MACHINE LEARNING MEETS PROGRAM SYNTHESIS

Nathanaël Fijalkow
CNRS, LaBRI, Bordeaux, France

POPL Tutorial Fest 2024

Context: Large Language Models (LLMs) are taking over the world, including program synthesis

Questions:

(1) Is program synthesis solved?

(2) How can we (= Programming Languages, Formal Methods) contribute?

An interesting terminology shift:

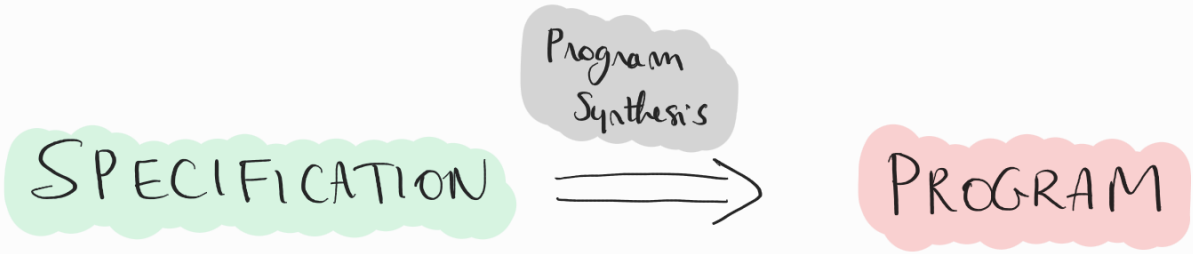
"program synthesis" has logical roots:
inductive programming, logic programming

a new term emerged with LLMs:

"code generation"

→ where is the specification?

WHAT IS PROGRAM SYNTHESIS ?



logical formula
examples
natural language description
demonstration
⋮

code
System
Chip
macro
⋮

- One of the oldest dream of AI
- For decades only for delusional researchers
- The upcoming revolution for | end users ?
  software (developpers) ?
  hardware ?

BACK IN 2011: FLASHFILL (Microsoft)

Summit Attendees

File Edit View Insert Format Data Tools Add-ons Help

	A	B	C	D	E	F
1	**Attendees**	**Office Address**	**Hi _____**	**E-mail**	**Zip Code**	
2	Gerald Parker	320 N Morgan St #600, Chicago, IL 60607	Hi Gerald			
3	Phillip Davis	111 8th Ave. New York, NY 10011				
4	Ralph Gonzales	2930 Pearl Street, Boulder, CO. 80301				
5	Shawn Jenkins	1160 Bordeaux Drive, Sunnyvale, CA 94089				
6	Thomas Murphy	1160 Bordeaux Drive, Sunnyvale, CA 94089				
7	Norma Long	2930 Pearl Street, Boulder, CO. 80301				
8	Beverly Harris	320 N Morgan St #600, Chicago, IL 60607				
9	Melissa Torres	1160 Bordeaux Drive, Sunnyvale, CA 94089				
10	Joshua Watson	111 8th Ave. New York, NY 10011				
11	Larry Wright	500 W 2nd St. Austin, TX 78701				
12	Lois Brooks	2930 Pearl Street, Boulder, CO. 80301				
13	Steve James	1160 Bordeaux Drive, Sunnyvale, CA 94089				
14	Helen Rivera	2930 Pearl Street, Boulder, CO. 80301				
15	Samuel Bailey	111 8th Ave. New York, NY 10011				
16	Kenneth Gonzalez	2930 Pearl Street, Boulder, CO. 80301				
17	Sharon Washington	1160 Bordeaux Drive, Sunnyvale, CA 94089				
18	Mildred Evans	1160 Bordeaux Drive, Sunnyvale, CA 94089				

Smartfill

Google's version

Limitations of FlashFill

- it doesn't know that Paris is in France!
- it doesn't output the program so we cannot trust the results

BACK IN 2021: CODEX (OpenAI / Github)

```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, value, currency).
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2016-01-02 -34.01 USD
        2016-01-03 2.59 DKK
        2016-01-03 -2.72 EUR
    """
    expenses = []
    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split(" ")
        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
                         float(value),
                         currency))
    return expenses
```

Copilot

Limitations of LLMs:

- treats code as raw text
- no guarantees (ambiguous specification)
- only works in scenarios with loads of data!

LLMs treat code as raw text

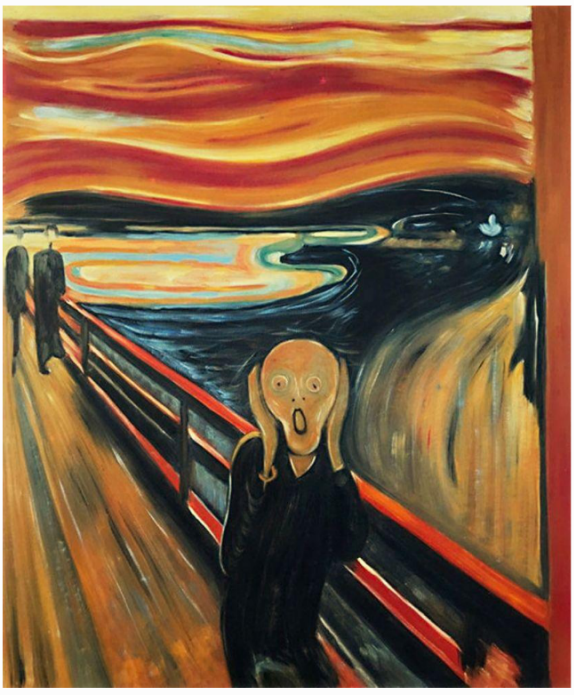
LLMs treat code as raw text

LLMs treat code as raw text

LLMs treat code as raw text

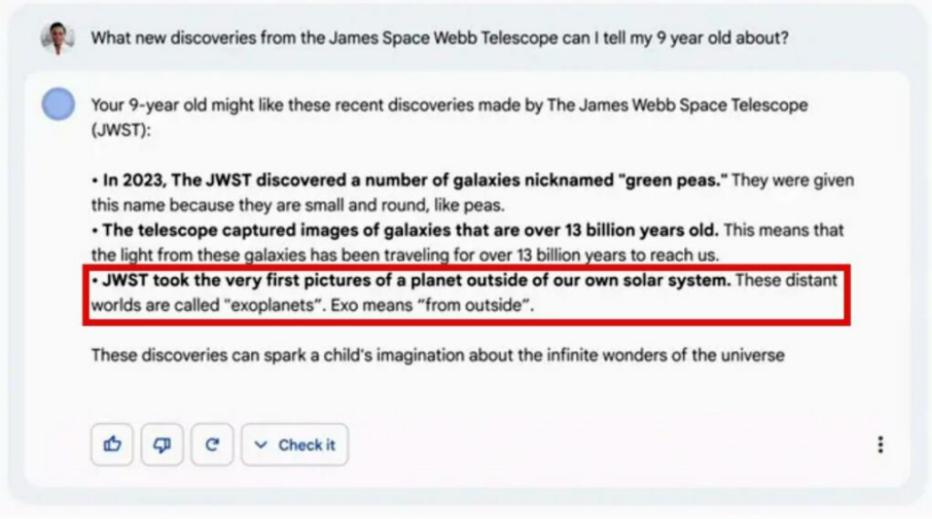
LLMs treat code as raw text

LLMs treat code as raw text



HAVE WE LEARNED NOTHING FROM 80+ YEARS OF PROGRAMMING LANGUAGES RESEARCH ?

LAST YEAR : BARD (Google)



Limitations of LLMs :

they doesn't know, they learn from biassed data

(also, they cannot do additions nor play Wordle)

CLEARLY, PROGRAM SYNTHESIS IS NOT SOLVED!

Goal of this tutorial:

present the techniques, challenges, and perspectives of neuro symbolic approaches to program synthesis rooted in programming languages, formal methods, algorithms

NEUROSYMBOLIC APPROACHES COMBINE

- neural networks (including LLMs) ← efficient
- logical reasoning / combinatorial search ← reliable

to achieve the best of both worlds

other names: neural guided, machine learned, neural } program synthesis

SOME RECENT TOOLS

Euphony

Bustle

DeepCoder

my group's

SyGuS competition

Herb

DeepSynth / ProgSynth

TF-Coder

Brute

DreamCoder

FlashFill / FlashMeta

GOAL

Present the common denominator to the tools above

OUTLINE

Part I : Machine Learning

Part II : Programming Languages

Part III : Algorithms

Part IV : Perspectives

IN THIS LLM WORLD,
NEUROSYMBOLIC PROGRAM SYNTHESIS
IS ONE OF THESE TWO :

a sinking ship

a messiah