Integrating Symbolic Modules, Constraints, and Knowledge Into Neural Language Models
LARGER LANGUAGE MODELS

The scale of text-generating neural networks is growing exponentially, as measured by the models' parameters (roughly, the number of connections between neurons).

- "Dense" models
- "Sparse" models

*Google's 1.6-trillion parameter 'sparse' model has performance equivalent to that of 10 billion to 100 billion parameter 'dense' models.

[Peters et al. '18, Radford et al. '19, Brown et al. '20, ...]
... The meaning of life

[Diagram: Graph showing the relationship between compute (PF-days, non-embedding) and test loss. The graph has a power law model equation: $L = (C_{min}/2.3 \cdot 10^8)^{-0.050}$.

[Kaplan et al. 2020, Scaling Laws for Neural Language Models]

Compute $\approx$ number of parameters $\times$ tokens processed
What is the meaning of life? 

The meaning of life is a question that has been asked by people throughout history. There is no one correct answer to this question.

I am a highly intelligent question answering bot.

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.

Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.
On the other hand, we find that scale has a reduced benefit for tasks in the Maths, Logical Reasoning, and Common Sense categories. Our results suggest that for certain flavours of mathematical or logical reasoning tasks, it is unlikely that scale alone will lead to performance breakthroughs. In some cases Gopher has a lower performance than smaller models—examples of which include Abstract Algebra and Temporal Sequences from BIG-bench, and High School Mathematics from MMLU.
Reasoning

Claim: One is a number that comes \textit{after} zero. TRUE

Claim: One is a number that comes \textit{before} zero. TRUE

Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations
	extit{EMNLP} 2022.
Problem: It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

Let's think step by step:

1. Jennifer spends 20 minutes grooming each of her 2 long hair dachshunds.
2. If she grooms her dogs every day, she spends 40 minutes grooming her dogs every day.
3. In 30 days, she spends 40 minutes x 30 days grooming her dogs, which is 1200 minutes, or 20 hours.

We can set up a proportion to solve this problem:

\[
\begin{align*}
20 \text{ minutes} & : 1 \text{ dog} \\
x \text{ minutes} & : 2 \text{ dogs}
\end{align*}
\]

We want to find \(x\), so we can cross multiply:

\[
20x = 2(20)
\]

\[
x = 40 \text{ minutes}
\]

Now we need to convert 40 minutes to hours. There are 60 minutes in an hour, so we can divide:

\[
40 \text{ minutes} / 60 \text{ minutes} = 0.67 \text{ hours}
\]

Jennifer spends 0.67 hours grooming her dogs in 30 days.
Reasoning

**MATH Dataset (Ours)**

**Problem:** Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

**Solution:** There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = 7$.

**Problem:** The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

**Solution:** Complete the square by adding $1$ to each side. Then $(x + 1)^2 = 1 + i = e^{\frac{\pi}{4}} \sqrt{2}$, so $x + 1 = e^{\frac{\pi}{8}} \sqrt{2}$. The desired product is then $\left( -1 + \cos \left( \frac{\pi}{8} \right) \sqrt{2} \right) \left( -1 - \cos \left( \frac{\pi}{8} \right) \sqrt{2} \right) = 1 - \cos^2 \left( \frac{\pi}{8} \right) \sqrt{2} = 1 - \frac{1 + \cos \left( \frac{\pi}{8} \right)}{2} \sqrt{2} = \frac{1 - \sqrt{2}}{2}$.

“Assuming a log-linear scaling trend, models would need around $10^{35}$ parameters to achieve 40% on MATH, which is impractical.”

[Hendrycks et al, NeurIPS 2021, Measuring Mathematical Problem Solving With the MATH Dataset]
Control

Real Toxicity Prompts

Example 'Very Toxic' GPT-2 generations

I swear, I just don't know it. And this.
What is the mass of Jupiter?
Overview

- **Modularity**
  - Single monolithic system → decomposed neural & symbolic modules

- **Constraints**
  - Discrete logical constraints

- **Knowledge**
  - Hand-crafted → generated and distilled
Modularity

- Conventional: generate from a single monolithic model

- **Rapidly expanding trend**: generate with multiple, composed modules. Modules can be neural or symbolic.
  - Expanded capabilities
    - Some functionality is difficult to learn, yet easy for symbolic modules (e.g. calculation, internet search).
  - Stronger generalization
    - Symbolic layer on top of noisy enumerator
Modularity

Language Model Cascade [Dohan et al 2022]

• View language model as a single module.
• Form a “cascade” of multiple modules that interact via text.
  • Module: string-valued random variable.
  • Interact: observed value.
It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?
It takes Jennifer 20 minutes to groom each of her 2 long hair dachschunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

```
def qa():
    q = yield s('question')
    a = yield s('answer', question=q)
    return a
```
It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

Sam memorized six more digits of pi than Carlos memorized...

... 

Carly collected 7 starfish with 5 arms each and one seastar with 14 arms. How...
Modularity

- Intermediate rationale $z$

\[ p(a \mid q) = \sum_z p(a \mid q, z)p(z \mid q) \]

- Approximation:

\[
\hat{z} \sim p(z \mid q) \\
\hat{a} \sim p(a \mid q, \hat{z})
\]

It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

Jennifer spends 40 minutes per day grooming her dachshunds. In 30 days she spends 1200 minutes. Thus the answer is 20 hours.

Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems
W. Ling, D. Yogatama, C. Dyer, P. Blunson
ACL 2017.
**Modularity**

- **Prompted intermediate rationale** $z$

$$p(a | q; D) = \sum_z p(a | q, z; D)p(z | q; D)$$

---

It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

Jennifer spends 40 minutes per day grooming her dachshunds. In 30 days she spends 1200 minutes. Thus the answer is 20 hours.

---

[Wei et al 2022]
Modularity | symbolic tools

\[ p(a \mid q) = \sum p(a \mid q, z)p(z \mid \text{exec}(z'), q)p(z' \mid q) \]
It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

Jennifer spends 40 minutes per day grooming her dachshunds.

In 30 days she spends $30 \times 40 = 1200$ minutes.

Thus the answer is 20 hours.
It takes Jennifer 20 minutes to groom each of her 2 long hair dachshunds. If she grooms her dogs every day, how many hours does she spend grooming her dogs in 30 days?

Jennifer spends 40 minutes per day grooming her dachshunds.
In 30 days she spends 30*40 = [CALCULATOR]

Thus the answer is 20 hours.

Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems
W. Ling, D. Yogatama, C. Dyer, P. Blunson
ACL 2017.
Modularity

- [Cobbe et al 2021]: GPT-3 + supervised rationales + calculator

**Problem:** Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?

**Solution:**

- Tina buys 3 12-packs of soda, for 3*12 = 36 sodas.
- 6 people attend the party, so half of them is 6/2 = 3 people.
- Each of those people drinks 3 sodas, so they drink 3*3 = 9 sodas.
- Two people drink 4 sodas, which means they drink 2*4 = 8 sodas.
- With one person drinking 5, that brings the total drank to 5+9+8+3 = 25 sodas.
- As Tina started off with 36 sodas, that means there are 36-25 = 11 sodas left.

**Final Answer:** 11
Modularity

Grade School Math 8k

- 6B language model Q → A
- + rationale + calculator
- + rationale + calculator + verifier
- w/ 175B generator + 6B verifier
Modularity

6B language model Q —> A
+ rationale + calculator
+ rationale + calculator + verifier
w/ 175B generator + 6B verifier
Modularity

Learned verifier

Grade School Math 8k

- 6B language model Q -> A
- + rationale + calculator
- + rationale + calculator + verifier
- w/ 175B generator + 6B verifier
Modularity

Learned verifier

Grade School Math 8k

- 6B language model Q → A
- + rationale + calculator
- + rationale + calculator + verifier
- w/ 175B generator + 6B verifier
Modularity

Learned verifier

Grade School Math 8k

- 6B language model Q -> A
- + rationale + calculator
- + rationale + calculator + verifier
- w/ 175B generator + 6B verifier
Modularity | other tools

- Tool Augmented Language Models
  [Parisi et al 2022]
Modularity | other tools

- Tool Augmented Language Models [Parisi et al 2022]
  - **Tool**: Retrieval/web-search

| Question: | when are hops added in brewing process? |
| Short Answer: | The boiling process. |

```latex
|question when are hops added in brewing process? |result The boiling process is where chemical reactions take place...including |output The boiling process. |
```
Modularity | other tools

- Tool Augmented Language Models [Parisi et al 2022]
  - **Tool**: Retrieval/web-search
• **Lila Benchmark** [Mishra et al 2022]
  Unifies 20 math datasets:

  • **‘Rationale’**: python program

  • **Tools**: libraries (numpy, ...), standard Python (variables, ...)

---

**Problem:**
The pirates plan to explore 4 islands. Two islands require walking 20 miles per day while the other two islands require 25 miles per day. How many miles will they have to walk if it takes 1.5 days to explore each island?

**Program:**
```python
a=20*2
b=25*2
c=a+b
d=c+1.5
answer=d
print(answer)
# == > 135.0
```

---

**Problem:**
Compute the nullity of $\begin{pmatrix} -9 \\ -2 \\ 3 \\ -1 \end{pmatrix}$.

**Program:**
```python
import numpy as np
a = np.array([[[-9], [-2], [3], [-1/2]])
r = np.linalg.matrix_rank(a)
print(len(a[0]) - r)
# == > 0.0
```
Modularity | other tools

- Lila Benchmark
  - Program + execution > answer
    - In-domain & OOD generalization

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Neo-A</th>
<th>Neo-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math ability</td>
<td>0.191</td>
<td>0.445</td>
</tr>
<tr>
<td>Language</td>
<td>0.189</td>
<td>0.429</td>
</tr>
<tr>
<td>Format</td>
<td>0.246</td>
<td>0.372</td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.206</td>
<td>0.331</td>
</tr>
<tr>
<td>Average</td>
<td>0.208</td>
<td>0.394</td>
</tr>
</tbody>
</table>
Problem
Let \( P_1(x) = x^2 - 2 \) and \( P_j(x) = P_1(P_{j-1}(x)) \) for \( j = 2, \ldots \) Prove that for any positive integer \( n \) the roots of the equation \( P_n(x) = x \) are all real and distinct.

Solution
I shall prove by induction that \( P_n(x) \) has \( 2^n \) distinct real solutions, where \( 2^{n-1} \) are positive and \( 2^{n-1} \) are negative. Also, for every root, \( |r| < 2 \).

Clearly, \( P_1(x) \) has 2 real solutions, where 1 is positive and 1 is negative. The absolute values of these two solutions are also both less than 2. This proves the base case.

Now assume that for some positive integer \( k \), \( P_k(x) \) has \( 2^k \) distinct real solutions with absolute values less than 2, where \( 2^{k-1} \) are positive and \( 2^{k-1} \) are negative.

Choose a root \( r \) of \( P_{k+1}(x) \). Let \( P_1(r) = s \) where \( s \) is a real root of \( P_k(x) \). We have that \(-2 < s < 2\), so \( 0 < r^2 < 4 \), so \( r \) is real and \( |r| < 2 \). Therefore all of the roots of \( P_{k+1} \) are real and have absolute values less than 2.

Note that the function \( P_{k+1}(x) \) is an even function, since \( P_1(x) \) is an even function. Therefore half of the roots of \( P_{k+1} \) are positive, and half are negative.

Now assume for the sake of contradiction that \( P_{k+1}(x) \) has a double root \( r \). Let \( P_1(r) = s \). Then there exists exactly one real number \( r \) such that \( r^2 - 2 = s \). The only way that this could happen is when \( s + 2 = 0 \), or \( s = -2 \). However, \( |s| < 2 \) from our inductive hypothesis, so this is a contradiction.

Therefore \( P_{k+1}(x) \) has no double roots. This proves that the roots of \( P_{k+1}(x) \) are distinct.

This completes the inductive step, which completes the inductive proof.
Modularity | bridging informal+formal reasoning

Natural language mathematics

Flexibility

Data

Verifiability

Grounding

Problem
Let \( P_1(x) = x^2 - 2 \) and \( P_j(x) = P_j(P_{j-1}(x)) \) for \( j = 2, \ldots \). Prove that for any positive integer \( n \) the roots of the equation \( P_1(x) = x \) are all real and distinct.

Solution
I shall prove by induction that \( P_n(x) \) has \( 2^n \) distinct real solutions, where \( 2^{n-1} \) are positive and \( 2^{n-1} \) are negative. Also, for even root \( r \), \( |r| < 2 \).

Clearly, \( P_1(x) \) has 2 real solutions, when solutions are also both less than 2. This shows that \( P_n(x) \) has \( 2^n \) distinct real solutions with absolute values less than 2.

Now assume that for some positive integer \( n \) there are exactly \( 2^n \) distinct solutions with absolute values less than 2.

Choose a root \( r \) of \( P_{n+1}(x) \). Let \( P_1(r) \) be the root of \( P_1(x) \). We have that \( -2 < r < 2 \) and all \( x \) are real and \( |x| < 2 \).

Note that the function \( P_{n+1}(x) \) is an even function, since \( P_1(x) \) is an even function. Therefore half of the roots of \( P_{n+1}(x) \) are positive, and half are negative.

Now assume for the sake of contradiction that \( P_{n+1}(x) \) has a double root \( r \). Let \( P_1(r) = x \). Then there exists exactly one real number \( r \) such that \( r^2 - 2 = x \). The only way that this could happen is when \( x = 2 \) or \( x = -2 \). However, \( |x| < 2 \) from our inductive hypothesis, so this is a contradiction.

Therefore \( P_{n+1}(x) \) has no double roots. This proves that the roots of \( P_{n+1}(x) \) are distinct.

This completes the inductive step, which completes the inductive proof.
Modularity | bridging informal+formal reasoning

Natural language mathematics
- Flexibility
- Data
- Verifiability
- Grounding

Formalized mathematics
- Flexibility
- Data
- Verifiability
- Grounding

Problem
Let $P_k(x) = x^2 - 2$ and $P_f(x) = P_f(P_{k+1}(x))$ for $j = 2, \ldots$. Prove that for any positive integer $n$ the roots of the equation $P_k(x) = 0$ are all real and distinct.

Solution
I shall prove by induction that $P_k(x)$ has $2^n$ distinct real solutions, where $2^{n-1}$ are positive and $2^{n-1}$ are negative. Also, for every root $x$, $|x| < 2$.

Clearly, $P_1(x)$ has 2 real solutions, where solutions are also both less than 2. Thus, for $n = 1$ the absolute values of these two solutions are both negative. The absolute values of these two solutions add up to be negative. The absolute values of these two solutions add up to be negative.

Now assume that for some positive integer $n$, $P_k(x)$ has $2^n$ distinct real solutions with absolute values less than 2, where $2^{n-1}$ are positive and $2^{n-1}$ are negative. We have that $-2 < s < 2$, so $0 < r^2 < 4$, so $r$ is real and $|r| < 2$.

Note that the function $P_{k+1}(x)$ is an even function, since $P_k(x)$ is an even function. Therefore half of the roots of $P_{k+1}(x)$ are positive, and half are negative.

Now assume for the sake of contradiction that $P_{k+1}(x)$ has a double root $r$. Let $P_f(r) = s$. Then there exists exactly one real number $s$ such that $r^2 - 2 = s$. The only way this could happen is when $\sqrt{2} \pm 2 = \sqrt{2}$, so $r = \sqrt{2}$. However, $|r| < 2$, thus the inductive hypothesis, so this is a contradiction. Therefore $P_{k+1}(x)$ has no double roots. This proves that the roots of $P_{k+1}(x)$ are distinct.

This completes the inductive step, which completes the inductive proof.
Modularity | bridging informal+formal reasoning

**Natural language mathematics**

- Flexibility
- Data
- Verifiability
- Grounding

- Flexibility
- Data
- Verifiability
- Grounding

Best of both worlds?

**Problem**

Let \( P_i(x) = x^2 - 2 \) and \( P_j(x) = P_j(P_{j-1}(x)) \) for \( j = 2, \ldots \). Prove that for any positive integer \( n \) the roots of the equation \( P_n(x) = 0 \) are all real and distinct.

**Solution**

I shall prove by induction that \( P_i(x) \) has \( 2^i \) distinct real solutions, where \( 2^{i+1} \) are positive and \( 2^{i-1} \) are negative. Also, for every root \( r, |r| < 2 \).

Clearly, \( P_1(x) \) has \( 2 \) real solutions, when solutions are also both less than 2. Thus, \( P_2(x) \) has \( 4 \) real solutions, when solutions are also both less than 2. This continues. Now assume that for some positive integer \( i \), \( P_i(x) \) has \( 2^i \) distinct real solutions with absolute values less than 2, where \( 2^{i-1} \) are positive and \( 2^{i+1} \) are negative. The absolute values of these two sets of roots are all less than 2.

Now choose a root \( r \) of \( P_{i+1}(x) \). Let \( P_1(r) \) be the root of \( P_1(x) \). We have that \( -2 < r < 2 \), so \( 0 < r^2 < 4 \), so \( r \) is real and \( |r| < 2 \).

Note that the function \( P_{i+1}(x) \) is an even function, since \( P_i(x) \) is an even function. Therefore half of the roots of \( P_i(x) \) are positive, and half are negative.

Now assume for the sake of contradiction that \( P_{i+1}(x) \) has a double root \( r \). Let \( P_1(r) = x \). Then there exists exactly one real number \( s \) such that \( r^2 - 2 = s \). The only way this could happen is when \( x = \pm 2 \). However, \( |s| < 2 \) from our inductive hypothesis, so this is a contradiction. Therefore \( P_{i+1}(x) \) has no double roots. This proves that the roots of \( P_{i+1}(x) \) are distinct.

This completes the inductive step, which completes the inductive proof.
Modularity | sketching

- Draft-Sketch-Prove [Jiang et al 2022]

Statement

If $\gcd(n, 4) = 1$ and $\text{lcm}(n, 4) = 28$, show that $n$ is 7.
Modularity | sketching

- **Draft-Sketch-Prove** [Jiang et al. 2022]

**Statement**
if \( \gcd(n, 4) = 1 \) and \( \text{lcm}(n, 4) = 28 \),
show that \( n \) is 7.

**Informal proof**
We know that \( \gcd(a, b) \cdot \text{lcm}(a, b) = ab \),
hence \( 1 \cdot 28 = n \cdot 4 \).
Then \( 1 \cdot 28/4 = 7 \),
completing the proof. \( \square \)

**Draft informal proof**

*Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs*
A. Jiang*, S. Welleck*, J. Zhou*, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, Y. Wu
Modularity | sketching

• Draft-Sketch-Prove [Jiang et al 2022]

Statement
if gcd(n, 4) = 1 and lcm(n, 4) = 28, show that n is 7.

Informal proof
We know that gcd(a, b) \cdot lcm(a, b) = ab, hence 1 \cdot 28 = n \cdot 4.
Then n = 1 \cdot 28/4 = 7,
completing the proof.

Formal sketch
have c1: “1*28 = n*4”
using asms
<proof>
then have c2: “n = 1*28/4”
<proof>
then show ?thesis
<proof>

Draft informal proof

Generate formal sketch

Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs
A. Jiang*, S. Welleck*, J. Zhou*, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, Y. Wu
Draft-Sketch-Prove [Jiang et al 2022]

Statement
If \( \gcd(n, 4) = 1 \) and \( \text{lcm}(n, 4) = 28 \), show that \( n \) is 7.

Informal proof
We know that \( \gcd(a, b) \cdot \text{lcm}(a, b) = ab \), hence \( 1 \cdot 28 = n \cdot 4 \).
Then \( n = 1 \cdot 28 / 4 = 7 \), completing the proof. ■

Formal sketch
have c1: "1*28 = n*4"
using asms

then have c2: "n = 1*28/4"

then show ?thesis

Verified formal proof
have c1: "1*28 = n*4"
using asms
by (smt (z3) prod_gcd_lcm_nat)
then have c2: "n = 1*28/4"
by auto
then show ?thesis
by auto

Draft informal proof

Generate formal sketch

Prove remaining gaps

---

Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs
A. Jiang*, S. Welleck*, J. Zhou*, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, Y. Wu
Modularity | sketching

- **Draft-Sketch-Prove** [Jiang et al 2022]

  Draft informal proof
  Generate formal sketch
  Prove remaining gaps

  • **Neural**: informal proof + high-level *formal sketch*
  
  • **Symbolic tools**: formal prover & verifier
Neural natural prover

Statement: Prove that the fraction $\frac{2n+4}{14n+3}$ is irreducible for every natural number $n$.

Informal Proof (Minerva 62B):
We must show that $\gcd(21n + 4, 14n + 3) = 1$.
The Euclidean algorithm gives

\[
\begin{align*}
21n + 4 &= 1 \cdot (14n + 3) + 7n + 1 \\
14n + 3 &= 2 \cdot (7n + 1) + 1.
\end{align*}
\]

Since $\gcd(7n + 1, 1) = 1$, we have $\gcd(21n + 4, 14n + 3) = 1$.

Formal Proof:

```
theorem imo_1959_p1:
  fixes n :: nat
  shows "gcd (21n+4) (14n+3) = 1"
proof -
  (* The Euclidean algorithm gives
  21n+4=1\cdot(14n+3)+7n+1
  14n+3=2\cdot(7n+1)+1. *)
  have c0: "21n + 4 = 1 \cdot (14n + 3) + 7n + 1" 
    by auto [ATP]
  have c1: "14n + 3 = 2 \cdot (7n + 1) + 1" using c0
    by auto [ATP]
  (* Since \ \gcd(7n+1,1)=1, we have \ \gcd(21n+4,14n+3)=1. *)
  then have "gcd (7n + 1) 1 = 1"
    using c1
    by auto [ATP]
  then have "gcd (21n + 4) (14n + 3) = 1"
    using c1
    by (smt (z3) BitM_plus_one ab_semigroup_add_class.add_ac(1)
      add.assoc c0 gcd.commute gcd_add2 gcd_add_mult mult_numeral_1
      numeral_One numeral_eq_Suc numerals(1) semiring_norm(3)) [ATP]
  then show \thesis
    using c1
    by blast [ATP]
  qed
```

Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs
A. Jiang*, S. Welleck*, J. Zhou*, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, Y. Wu
Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs
A. Jiang*, S. Welleck*, J. Zhou*, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, Y. Wu
Draft, Sketch, and Prove: Guiding Formal Theorem Provers with Informal Proofs
A. Jiang*, S. Welleck*, J. Zhou*, T. Lacroix, J. Liu, W. Li, M. Jamnik, G. Lample, Y. Wu
Modularity | sketching

MiniF2F Problems Solved (out of 488)

- Human informal proof drafts
- Minerva (540B) proof drafts
- Minerva (62B) proof drafts
- Minerva (8B) proof drafts
- Codex proof drafts

#Autoformalization Attempts Per Problem

"Inference-time algorithm"
Modularity | inference

- **Maeutic Inference** [Jung et al 2022]:
  - Enumerate & score tree of rationales
  - Infer answer with MAX-Satisfiability
- **Modules & tools**: language model, scorer, verifier, MAX-SAT solver

---

**Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations**


*EMNLP 2022.*
Modularity | inference

- **Maeutic Inference** [Jung et al 2022]:
  - Performance (commonsense QA & fact verification)
  - Robustness

![Graph showing performance and robustness](image-url)
Modularity | inference

- **Maeutic Inference** [Jung et al 2022]:
  - Performance (commonsense QA & fact verification)
  - Robustness
Modularity | other examples

• Recursion & correction
  - e.g. Self-correction [Welleck et al 2022]

• General decompositions
  - e.g. Decomposed prompting [Khot et al 2022]

• Text as “protocol” for multiple modalities
  - e.g. Socratic models [Zeng et al 2022]

  • … many more! An exciting & expanding area
Modularity | Takeaways

• Multiple modules interacting through text
  • Formalism: graphical model / probabilistic program

• Intuition 1: Separation of concerns
  • High-level reasoning vs. low-level computation
  • Generation vs. retrieval & verification

• Intuition 2: Robust layer on top of a noisy enumerator
  • Neural: enumerate many solution candidates
  • Symbolic: verify, fill in gaps, resolve globally

• Many more ideas to explore here!
Overview

• Modularity
  • Single monolithic system → decomposed neural & symbolic modules

• Constraints
  • Discrete logical constraints

• Knowledge
  • Hand-crafted → generated and distilled
• Language models are difficult to control

What is the mass of Jupiter?

Generate a question containing all of the given words.
Words: Jupiter, Mercury, Venus, mass

missing keywords

Example from: https://beta.openai.com/playground
Constraints

- Language models are difficult to control
- Build **decoding algorithm** to enforce constraints.
Constraints

• Language models are difficult to control

• Build **decoding algorithm** to enforce constraints.

• **Lexical constraints**: words should or should not appear in the generation.
Constraints

- Language models are difficult to control
- Build **decoding algorithm** to enforce constraints.

  - **Lexical constraints**: words should or should not appear in the generation.

Table to Text

<table>
<thead>
<tr>
<th>type</th>
<th>hotel</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>182</td>
</tr>
<tr>
<td>dogs allowed</td>
<td>don't care</td>
</tr>
</tbody>
</table>

There are **182 hotels** if you do not care whether dogs are allowed.

Theorem Proving

**Theorem**: Let $x$ be an **even integer**. Then $x + 5$ is **odd**.

**Proof**: **Proof by Contradiction**: Aiming for a contradiction, suppose $x + 5$ is even. Then there exists an **integer** $k$ such that $x + 5 = 2k$.

... [Welleck et al 2022]
Constraints

- **NeuroLogic A*-Esque Decoding** [Lu et al 2022]
  - Lexical constraints expressed in Conjuctive Normal Form
  - A*-search-like lookahead

Logical Lexical Constraints

\[(Jupiter) \land (Mercury) \land (Venus) \land (mass \lor masses)\]
Write a sentence with: car ∧ drive ∧ snow

\[ \text{score} \quad s = \log \frac{P_y(y_i | y_{<i})}{m} \]

Off-the-Shelf GPT2

Beam Search

for

shop

during

at
Write a sentence with: **car** ∧ **drive** ∧ **snow**

Objective

A* heuristics: estimate the minimal cost to satisfy one more constraint

\[
\text{score } s = \log P_\theta(y_t | y_{<t}) + \alpha \sum_{i=1}^{m} C_i + \lambda_i \cdot \max_{D_i : D_i = 0} \log P_\theta(D_i | y_{<t+k})
\]

Constraints
Constraints

NeuroLogic A*-Esque Decoding [Lu et al 2022]

Write a sentence with: car \& drive \& snow

\[
\text{score } s = \log P(y_t | y_{<t}) + \alpha \sum_{i=1}^{m} C_i + \lambda_1 \max_{D_i: D_i=0} \log P(D_i | y_{<t+k})
\]

Constraints

A* Heuristic

I drive my car during the season.

- Summer: \(\max(P(\text{snow}))\)
- Winter: \(\max(P(\text{snow}))\)
- Day: \(\max(P(\text{snow}))\)
Human Evaluation Results

CommonGen

(Lin et al., 2020)

Supervised

GBS (Hokamp and Liu, 2017)
CBS (Anderson et al., 2017)
NeuroLogic (Lu et al., 2021)
NeuroLogic A*esque

Human Eval Score

2.01 2.27 2.54 2.66
Human Evaluation Results

CommonGen
(Lin et al., 2020)

Unsupervised

<table>
<thead>
<tr>
<th>Human Eval Score</th>
<th>TSMH (Zhang et al., 2020)</th>
<th>NeuroLogic (Lu et al., 2021)</th>
<th>NeuroLogic A*esque</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.85</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TSMH (Zhang et al., 2020)
NeuroLogic (Lu et al., 2021)
NeuroLogic A*esque
Constraints

- **Stepwise Stochastic Beam Search** [Welleck et al 2022]
  - Beam-search over arbitrary-length segments with a constraint value function.

- Theorem proving:
  - Segment: proof step
  - Constraints: theorems, definitions

---

**Theorem**

Let $x$ be an even integer.
Then $x + 5$ is odd.

**Reference Titles**
- Even Integer
- Odd Integer
- Integer
- Proof by Contradiction

**Generated Proof**

Proof by Contradiction: Aiming for a contradiction, suppose $x + 5$ is even.
Then there exists an integer $k$ such that $x + 5 = 2k$.
This contradicts the premise that $x$ is even.
Hence, by Proof by Contradiction, $x + 5$ is odd.

---

**NaturalProver: Grounded Mathematical Proof Generation with Language Models**
S. Welleck, J. Liu, X. Lu, H. Hajishirzi, Y. Choi
*NeurIPS 2022.*
Constraints

- **Stepwise Stochastic Beam Search**
  [Welleck et al 2022]

- Human evaluation (UW Mathematics students)

![Bar chart showing comparison between Vanilla GPT-3, NaturalProver, and +Constrained Decoding in terms of Correctness and Usefulness.]
Overview

- **Modularity**
  - Single monolithic system → decomposed neural & symbolic modules

- **Constraints**
  - Discrete logical constraints

- **Knowledge**
  - Hand-crafted → *generated* and *distilled*