Igor T. Podolak, prof. dr hab.
Faculty of Mathematics and Computer Science UJ
Phone: 607 913 367
e-mail: igor.podolak@uj.edu.pl

Igor T. Podolak

Igor T. Podolak, prof. dr hab. • Faculty of Mathematics and Computer Science UJ

Discipline Council for Computer Science
University of Warsaw

Kraków, 17 January 2026

Review of the doctoral thesis

Visual and language reasoning in deep learning models

by Spyridon Mouselinos,

under the supervision of
dr hab. Henryk Michalewski, University of Warsaw, Google,

and
dr Mateusz Malinowski, Moonvalley

1 Content and novelty

The dissertation focuses on deep learning reasoning problems and their limitations. The author concentrates on the weaknesses and deficiencies of reasoning in deep learning models, including limited understanding and how it may be enhanced in large language models for different objectives: scene analysis, code generation, and geometry understanding. Various frameworks are evaluated, some insight into the process of reasoning is provided. The topics cover some of the most important current research challenges in machine learning.

1.1 Bibliometric indicators

The thesis comprises three publications from 2022 to 2024 and an accompanying 70-page introduction. The dissertation consists of papers published at top-tier conferences; see the table below. All articles were written in small teams that included both supervisors. In each paper, the PhD candidate was the first author. The doctoral student is a co-author of some other papers, which focused

Dissertation papers: place, year, Ministry of Science and Higher Education (MNiSzW) points, conference rank, number of citations (excluding self citations), Google H-5 and impact factor IF.

	name	year	MNiSzW	CORE rank	citations	Google H-5	IF
I	ICLR	2022	200	A+	2	253	48.9
II	ACL	2023	200	A	16	218	n/a
III	EMNLP	2024	140	A+	10	387	n/a

on different problems and, therefore, were not included in this dissertation.

2 Objectives, hypotheses, degree of achievement

The author observes in the introduction that large deep ML models often succeed in solving large and complicated problems, only to fail on some seemingly simple or even trivial ones. The author examines problems in the visual, language, and multimodal domains that require the application of logical rules, but are prone to distraction. The research question posed in the introduction is: *How can AI systems progress beyond pattern recognition to cultivate strong reasoning capabilities across multiple domains with diverse contexts and modalities?* Asking this question directly is both bold and important, particularly when part of the ML community mainly focus on tuning simpler problems.

3 The publications included in the thesis

The included papers address the research question to varying degrees; it is now necessary to consider how deeply each of them engages with and answers it.

Measuring CLEVRness: black-box testing of visual reasoning models, ICLR 2022

The paper addresses the problem of measuring reasoning capabilities in visual tasks. The PhD candidate (together with his supervisors) proposes a black-box mechanism to test models based on an RL agent that modifies a scene and compares the performance of several models. In the proposed framework, the visual and adversarial models compete against each other: the adversarial RL agent performs scene manipulations, testing the visual model's response. The CLEVR dataset is used for evaluation (Johnson, 2017).

The paper demonstrates that some tested models are considerably more susceptible to scene modifications than would have been expected given their originally reported accuracy results. In my opinion, the paper primarily presents a benchmarking framework, which is likely to be quite useful for model comparison. However, I believe that the paper does not adequately address the problem of model *reasoning*, as suggested by the title. Although results for the models tested are reported, the authors do not attempt to explain *why* a given model—or, more precisely, its reasoning—is inferior to another. The paper does not consider different types of reasoning (e.g., deductive, inductive, abductive) that models might employ. In my view, reasoning differs from inference in that it is a structured, multi-step process that provides insight into the rationale and structure underlying how an answer is reached. The proposed benchmarking framework does not attempt to provide such insights. That said, perhaps the proposed RL approach could prove to be a suitable tool for addressing these questions in future work?

A simple, yet effective approach to finding biases in code generation, ACL 2023

In this paper, the author considers the problem of reasoning capabilities in the area of LLM use for code generation. The paper introduces a new framework, proposing a division of code into *blocks of influence* (name-, description-, and example-blocks) and examining different perturbation alternatives within a similar adversarial framework. Several benchmark datasets are used.

The authors show that a high number of modifications result in a drop in code correctness (measured with the Pass@k metric). They demonstrate, for example, that reducing the number of examples reduces correctness; similarly, perturbations of other blocks reduce code accuracy. From these findings and from training on augmented code snippets, they deduced that well-structured and

commented code would help LLMs generate better code, a conclusion that is likely correct. They also suggested that example removal or function anonymization may reveal poor reasoning in code generation LLMs. However, there is no attempt at deeper insight into the reasoning nature of these models, stopping at the level of better versus worse coding performance.

Beyond lines and circles: unveiling the geometric reasoning gap in large language models, EMNLP 2024

This paper attempts to build and describe a thought process for a computational model – an LLM that reasons using a range of tools. The area of interest is planar geometry, a domain in which language models still struggle. The author defines a set of tools to be used, an additional tuning model based on Euclid's Elements, and the process of selecting appropriate tools during reasoning. Viewed in this way, this paper provides the most in-depth analysis of solution search and the most thorough description of a model's *reasoning* process among the papers in this dissertation. It proposes concepts including example grouping and selection, extended prompt engineering, the addition of multiple conversational agents (simulacra) to help the LLM choose the most appropriate tools and arrive at correct answers, and the extension of *spatial awareness* through geometry-oriented tools. By the end of 2025, these approaches had become common, but at the time of publication they were considerably rarer. Different approaches and modifications are compared, with simulacra shown to achieve substantial improvements.

The paper introduces an interesting observation regarding LLM efficiency: these models perform well when generating solution ideas as a plan, yet exhibit a notable inability to execute such plans using geometric tools. The author attributes this to the limited geometric understanding of LLMs and proposes building two-agent systems to address this limitation. This solution was not well known at the time of publication. The paper also proposes a method for building better-structured prompts by using specialised tools to understand the scene and supplementing that information in the prompt.

3.1 Papers summary and comments

All papers cover the same subject and research area and clearly form a coherent series. All were published at high-tier ML conferences. The introduction to the series of works is thorough and comprehensive. The works were written and published in a short period of three years, from 2022 to 2024. As such, the

thesis fully meets the requirements for a doctoral dissertation consisting of a series of publications in the field.

Comments, remarks, questions.

1. Nowhere in the work is any attempt to define what reasoning is or how the reader should understand this construct. In the introduction, citing the Clever Hans paradox, the author notes that this concept can be understood in different ways, and in particular that we may 'perceive' reasoning where, in reality, there is none. How does the author understand this concept?

2. In the introduction, the author states an intriguing and important research question: *How can AI systems progress beyond pattern recognition to cultivate strong reasoning capabilities across multiple domains with diverse contexts and modalities?* In practice, only the last paper directly addresses this question by analysing a geometric modality, while the first two are primarily propositions for benchmarking frameworks. Could you comment on current definitions of reasoning in machine learning models?

3. In the paper *Measuring CLEVRness: Black-Box testing ...*, a *Visual-QA Player* framework is proposed. The *Adversarial Player* is RL-trained with the episode length set to one (see Section 5, paragraph "Training algorithm", page 6). Thus, is RL used solely for the sake of employing a reward-based loss function? The problem is defined in the form of mini-games (see Section 3). Why is this the case, since examples within individual mini-games appear not to have anything to do with each other?

4. In the paper on code generation using LLMs *A simple, yet effective approach ...*, most of the analysis is conducted by *removing* or *modifying* supposedly correct blocks of influence. Less attention is devoted to improving knowledge – or at least this is less evident in the paper. Why is this the case? Is the code generation ability connected to the in-context learning (ICL) capabilities of LLMs? I would like to know your opinion on the limits of ICL in LLMs[1]. In modern mechanistic interpretability approaches, the notion of *circuits* is popular[2]. Could a similar analysis be conducted for code generation, particularly given that attention scores are illustrated in the paper?

[1] See, e.g., Singh, *The Transient Nature of Emergent In-Context Learning in Transformers*, NeurIPS 2023.

[2] See, e.g., Wang et al., *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small*, ICLR 2022.

5. The paper on LLM use in geometry *Beyond lines and circles ...*, understanding addresses the challenging issue of reasoning to the greatest extent. It presents several interesting observations, for example that LLMs may be better at devising a solution plan than at executing it. The paper includes several propositions that have since become common in current LLM systems, such as a more widespread use of tools. This is the most advanced and reasoning-focused paper in the series.

4 Conclusions

The group of papers presented, together with the comprehensive introduction, form a well-founded thesis. I have raised some comments and concerns. These relate primarily to the fact that the author devoted less attention to defining and analysing the reasoning process, limiting the investigation to demonstrating that inference in such models can be disturbed. The issue of reasoning in ML models is very challenging and complex, and particularly important at present.

I assess the thesis as fully meeting all the requirements for doctoral dissertations set out in the relevant legislation. I recommend that it be admitted to the next stages of the procedure.

Sincerely,

Igor T. Podolak, prof. dr hab.