.911 ± 0.082	0.990 ± 0.008	0.834 ± 0.031	0.785 ± 0.046	0.894 ± 0.079
gpt-o3 (5-shot, w-CoT)	0.939 ± 0.072	0.998 ± 0.004	0.884 ± 0.094	0.858 ± 0.075	0.922 ± 0.086
gpt-o3 (5-shot, wo-CoT)	0.925 ± 0.078	1.000 ± 0.000	0.842 ± 0.019	0.830 ± 0.033	0.938 ± 0.085
gpt-o3-mini (0-shot, w-CoT)	0.754 ± 0.026	0.629 ± 0.028	0.606 ± 0.044	0.431 ± 0.041	0.641 ± 0.034
gpt-o3-mini (0-shot, wo-CoT)	0.745 ± 0.047	0.644 ± 0.070	0.636 ± 0.052	0.410 ± 0.014	0.573 ± 0.101
gpt-o3-mini (1-shot, w-CoT)	0.809 ± 0.015	0.863 ± 0.055	0.673 ± 0.024	0.542 ± 0.060	0.701 ± 0.063
gpt-o3-mini (1-shot, wo-CoT)	0.812 ± 0.023	0.866 ± 0.059	0.647 ± 0.068	0.551 ± 0.044	0.730 ± 0.068
gpt-o3-mini (5-shot, w-CoT)	0.894 ± 0.076	0.989 ± 0.013	0.774 ± 0.042	0.638 ± 0.044	0.767 ± 0.053
gpt-o3-mini (5-shot, wo-CoT)	0.881 ± 0.086	0.995 ± 0.005	0.764 ± 0.045	0.622 ± 0.053	0.786 ± 0.047
llama-3.2-3B-Instruct (0-shot, w-CoT)	0.093 ± 0.035	0.478 ± 0.033	0.109 ± 0.014	0.118 ± 0.015	0.061 ± 0.033
llama-3.2-3B-Instruct (0-shot, wo-CoT)	0.105 ± 0.042	0.501 ± 0.034	0.112 ± 0.018	0.113 ± 0.015	0.066 ± 0.037
llama-3.2-3B-Instruct (1-shot, w-CoT)	0.524 ± 0.079	0.761 ± 0.031	0.388 ± 0.044	0.279 ± 0.053	0.283 ± 0.049
llama-3.2-3B-Instruct (1-shot, wo-CoT)	0.532 ± 0.055	0.757 ± 0.032	0.374 ± 0.039	0.271 ± 0.057	0.269 ± 0.036
llama-3.2-3B-Instruct (5-shot, w-CoT)	0.456 ± 0.071	0.882 ± 0.045	0.330 ± 0.051	0.317 ± 0.043	0.308 ± 0.031
llama-3.2-3B-Instruct (5-shot, wo-CoT)	0.491 ± 0.063	0.890 ± 0.038	0.313 ± 0.041	0.340 ± 0.051	0.319 ± 0.026
llama-3.3-70B-Instruct (0-shot, w-CoT)	0.304 ± 0.083	0.700 ± 0.015	0.340 ± 0.034	0.337 ± 0.026	0.238 ± 0.110
llama-3.3-70B-Instruct (0-shot, wo-CoT)	0.298 ± 0.087	0.687 ± 0.017	0.309 ± 0.021	0.318 ± 0.018	0.223 ± 0.106
llama-3.3-70B-Instruct (1-shot, w-CoT)	0.774 ± 0.021	0.845 ± 0.046	0.515 ± 0.036	0.491 ± 0.042	0.592 ± 0.073
llama-3.3-70B-Instruct (1-shot, wo-CoT)	0.783 ± 0.047	0.783 ± 0.080	0.499 ± 0.015	0.447 ± 0.046	0.594 ± 0.066
llama-3.3-70B-Instruct (5-shot, w-CoT)	0.964 ± 0.055	0.996 ± 0.001	0.875 ± 0.084	0.737 ± 0.035	0.952 ± 0.045
llama-3.3-70B-Instruct (5-shot, wo-CoT)	0.942 ± 0.071	0.976 ± 0.051	0.870 ± 0.091	0.734 ± 0.021	0.931 ± 0.056
qwen3-32B (0-shot, w-CoT)	0.786 ± 0.122	0.596 ± 0.222	0.736 ± 0.029	0.441 ± 0.009	0.538 ± 0.074
qwen3-32B (0-shot, wo-CoT)	0.659 ± 0.088	0.662 ± 0.059	0.752 ± 0.069	0.429 ± 0.016	0.535 ± 0.065
qwen3-32B (1-shot, w-CoT)	0.808 ± 0.011	0.797 ± 0.018	0.792 ± 0.018	0.532 ± 0.014	0.729 ± 0.046
qwen3-32B (1-shot, wo-CoT)	0.811 ± 0.021	0.788 ± 0.009	0.800 ± 0.012	0.525 ± 0.028	0.740 ± 0.046
qwen3-32B (5-shot, w-CoT)	0.839 ± 0.003	0.896 ± 0.070	0.830 ± 0.017	0.636 ± 0.049	0.815 ± 0.007
qwen3-32B (5-shot, wo-CoT)	0.833 ± 0.009	0.917 ± 0.067	0.842 ± 0.063	0.619 ± 0.021	0.808 ± 0.010
qwq-32B (0-shot, w-CoT)	0.776 ± 0.035	0.724 ± 0.023	0.747 ± 0.026	0.504 ± 0.018	0.672 ± 0.076
qwq-32B (0-shot, wo-CoT)	0.752 ± 0.060	0.732 ± 0.011	0.731 ± 0.036	0.462 ± 0.042	0.652 ± 0.076
qwq-32B (1-shot, w-CoT)	0.843 ± 0.010	0.821 ± 0.015	0.793 ± 0.007	0.598 ± 0.042	0.762 ± 0.056
qwq-32B (1-shot, wo-CoT)	0.843 ± 0.009	0.822 ± 0.015	0.792 ± 0.011	0.600 ± 0.047	0.755 ± 0.048
qwq-32B (5-shot, w-CoT)	0.863 ± 0.006	0.863 ± 0.004	0.831 ± 0.017	0.691 ± 0.021	0.846 ± 0.021
qwq-32B (5-shot, wo-CoT)	0.864 ± 0.002	0.863 ± 0.003	0.829 ± 0.009	0.680 ± 0.017	0.836 ± 0.027

Table 4: Performance (F1 scores in % ± standard deviation) of all LLMs validated across all configurations.