Boosting Cognitive Modelling for Human Reasoning

Meghna Bhadra (meghna.bhadra@tu-dresden.de)
Computational Logic Group, Technische Universität Dresden, Germany

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

AI models are often developed to solve reasoning problems optimally. In contrast, cognitive models focus on explaining and predicting replicative cognitive patterns of human information processing. And while many of the theories aim to explain an assumed 'general' human reasoner, only few are aimed at the individual. This paper addresses the challenge of the latter by investigating the automatic generation of individualised predictive algorithms using transformer-based models. These models which have been trained on huge amounts of human data, potentially exhibit built-in cognitive patterns. Leveraging such characteristics and architecture of transformer-based models, we outline a generalized methodology for establishing a human-AI collaborative framework, to generate explainable and reproducible algorithms with cross-domain applicability. While predictive accuracy and generalizability pose less of a problem, the bigger challenges in using machine learning approaches or transformer-based models may be explainability and replicability. Hence, instead of 'just' using such a model for directly fitting the data, we use it to extract features and to propose cognitive algorithms that are executable in systems outside of the model. Using two datasets pertaining to syllogistic and spatial reasoning, the predictive algorithms thus generated applying the presented framework, achieve mean accuracies of 68% and 81%, respectively. Both algorithms outperform other established, state-of-the-art cognitive models by far, surpassing the (previously) best state-of-the art models in syllogistic and spatial human reasoning by 19% and 13%, respectively.

Keywords: Cognitive Modelling; Artificial Intelligence; Syllogistic Reasoning; Spatial Reasoning; Predictive Models

Introduction

Cognitive science is built on the idea of developing algorithms that can predict and explain human behaviour. And human behaviour systematically deviates from normative accounts such as formal logic or probability theory. Consider the following *premises: Some A are B, some B are C*. The conclusion *some A are C* may seem intuitive and is indeed frequently selected by participants in experiments (Khemlani & Johnson-Laird, 2012). However, it is logically invalid because counterexamples are possible and consequently, an algorithm built on the principles of logic would *not* derive a deviating human conclusion.

In the domain of syllogistic reasoning alone more than eleven theories have been developed (for an overview see (Khemlani & Johnson-Laird, 2012)). One prominent heuristic theory is the *Atmosphere Theory* (AT) (Woodworth & Sells, 1935). It proposes that reasoning with syllogisms is dependent on an "atmosphere" created by the premises. For e.g., it proposes that premises having universal affirmative qualifiers, such as *"All A are B, All B are C"*, create a predisposition towards an All-qualified conclusion, and other such heuristics. The *Probability Heuristics Model* (PHM) assumes that the informativity of the qualifier plays a role. The *Mental Models Theory* (MMT) and its implementation the *mReasoner*, assumes that human reasoners construct analogous "model-like" semantic representations. Which conclusions follow from the reasoning episode may be determined by inspecting these models. And errors in reasoning can occur when individuals fail to consider all possible models. The model mReasoner also contains additional parameters that control model size, stochasticity etc. Aside this, there are also works such as (Tessler, Tenenbaum, & Goodman, 2022) which argue that human syllogistic reasoning is best understood as some rational probabilistic inference, influenced by the pragmatics of the context.

Generally, cognitive models for human reasoning have been developed using a "one-size-fits-all" approach, i.e., a single "static" model that does not account for individual differences. While some approaches have aimed to identify group differences, e.g., (Khemlani & Johnson-Laird, 2016), even fewer have analysed individual behaviour, e.g., (Riesterer, Brand, & Ragni, 2020). Although these models achieve approximately 80% accuracy on average (Khemlani & Johnson-Laird, 2012), their performance drops to less than 40% for individual cases as discussed in Riesterer et al. (2020).

The overall aim, however, has always been to understand and model the individual reasoner. In cognitive science, this has sometimes been approached as a parameter problem (e.g., like in mReasoner), but what makes modelling human reasoning challenging is that there can be a diverse range of cognitive representations and reasoning approaches, such as heuristic or model-based approaches (Khemlani & Johnson-Laird, 2012). Another problem is that implementable algorithms might not capture individual differences comprehensively – a need that was identified early on (Stenning & Cox, 2006; Fugard & Stenning, 2013). Hence, what is needed is a methodology to tailor cognitive models to individual users and explain their underlying reasoning patterns.

With the advent of large language models (LLMs) (Vaswani et al., 2017) comparison of human reasoning with

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

that simulated by LLMs has gained interest (Eisape et al., 2024). Generally LLMs aim to generate contextually appropriate text that may be indistinguishable from human generated text, serving as a kind of a Turing test (Johnson-Laird & Ragni, 2023). But LLMS have demonstrated further capabilities, for e.g., PaLM 2 seems to surpass human performance in logical inference tasks (Eisape et al., 2024), Llama 3 has demonstrated strong performance in common-sense reasoning tasks (Krause & Stolzenburg, 2024) – a domain in the forte of humans. LLMs also apparently exhibit inherent cognitive biases (Eisape et al., 2024; Yax, Anlló, & Palminteri, 2024) – which indicates implicitly "built-in" cognitive patterns – although this also necessitates better benchmarks (Wu et al., 2023). Given such motivations the question that we now pose is the following: can transformer-based models be effectively used to predict future reasoning patterns of an individual based on their past data? We propose that it is indeed possible to have human-AI collaborative frameworks with transformer-based models to generate predictive *cognitive algorithms* adapted to suit individualized reasoning processes. This is based on the notion that prompting provides the scope to generate, test and refine the predictive quality of the resulting model, in a transparent and explainable manner.

In the following, we outline a general protocol for establishing a human-AI collaborative framework to generate predictive algorithms using transformer-based models. We then report our results, applying this protocol to develop algorithms that predict individual reasoning in syllogistic and spatial tasks – two well-studied domains in cognitive science. Finally, we present details of the resulting predictive model and compare it to state-of-the-art cognitive models.

Properties of good cognitive models, Testing Framework, and Benchmarks

Many experimental findings demonstrate a variety of reasoning patterns, i.e., rational "logical" processes, biases (e.g., the belief bias), and heuristics. This shows that a general cognitive model fitting the average reasoner may often fail to capture the variety of individual reasoning strategies. And it implies a fundamental challenge and our research question: how can we accurately predict an *individual's* reasoning process based on their past inferences? Addressing this challenge requires a methodology capable of automatically generating personalized cognitive models. However, building such models entails prescribing clear criteria for what constitutes a "good" cognitive model. Among the many, we consider the following to be important:

1. **Predictive accuracy** when predicting an individual's response pattern, and the flexibility to adapt to new data or potential changes in reasoning patterns over time.

2. **Explainability** of the underlying processes.

3. **Replicability and consistency** in predictions when applied to the same data.

4. **Generalizability** of the methods across domains.

These features can create challenges to using machine learning approaches and especially LLMs. While predictive accuracy and generalizability pose less of a problem, explainability and replicability are greater challenges. Hence, instead of 'just' using an LLM directly for fitting the data, we use them to extract features and to automatically generate cognitive algorithms that can be executed outside of the LLMs. To address the current goals, we use datasets from syllogistic and spatial reasoning.

Dataset Description - Syllogistic Reasoning We used the data set introduced by (Ragni, Dames, Brand, & Riesterer, 2019),[1] that was collected through an Amazon Mechanical Turk web experiment in 2016 and has also been used by several other researchers (Wu et al., 2023; Eisape et al., 2024; Tessler et al., 2022). It contains the details of 139 participants who solved a set of 64 syllogistic problems (comprising qualifiers: All (A), No (E), Some (I), Some-not (O)), including their choice of response. The fields in the dataset are the following: *individual_id*: unique identifier for each individual; *problem_id*: unique identifier for each syllogistic problem; *premises*: a pair of syllogistic premises, delimited using '/', e.g., *"All bakers are chefs/Some doctors are chefs"*; *choices*: a list of nine possible choices of conclusions, including "NVC" (no valid conclusion), delimited using '|'; *individual_response*: conclusion selected by an individual.

Dataset Description - Spatial Reasoning For spatial reasoning, we used the dataset introduced by (Ragni, Brand, & Riesterer, 2021).[1] It contains the details of 49 participants who solved a set of 64 spatial relational problems, including their response, to eight given choices of conclusions per problem. While the original dataset consists of many fields, we cleaned it of extraneous details and formatted it. The following are the resultant fields: *individual_id*: unique identifier for each individual; *problem_id*: unique identifier for each spatial problem; *premises*: a pair of premises, delimited using '/', which specify the location of three entities with respect to one another, e.g., *"the mall is south of the park/the park is west of the bridge"*; *task*: it asks for the directional relation of the third entity with respect to the first, e.g., *"the bridge is ____ of the mall"*; *choices*: a list of eight choices of conclusions, delimited by '|', such as: *"the bridge is north-east of the mall|the bridge is east of the mall|... "* – each choice pertains to a cardinal direction; *individual_response*: the choice of response of the individual.

Automatic Generation of Cognitive Models

We now provide a general protocol (in collaboration with ChatGPT) to evaluate and utilize a transformer-based model's ability to predict an individual's responses to reasoning problems based on their previous responses (problems pertaining to a certain domain of human reasoning). The model's task

[1] GitHub link: https://github.com/CognitiveComputationLab/ccobra/

Table 1: Comparison table depicting features of various state-of-the-art cognitive models that predict syllogistic and spatial relational reasoning, including the presented transformer-based model CMA-TM (using ChatGPT o1-mini).

State-of-Art Models for **Syllogistic Reasoning**	State-of-Art Models for **Spatial Reasoning**
Atmosphere Theory (AT) (Woodworth & Sells, 1935): 1. Premises set atmosphere for conclusions (as TransSet) 2. Ordering independence of premise qualifiers	**Most-Frequent Answer (MFA)** Empirical baseline measure: Returns the most frequent response from the training set
PHM (Oaksford & Chater, 2007): 1. Fixed order of informativeness for qualifiers (A>I>E≫O) 2. Conclusion type based on heuristics (e.g., max, min, etc.)	**Transitive Closure**: A logic-based mechanism: 1. Uses rules such as transitive inferences 2. Determines membership of conclusion in closure of inferred spatial relations
TransSet (Brand, Riesterer, & Ragni, 2022): 1. Uses a two-phase process: direction and quantifier selection 2. Infers direction from term order and transitive paths in figures 3. Quantifiers are combined heuristically to derive a conclusion 4. Returns "NVC" if transitivity fails or specific rules (e.g., negativity, particularity) apply	**PRISM** (Ragni & Knauff, 2013): 1. Builds a preferred model in a spatial array 2. Draws inferences by model construction/variation 3. Variation applies minimal changes 4. Reasoning difficulty depends on focus operations
mReasoner (MMT) (Khemlani & Johnson-Laird, 2016): 1. Uses heuristics and strategies to construct mental models 2. Generation of weaker conclusions on falsification of initial conclusions and exploration of other mental models 3. Utilizes parameters for mental model representations (e.g., ω)	**Verbal** (Krumnack, Bucher, Nejasmic, & Knauff, 2010): 1. "Mental queue" representing objects 2. Directional encodings and operations 3. Conclusions based on relations in mental queue
CMA-TM: 1. Maps premise-pair qualifiers to conclusion qualifiers while tracking frequency 2. Ordering independence during training data mapping (cp. AT) 3. Selection of most common conclusion types 4. Fallback, e.g., conclusion type with highest bias in training	**CMA-TM**: 1. Geometric generation of all possible directions via grid-based (vector) sampling 2. Grouping pairs of directional relations in premises using sets of resulting possible directions 3. Also uses direction with most bias in a group

is to analyse the data for underlying cognitive patterns and features (if a sample dataset is available), and generate a predictive algorithm with aspects suitable for the goal.

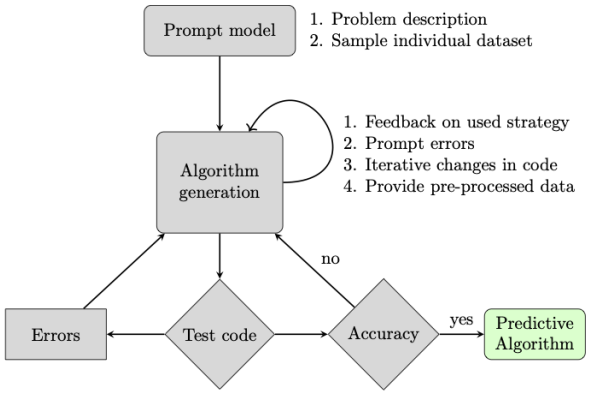

Figure 1: General Protocol for the automatic generation of predictive algorithms using transformer-based models.

In the following, we briefly outline the steps of the iterative and collaborative process which begins with *prompting the transformer-based model*. This involves providing the model with an adequate description of the prediction task, along with a description of any available data, such as the fields of the dataset. A sample individual dataset with training and test problems that adheres to the model's input constraints – e.g., size, format etc. may also be provided to the model.[2] The transformer-based model then returns a code in the specified programming language, and an explanation for its choice of algorithm. This can be followed by an *iterative cycle of feedback and code improvement*, including providing pre-processed data, example-based prompts or a local data analysis for verification of the algorithm's features.

Following the above, a *local test of the generated code* is performed (after one or more iterations), in which case: errors, if any, are prompted back to the transformer-based model which then generates an alternate code. In the event of predictions with a single qualifier overfitting the data, this is prompted back to the model, for e.g., if predictions for all syllogistic tasks are NVC, or bear the same qualifier like "All". Finally, the predictions are locally compared with the actual responses of the individuals and the accuracy of the algorithm

[2]The prompts are available via the OSF link: https://osf.io/6swjp/

is calculated. Depending on the accuracy, the transformer model is either (re)prompted to keep looking for cognitive patterns in the provided dataset, adopt alternative strategies etc., or the algorithm is finalised. And a suitable prompt may be reconstructed which can help reproduce the generated algorithm in the future, across other transformer-based models. Figure 1 illustrates the steps of the above generalised strategy.

The process is susceptible to some problems, such as parts of the generated algorithm deviating from the instructions provided in the prompt and its goal. Potential workarounds may be human code reviews, usage of descriptive comments, appropriate variable names, prompting with higher level of detail to prevent any misinterpretations, or adjusting temperature, etc.

Evaluation

We followed the generalised protocol outlined above to assess the generation of predictive algorithms using transformer-based models. We used the earlier presented syllogistic and spatial reasoning datasets and selected ChatGPT models o1-mini and o1 (at the time of writing this paper) for the purpose. We have discussed the particulars of the process which varies with the distinct nature of each domain, below. In both cases however, the data of each individual is split such that 5/8th of the data, i.e., the *first* 40 problems (selected sequentially), are fed to the algorithm as training data. And 3/8th of the data, i.e., the remaining 24 problems, constitute test problems that the algorithm is required to predict. Finally, the accuracy of the predictions is evaluated. For brevity, we henceforth refer to this model as CMA-TM (**C**ognitive **M**odel **A**ugmentation with **T**ransformer-based **M**odels).

Syllogistic Predictive Algorithm

Prompt As stated in the protocol, the development of a prompt capable of producing consistent and reproducible results with transformer-based models is iterative. We started the process with ChatGPT o1-mini, using base prompts, which involved providing the model with a sample individual dataset, description of the intended goal, and instructions like developing a cognitive profile of the individual, looking for observable biases for certain conclusion types, logical consistency etc. We also added some extra fields to the original dataset, such as: *general_form*, which comprised the generalised structure of the pair of premises, and *term_format*, which comprised the figure (1-4) of the syllogistic problem.

Features in Algorithm The algorithm generated to predict individuals performing syllogistic reasoning has been illustrated in Algorithm 1. Based on our analysis and query to ChatGPT o1-mini, the algorithm comprises several primary features highlighted in various steps. In the training phase each pair of qualifiers extracted from the premises in the training data is *mapped* to the qualifier types of their conclusions, while keeping track of the most commonly or frequently mapped conclusion types (steps 3 and 4). The *most common* conclusion types in turn qualify as the predicted conclusion types, for test data having similar qualifier pairs (step 6) – it is possible to have *multiple* qualifiers that have the highest frequency. Moreover, the training mapping of premise qualifiers is *independent* of their *ordering* (step 3). For e.g., All/Some is treated the same as Some/All. This approach allows the model to recognize that the logical impact of the premises remains the same regardless of their sequence. Some *fallback measures* are also used, e.g., choosing the conclusion type with highest training bias if the predicted conclusion type does not match available choices, or predicting NVC if test premise-pair types are unseen in training (steps 7 and 9).

Algorithm 1 Predictive Algorithm for Individual Responses to Syllogistic Problems (by ChatGPT o1-mini – Abstracted)

Input: CSV dataset with columns: individual_id, problem_id, premises, general_form, term_format, choices, response
Output: CSV file containing predicted responses with fields: individual_id, problem_id, predicted_response.
Steps:

1: For all problems, classify the type (A, E, I, O, NVC, Unknown) of each premise and conclusion based on keywords like "all", "no", "some", and "not".
2: Divide each individual data into training and test sets.
3: In training, create order-independent mapping from premise-pair types to the conclusion types.
4: For training premise-pair types, identify the conclusion type(s) with the highest frequency (multiple possible).
5: **function** process_syllogistic_reasoning(Data Frame)
6: For test problems, classify premise types and retrieve the most common conclusion types from training mapping.
7: If premises are unseen, default to NVC.
8: **function** select_conclusion(predicted_qualifier_type, choices, term_format, premises, training_bias):
9: For each individual, select conclusion(s) from choices based on predicted type(s), applying fallback on type with (highest) training bias when necessary.
10: Save the predictions in a CSV file.

Performance The CMA-TM Algorithm 1, achieves a mean prediction accuracy of 68% – the highest among state-of-the-art cognitive models for syllogistic reasoning – with maximum accuracy of 96% for some individuals. The graph in Figure 2 visualises the distribution of the predictions made by the model for 50 individuals, in comparison with the other cognitive models, such as PHM, mReasoner etc. The CMA-TM model achieves the highest performance with 68%, followed by TransSet at 49%. The AT model reaches 46%, while both MMT and PHM perform similarly at 45%. The model randomly selecting a response (uniform distribution) has the lowest accuracy of 12%. Table 2 illustrates the mean accuracies across the various models. Table 1 illustrates features characterizing each. The predictive syllogistic algo-

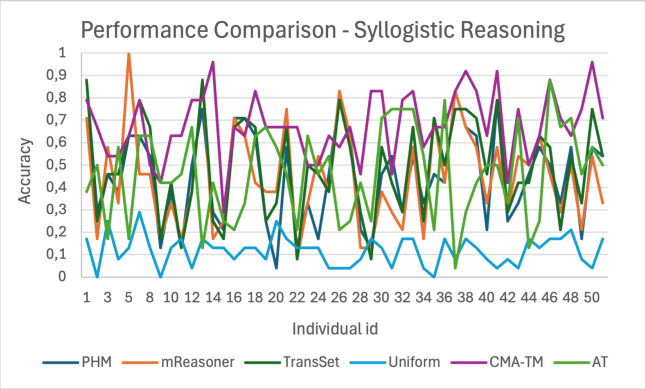

Figure 2: Performance comparison of CMA-TM with state-of-the-art cognitive models for Syllogistic reasoning.

Table 2: Mean Accuracy Percentage of CMA-TM and other state-of-the-art cognitive models. We use the following notations: AT: Atmosphere Theory; MMT: mReasoner; PHM: Probability Heuristics Model; Uni: Uniform.

CMA-TM	TransSet	AT	MMT	PHM	Uni
68%	49%	46%	45%	45%	12%

rithm takes a heuristic based approach (like AT, PHM, TransSet etc.) with the qualification that it is based on heuristics that vary across individuals, albeit the search space depending on the training data. For e.g., AT predicts that premise qualifiers of types OO and EE, induce response types O and E, respectively. However, analysing individual datasets shows that individual 30 uses a heuristic where both premise types OO and EE are commonly mapped to response type I. The predictive accuracy of the AT for this individual is 71%, while CMA-TM reaches 83%.

Spatial Predictive Algorithm

Prompt As with the syllogistic algorithm, in this case too the prompt was developed iteratively. During the iterations we pre-processed the data by grouping the direction-pair extracted from the premise-pair of each problem id. The grouping was done based on the sets of directions logically possible between the entities in the task field of the problem. For e.g., the direction pairs *North-East/East* and *East/North* extracted from two premise-pairs belong to the same group, as the logically possible direction for the task in both cases is *South-West*. After grouping, the data was then supplemented to the model in successive prompts.

Features in Algorithm The algorithm generated to predict individuals performing spatial reasoning has been illustrated in Algorithm 2. The task field could be left out of the input dataset as it was redundant in the final algorithm, given the choices field. Based on our analysis and query to ChatGPT o1-mini, the algorithm comprises several primary features: the direction pairs from the premises of each problem are first extracted. Then, using *vector logic*, the set of all *logically possible directions* that can be inferred between the entities in the corresponding task field is computed (step 3). And the direction pairs from all premises are *grouped* based on the aggregated set of logical possibilities (steps 4 and 5). During training, the direction-pair extracted from each premise-pair is matched against the groups, assigning the direction of the corresponding conclusion to the matching group, while also maintaining a *frequency count* (step 6). This serves the purpose of tracking the inferred (conclusive) directions with the *highest bias* within a group, leveraging patterns in the training set. Finally the algorithm extracts the direction-pair from the test premises, identifies its matching group, and uses the inferred direction with the most bias in the group, and the aggregated set of possible directions as the prediction (step 7).

Algorithm 2 Predictive Algorithm for Individual Responses to Spatial Problems (by ChatGPT o1-mini – Abstracted)

Input: CSV dataset with columns: individual_id, problem_id, premises, choices, individual_response
Output: CSV file containing predicted responses with fields: individual_id, problem_id, predicted_responses.
Steps:

1: Load the CSV file, grouping rows by individual ID.
2: Divide each individual data into training and test sets.
3: For each problem, use vector logic to compute all logically possible directions for the conclusion.
4: Group all premise-pair directions, based on sets of all logically possible directions for conclusion.
5: Also maintain an aggregate mapping of the groups to the set of logically possible directions.
6: For each training problem, identify group for extracted premise-pair directions, and assign its conclusion direction to the group, while keeping count of frequency.
7: For each test problem, identify its group and merge the aggregated possible directions for that group with the most frequent training directions for the same.
8: Save the final predicted responses to a CSV file.

Performance The CMA-TM Algorithm 2, achieves a mean prediction accuracy of 81%, which is the highest among state-of-the-art cognitive models for spatial relational reasoning, with maximum accuracy of 100% for some individuals. The graph in Figure 3 visualises the distribution of the predictions made by CMA-TM for 49 individuals, in comparison with other cognitive models, such as PRISM, Verbal Reasoning etc. The CMA-TM model achieves the highest performance at 81%, followed by Verbal with 68%. Both PRISM and MFA perform similarly at 67%, while TC reaches 42%. The Random model shows the lowest performance, achieving only 11%. Table 3 illustrates the mean accuracies across the

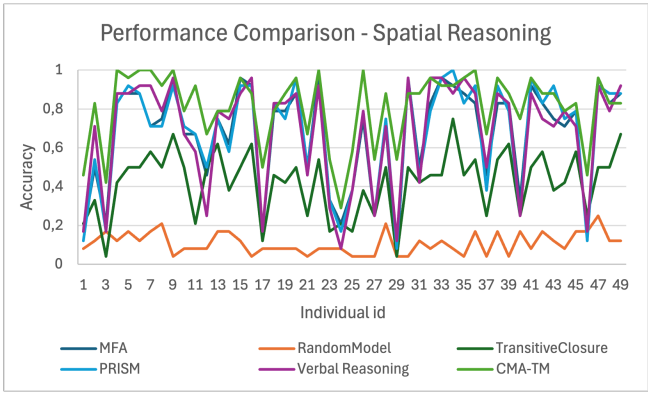

Figure 3: Performance comparison of CMA-TM with state-of-the-art cognitive models for Spatial reasoning.

various models. Features characterising each model can be found in Table 1.

Table 3: Mean Accuracy Percentage of CMA-TM and other state-of-the-art cognitive models. We use the abbreviations: MFA: Most-Frequent Answer; TC: Transitive Closure.

CMA-TM	Verbal	PRISM	MFA	TC	Random
81%	68%	67%	67%	42%	11%

Upper limits - Replicability across LLMs

To ensure that the protocol presented in this paper is generalizable across LLMs, we conducted tests with CoPilot Pro and Gemini 1.5 Pro. It was possible to regenerate the predictive algorithms using both models, albeit with some (minor) limitations. CoPilot Pro currently has a size constraint (10240 characters) for input prompts. While the prompt for the syllogistic algorithm was within the range, the prompt for the spatial algorithm had to be divided between two prompt iterations. In case of Gemini 1.5 Pro, the algorithms generated for both prompts had some code-completion issues, e.g., incomplete declaration of lists. However, with these issues fixed locally, the resulting codes produced the expected outputs. When testing the prompt with Gemini 2.0 Flash, the model failed to generate the entire algorithm.

General Discussion

Cognitive modelling for individuals can be a challenge due to the variety of reasoning patterns, biases, interpretations etc., in a population. Most traditional cognitive theories often adopt a "one-size-fits-all" approach that fails to account for such diversity at the individual-level. In comparison, only some models provide cognitive modelling for the individual (e.g., mReasoner, TransSet, MFA, Verbal etc.; see Table 1). The goals of this paper are thus two-fold: to build cognitive algorithms that can be used to predict an individual instead of being limited to an average reasoner, and importantly, to automatically generate such algorithms in a collaborative setting using transformer-based models. Training with a vast amount of human data, and various parameters used in the process, make LLMs potentially powerful models that can build upon implicitly learned reasoning patterns and biases.

To that end, we present a general methodological approach that can be applied across various reasoning domains, to evaluate and harness a transformer-based model's ability to generate algorithms capable of predicting individuals. The outlined approach leads us to what we call the CMA-TM model, using ChatGPT o1-mini and o1. We evaluate the model against two diverse domains of reasoning – syllogistic and spatial. The predictive and explainable algorithms generated by CMA-TM outperform the existing state-of-the-art cognitive models by far. The syllogistic predictive algorithm achieves an accuracy of 68%, showing an improvement of 19% over the model TransSet, which has an accuracy of 49%, (previously) the highest among the existing syllogistic models. The spatial predictive algorithm achieves an accuracy of 81%, showing an improvement of 13% over the model Verbal Reasoner, which has an accuracy of 68%, (previously) the highest among the existing spatial models.

Both CMA-TM algorithms are largely heuristics-based, rather than using statistical or machine learning models to make predictions. The predictive syllogistic algorithm trains using order-independent mapping between premise-types and conclusion-types. For a test problem, it uses the most common conclusion types learned from the training mapping for its predictions. The algorithm could potentially be adapted to include a wider range of predictions through adjustments to the minimum frequency threshold. The predictive spatial algorithm applies geometric reasoning on directional relations in premises, to derive logically possible directional relationships for the conclusion; groups them accordingly, and tracks the direction with highest training bias in each group, to arrive at its predictions.

Reproducibility of the presented syllogistic and spatial predictive algorithms using the respective prompts was successfully validated in the GPT playground (using multiple GPT models like 4o-mini and 4.1-mini), at temperature 0. The structure of the presented protocol used to generate these algorithms can also be used to generate other such explainable and replicable predictive algorithms using transformer-based models in different domains of cognitive science, including decision making. It combines analysis of experimental data and the framework of transformer-based models to boost cognitive modelling for human reasoning, and helps to address the greatest challenge in cognitive modelling of the individual, namely the need for a specific model for each individual. Future work will investigate how ensemble models and more hybrid AI-cognitive approaches can be used to reach the upper empirical limits (Riesterer et al., 2020).

Acknowledgments

This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) and DAAD (German Academic Exchange Service) in project 57616814 (SECAI, School of Embedded and Composite AI), which is gratefully acknowledged. Funding by grants to MR in the DFG projects 529624975 and 283135041 is gratefully acknowledged.

References

Brand, D., Riesterer, N., & Ragni, M. (2022). Model-based explanation of feedback effects in syllogistic reasoning. *Topics in Cognitive Science*, *14*(4), 828–844.

Eisape, T., Tessler, M., Dasgupta, I., Sha, F., van Steenkiste, S., & Linzen, T. (2024). A systematic comparison of syllogistic reasoning in humans and language models. In *Proc. of naacl 2024: Human language technologies* (pp. 8425–8444).

Fugard, A. J. B., & Stenning, K. (2013). Formal models of reasoning in cognitive psychology. *Argument & Computation*, *4*(1), 89–102. (Received 6 December 2011, Accepted 6 March 2012, Published 1 March 2013)

Johnson-Laird, P., & Ragni, M. (2023, 11). What should replace the turing test? *Intelligent Computing*, *2*.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.

Khemlani, S., & Johnson-Laird, P. N. (2016). Reasoning about possibilities: Individual differences in the evaluation of syllogisms. *Journal of Cognitive Psychology*, *28*(4), 489–511.

Krause, S., & Stolzenburg, F. (2024). From data to commonsense reasoning: The use of large language models for explainable ai. *arXiv preprint arXiv:2407.03778*.

Krumnack, A., Bucher, L., Nejasmic, J., & Knauff, M. (2010). Spatial reasoning as verbal reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Ragni, M., Brand, D., & Riesterer, N. (2021). The predictive power of spatial relational reasoning models: A new evaluation approach. *Frontiers in Psychology*, *12*, 626292.

Ragni, M., Dames, H., Brand, D., & Riesterer, N. (2019). When does a reasoner respond: Nothing follows? In *Proceedings of the cogsci 2019* (pp. 2640–2546).

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561-580.

Riesterer, N., Brand, D., & Ragni, M. (2020). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, *12*(3), 960–974.

Stenning, K., & Cox, R. (2006). Reconnecting interpretation to reasoning through individual differences. *Quarterly Journal of Experimental Psychology*, *59*(8), 1454–1483.

Tessler, M. H., Tenenbaum, J. B., & Goodman, N. D. (2022). Logic, probability, and pragmatics in syllogistic reasoning. *Topics in Cognitive Science*, *14*(3), 574-601.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*(2017).

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451.

Wu, Y., Han, M., Zhu, Y., Li, L., Zhang, X., Lai, R., ... Cao, Z. (2023). Hence, socrates is mortal: A benchmark for natural language syllogistic reasoning. In *Findings of the association for computational linguistics: Acl 2023* (pp. 2347–2367).

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. *Communications Psychology*, *2*, 51.