FormalAlign: Automated Alignment Evaluation for Autoformalization

Jianqiao Lu*, Yingjia Wan*, Yinya Huang, Jing Xiong, Zhengying Liu, Zhijiang Guo

*Leading co-authors with equal contribution

Introduction

Autoformalization:

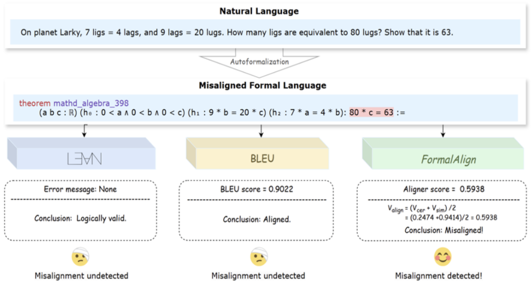
- The task of automatically converting informal math theorems and proofs into machine-verifiable formal languages.
- Promising direction for developing LLM reasoning and formal verification.

Challenge:

- Evaluating semantic alignment between the autoformalization input and output is difficult. The lack of effective evaluation methods hinders the development of robust autoformalization models.

Prior Evaluation Methods:

- *Automated Formal Language Compiler :* focus solely on **logical validity** of formal output
- *BLEU Score :* struggles with **semantic alignment/logical equivalence** between informal input and formal output
- *Manual Verification :* expensive, labor-intensive, and not scalable



Evaluator Framework

FormalAlign: The first method for automatically evaluating semantic alignment between informal and formal languages in autoformalization.

Training: $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{CL}$

1 **Autoformalization Task :**
$$\mathcal{L}_{CE} = -\sum_{j=1}^{n} \log P_\phi(\mathbf{FL}_{i,j} \mid \mathbf{FL}_{i,j'|j'<j}, \mathbf{NL}_i)$$

cross-entropy loss: minimizes the error in predicting formalized output.

2 **Alignment Task :**
$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\cos(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathbf{u}_i, \mathbf{v}_j)/\tau)}$$

contrastive loss: encourages the cosine similarity between the hidden states of aligned informal-formal pairs (u_i, v_i) to be higher than that of non-aligned pairs (u_i, v_i') .

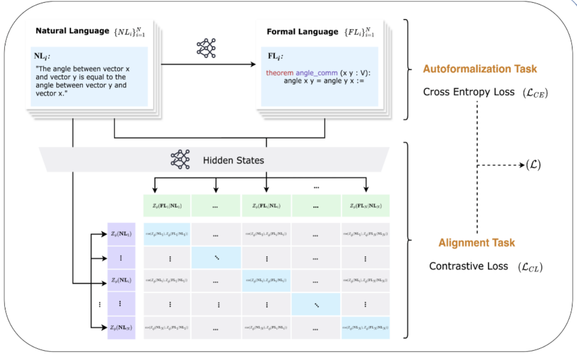
Inference: $\mathcal{V}_{align} = (\mathcal{V}_{cer} + \mathcal{V}_{sim}) / 2$

1 **Certainty Score $\mathcal{V}_{cer}$:**
$$\mathcal{V}_{cer} = \exp\left(\frac{1}{n}\sum_{j=1}^{n} \log P_\phi(\mathbf{FL}_{i,j} \mid \mathbf{FL}_{i,<j}, \mathbf{NL}_i)\right)$$

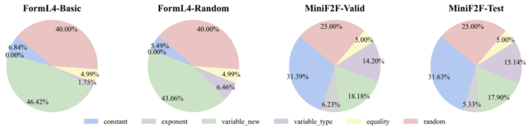
measures the trained model's confidence in predicting the formal output sequence.

2 **Similarity Score $\mathcal{V}_{sim}$:**
$$\mathcal{V}_{sim} = \cos(Z_\phi(\mathbf{NL}_i), Z_\phi(\mathbf{FL}_i \mid \mathbf{NL}_i))$$

measures alignment between the embedding representations of the informal input and the formal output.



Distribution of Misalignment Types



Evaluation

Misalignment Construction Strategies

Constant Modification (constant)
This type of misalignment involves changing a constant value within the expression.

```
theorem mathd_algebra_478
 (b h v : ℝ)
 (h_0 : 0 < b ∧0 < h ∧0 < v)
 (h_1 : v = 1 / 3 * (b * h))
 (h_2 : b = 31) -- changed constant
 (h_3 : h = 13 / 2) :
 v = 65 :=
```

Exponent Modification (exponent)
This misalignment targets the exponents in the expression.

```
theorem mathd_algebra_478
  (b h v : ℝ)
  (h_0 : 0 < b ∧0 < h ∧0 < v)
  (h_1 : v = 1 / 3 * (b^2 * h)) --
     changed exponent
  (h_2 : b = 30)
  (h_3 : h = 13 / 2) :
  v = 65 :=
```

Introduction of a New Variable (variable_new)
This misalignment introduces a completely new variable into the expression.

```
theorem mathd_algebra_478
  (b h v x : ℝ) -- added a new
     variable x
  (h_0 : 0 < b ∧0 < h ∧0 < v)
  (h_1 : v = 1 / 3 * (b * h))
  (h_2 : b = 30)
  (h_3 : h = 13 / 2) :
  v = 65 :=
```

Modification of Equality (equality)
This misalignment switches between equality = and inequality ≠ symbols within the expression.

```
theorem mathd_algebra_478
  (b h v : ℝ)
  (h_0 : 0 < b ∧0 < h ∧0 < v)
  (h_1 : v ≠1 / 3 * (b * h)) --
     swapped inequality
  (h_2 : b = 30)
  (h_3 : h = 13 / 2) :
  v = 65 :=
```

Random Pairing (random)
This creates a mismatch between the informal input and its formal output. Instead of pairing the informal input with its correct formal output, this strategy randomly selects a formal output from other examples.

Change of Variable Type (variable_type)

In this case, the misalignment involves changing the type of a variable within the expression. The function identifies the type of a randomly selected variable and changes it to a different type from a predefined list of types.

```
theorem mathd_algebra_478
(b h v : ℤ) -- changed type to ℤ
(h₀ : 0 < b ∧0 < h ∧0 < v)
(h₁ : v = 1/3 * (b * h))
(h₂ : b = 30)
(h₃ : h = 13/2) :
v = 65 :=
```

Evaluation Metrics

- **Alignment Selection (AS): $\mathbf{V}_{align}$,** i.e., how well an evaluator selects the aligned formal output from multiple candidates when given an informal input.
- **Alignment Detection:** We set a predefined threshold θ to detect alignment:
  - $\mathbf{V}_{align} \geq \theta$: the evaluator detects alignment;
  - $\mathbf{V}_{align} < \theta$: the evaluator detects misalignment.

Results

Automated Alignment Evaluation

Datasets	FormL4-Basic			FormL4-Random			MiniF2F-Valid			MiniF2F-Test		
	AS	Prec.	Rec.	AS	Prec.	Rec.	AS	Prec.	Rec.	AS	Prec.	Rec.
GPT-4	90.23	42.68	88.15	**91.85**	45.72	**89.95**	67.24	59.85	**89.87**	**70.82**	62.45	**92.88**
GPT-3.5	50.23	25.21	**90.83**	47.00	23.42	67.26	47.32	22.29	62.55	40.74	21.97	61.73
FormalAlign	**99.21**	**93.65**	86.43	85.85	**86.90**	89.20	**66.39**	**68.58**	60.66	64.61	**66.70**	63.37

Ablation of Backbones (AS)

Datasets	FormL4		MiniF2F	
	Basic	Random	Valid	Test
Phi	80.77	71.07	31.56	32.51
DeepSeek	90.29	77.08	54.66	55.19
LLaMA	98.08	76.42	54.51	57.20
Mistral	99.21	85.85	66.39	66.70

Ablation of Training Losses (AS)

Datasets	FormL4		MiniF2F	
	Basic	Random	Valid	Test
w/ cer	98.98	85.64	53.69	55.55
w/ sim	45.25	20.75	20.49	21.81
Ours	**99.21**	**85.85**	**66.39**	**66.70**