HANYANG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

CLAWS: Creativity detection for LLM-generated solutions using Attention Window of Sections

Keuntae Kim*, Eunhye Jeong*, Sehyeon Lee, Seohee Yoon, Yong Suk Choi†

Hanyang University, Seoul, Korea

(*: Equal contribution, †: Corresponding author)

HYU Artificial Intelligence Laboratory

Contact us: ktkpv94@hanyang.ac.kr, jeh0826@hanyang.ac.kr

Introduction

Background

- Recent **LLM progress is most notable in reasoning ability**, especially mathematical problem solving
- Latest frontier **LLMs are approaching human-level intelligence** in **mathematical problem solving ability**

Motivation

- **Human intelligence is not defined solely by accuracy**; It also includes diverse aspects such as **Creativity**
- There is a **need** to extend from **Hallucination detection to Creativity detection**

Introduction

Experimental Framework

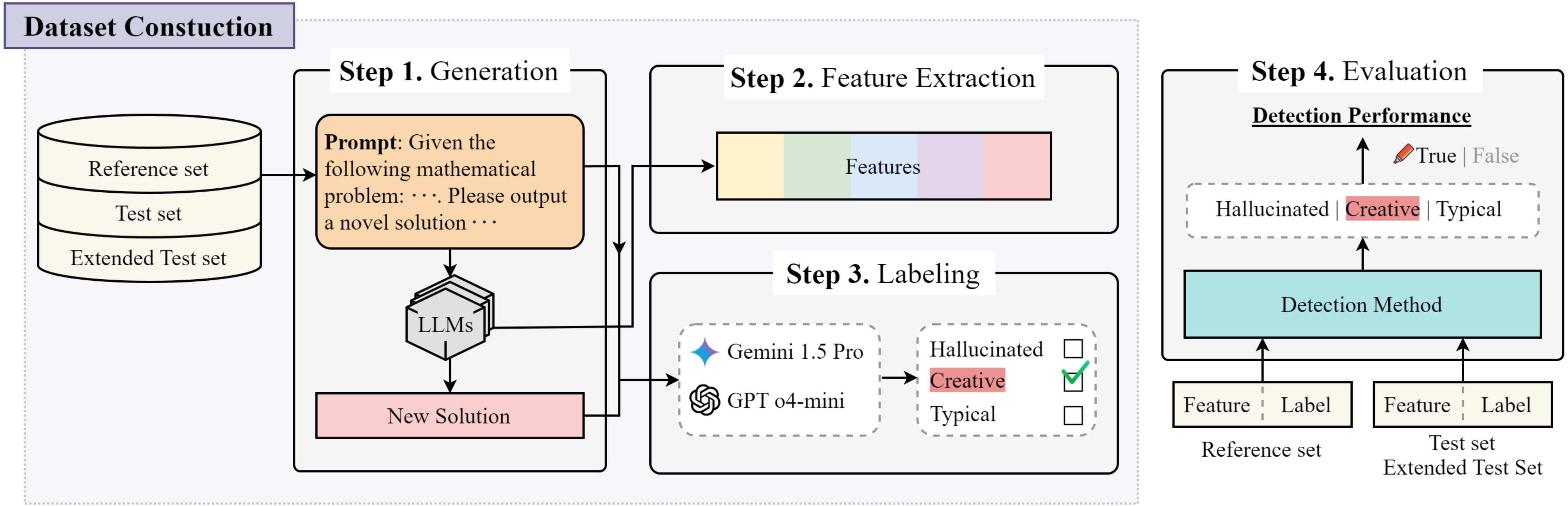
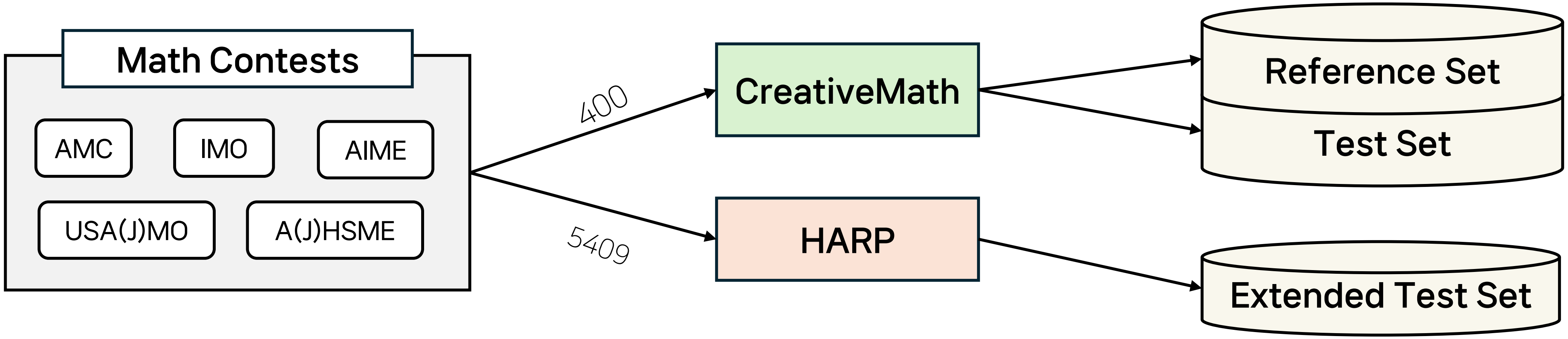


Figure 1. Overview of the experimental framework.

Proposed Method – CLAWS

Dataset Construction - Preparing Input Data for LLM Generation



- **Reference Set :** A set of 29 problems in CreativeMath [1]
- **Test Set :** A set of 371 problems in CreativeMath [1]
- **Extended Test Set :** A set of 4545 problems in HARP [2] (excluding overlaps with CreativeMath)

[1] Assessing the Creativity of LLMs in Proposing Novel Solutions to Mathematical Problems, AAAI 2025
[2] HARP: A challenging human-annotated math reasoning benchmark, 2024

Proposed Method – CLAWS

Dataset Construction - Step 1: Generation

- To generate mathematical problem solutions, we select five Reasoning Language Models
  - DeepSeek-Math-7B-RL, Qwen2.5-Math-7B-Inst, Mathstral-7B, OpenMath2-Llama3.1-8B, and OREAL-7B

Model	DeepSeek				Mathstral				OpenMath2				OREAL				Qwen-2.5			
Dataset	Ha	Cr	Ty	Total	Ha	Cr	Ty	Total	Ha	Cr	Ty	Total	Ha	Cr	Ty	Total	Ha	Cr	Ty	Total
REF	868	206	649	1723	1192	175	437	1804	923	103	785	1811	1244	83	379	1706	631	324	752	1707
TEST	798	160	456	1414	961	154	337	1452	815	97	551	1463	932	89	369	1390	578	203	579	1360
AMC	1197	530	1373	3100	1679	434	1049	3180	1330	291	1578	3199	1928	237	935	3100	637	629	1784	3050
AIME	772	126	262	1160	917	68	221	1206	644	67	501	1212	911	47	161	1119	529	159	373	1061
A(J)HSME	657	424	763	1844	945	354	606	1905	723	248	943	1914	1005	161	656	1822	281	491	1054	1826

Table 1. Overview of the generation results. Number of samples per class (Hallucinated, Creative, Typical) for each dataset and model.

Proposed Method – CLAWS

Dataset Construction - Step 1: Generation

- **Split** each input prompt and its corresponding output **into five predefined semantic sections**:
  - (Input prompt) Guideline | Problem | Solution | Instruction
  - (Output) Response
- These sections are **used for section-wise attention analysis**

Criteria for evaluating the difference between two mathematical solutions include:
i). If the methods used to arrive at the solutions are fundamentally different, such as algebraic manipulation versus geometric reasoning, they can be considered distinct;
ii). Even if the final results are the same, if the intermediate steps or processes involved in reaching those solutions vary significantly, the solutions can be considered different;
iii). If two solutions rely on different assumptions or conditions, they are likely to be distinct;
iv). A solution might generalize to a broader class of problems, while another solution might be specific to certain conditions. In such cases, they are considered distinct;
v). If one solution is significantly simpler or more complex than the other, they can be regarded as essentially different, even if they lead to the same result.

Given the following mathematical problem:
What is the largest power of 2 that is a divisor of $13^4 - 11^4$?

And some typical solutions:
1. First, we use the difference of squares on $13^4 - 11^4 = (12^2)^2 - (11^2)^2. \cdots$
2. Just like in the above solution, we use the difference-of-squares factorization, but only once to get $13^4 - 11^4 = (13^2 - 11^2)(13^2 + 11^2). \cdots$

Please output a novel solution distinct from the given ones for this math problem.

Figure 2. Input prompt consists of four sections: Guideline (G, yellow), Problem (P, green), Reference Solutions (S, blue), and Instruction (I, purple).

Proposed Method – CLAWS

Dataset Construction - Step 2: Feature Extraction

- Computes attention weights for each section during generation
- Quantifies **how much influence each section exerts on the response**

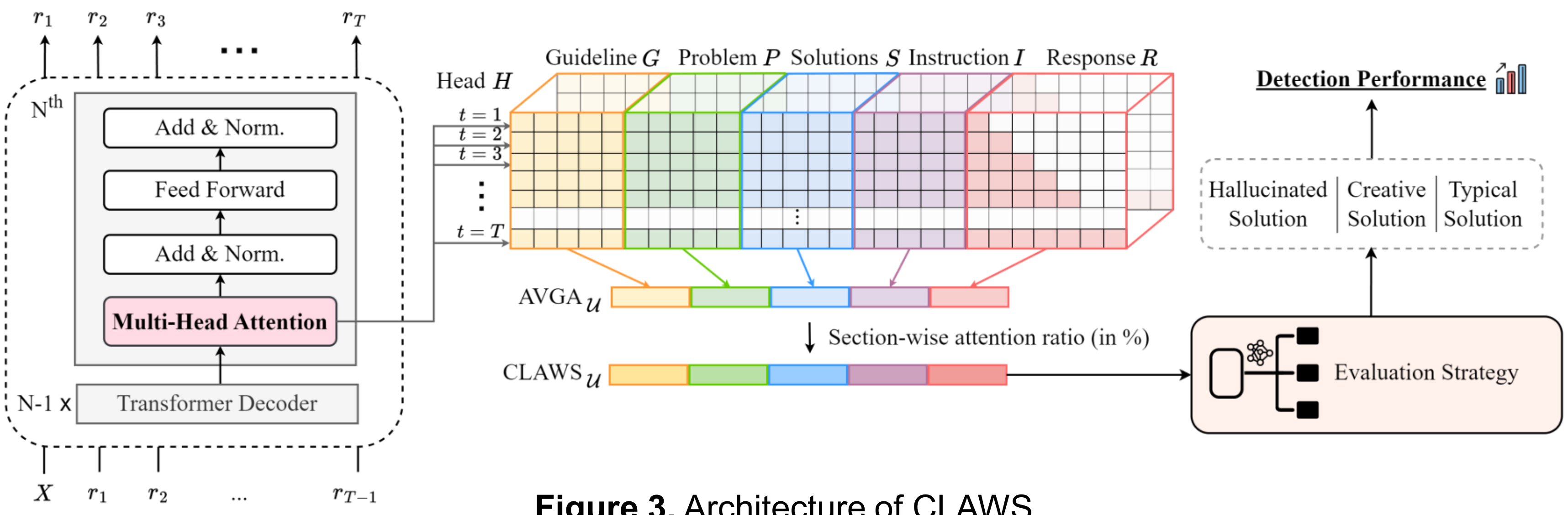


Figure 3. Architecture of CLAWS

Proposed Method – CLAWS

Dataset Construction - Step 2: Feature Extraction

① Compute average attention for each section $\mathcal{U}$ across all heads H and time steps T, section-specific token positions $\mathcal{I}_{\mathcal{U}}$

$$\mathrm{AVGA}_{\mathcal{U}} = \frac{1}{H \cdot T \cdot |\mathcal{I}_{\mathcal{U}}|} \sum_{h=1}^{H} \sum_{t=1}^{T} \sum_{i \in \mathcal{I}_{\mathcal{U}}} A_h^{(L)}[t, i], \ \text{for } \mathcal{U} = \{G, P, S, I, R\}$$

② Normalize the averaged attention values as

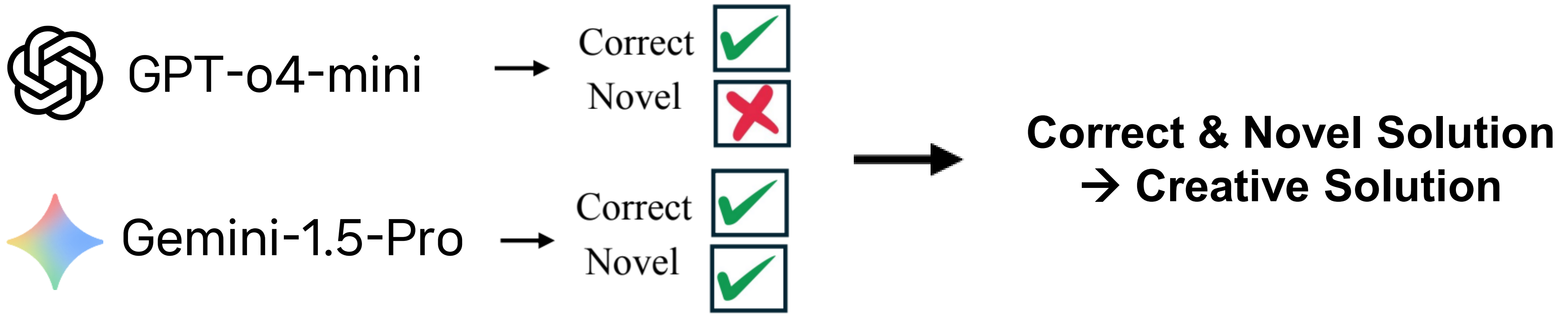
$$\mathrm{CLAWS}_{\mathcal{U}} = \frac{\mathrm{AVGA}_{\mathcal{U}}}{\sum_{\mathcal{U}' \in \{G, P, S, I, R\}} \mathrm{AVGA}_{\mathcal{U}'}}$$

so that the sum of all $\mathrm{CLAWS}_{\mathcal{U}}$ equals 1, **representing the relative contribution of each section to the model's reasoning pattern**

Proposed Method – CLAWS

Dataset Construction - Step 3: Labeling

To assess model generations, we employ two frontier-level LLMs – **GPT-o4-mini and Gemini-1.5-Pro** – as LLM Evaluators.



① **Evaluate 'Correctness'** - Solutions that both evaluators judged as 'Correctness', those that neither evaluator judged as 'Creativity'.

② **Evaluate 'Novelty (Creativity)'** - Solutions that both evaluators judged to be 'Correctness', if even one evaluator judged them to be 'Creativity'. This is a criterion for inclusive acceptance of Creativity.

Proposed Method – CLAWS

Step 4: Evaluation

To address **various models, imbalanced datasets, and multiple baselines**, sufficient **Evaluation Strategies** and **Metrics** are established.

Evaluation Strategies

- Applied five strategies (**Threshold, Prototype, XGBoost, MLP, TabM**) to evaluate performance using features

Evaluation Metrics

- **Weighted F1 ($F1_w$), Macro F1 ($F1_m$), AUROC, Macro AP (AP_m)** were used to evaluate performance for imbalanced datasets.

Proposed Method – CLAWS

Step 4: Evaluation

Baselines

- PPL (Perplexity), LE (Logit Entropy), WE (Window Logit Entropy), HS (Hidden Score), AS (Attention Score) [3]

Datasets

- Each method was evaluated using the dataset generated through Steps 1 to 3

① Three-class Dataset

- Imbalanced
- Balanced
  - random sampling with an equal number of samples per class

② Two-class Dataset

- Non-Hallucinated (Typical + Creative) & Hallucinated

[3] LLM-Check: Investigating Detection of Hallucinations in Large Language Models, NeurIPS 2024

Results & Analysis

For Creativity Detection

- **Imbalanced Dataset**
- Evaluation strategies
  - **Threshold (for PPL, WE, LE, HS, AS)**
  - **Prototype(for CLAWS).**
- **CLAWS** Outperformed all models on the all dataset across all four metrics, **achieving superior creativity detection performance compared with five white-box baselines**.

Table 2. Results for Creativity detection. Bold values indicate the best performance, underlined values denote the second best, and gray-shaded cells correspond to cases where the model detected only two out of the three classes.

Dataset		TEST				AMC				AIME				A(J)HSME			
Model	Method	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC
Deepseek	PPL	48.09	35.77	37.07	56.49	44.56	35.63	36.28	55.12	55.93	36.70	**36.59**	**56.63**	**42.34**	36.52	**37.49**	**56.82**
	WE	18.59	23.20	35.89	53.89	27.52	26.03	34.67	52.26	9.07	16.36	33.67	50.31	28.72	27.70	34.44	51.92
	LE	40.56	33.21	34.00	50.90	35.69	32.96	33.42	50.31	28.99	25.28	33.27	48.89	34.89	33.46	33.45	50.18
	HS	29.56	25.18	32.40	45.03	38.44	32.96	33.60	50.22	38.30	29.65	33.71	50.67	38.61	35.80	34.54	52.40
	AS	33.95	24.99	30.98	42.80	33.51	29.18	33.43	50.19	43.92	32.89	33.58	50.97	28.63	26.83	33.19	49.70
	CLAWS	**58.66**	**46.01**	**41.17**	**62.09**	**46.71**	**40.99**	**37.16**	**56.40**	**56.90**	**38.12**	35.38	54.47	38.82	**37.64**	36.25	54.40
Mathstral	PPL	42.45	25.94	31.37	43.26	36.50	25.21	31.81	45.89	56.90	29.97	32.76	47.49	34.79	25.58	32.58	48.15
	WE	46.19	28.89	32.58	46.68	40.71	30.02	32.73	48.44	52.20	30.20	32.91	48.33	40.34	31.79	33.86	51.05
	LE	41.62	28.17	32.11	45.66	35.20	29.55	32.05	46.93	44.77	28.56	33.47	50.50	35.46	30.56	32.40	47.77
	HS	49.86	26.53	32.49	47.07	37.37	23.46	33.33	49.97	**65.96**	31.13	33.46	50.23	33.42	22.65	33.42	50.14
	AS	38.41	24.50	31.23	42.22	36.92	27.53	32.02	46.69	57.35	31.95	33.51	49.82	35.26	27.57	32.41	47.60
	CLAWS	**63.20**	**46.05**	**41.75**	**63.70**	**51.47**	**41.45**	**37.89**	**57.69**	65.25	**36.05**	**34.43**	**52.73**	**49.13**	**42.29**	**38.20**	**58.18**
OpenMath2	PPL	36.47	27.52	32.72	47.30	41.10	31.45	33.12	49.24	40.44	30.49	32.18	47.57	39.22	30.05	33.13	48.56
	WE	40.89	32.14	33.84	50.50	43.44	34.48	33.93	51.19	40.55	31.17	33.37	50.00	42.45	34.16	33.99	51.08
	LE	47.48	35.96	35.15	53.15	43.17	36.18	34.28	52.62	41.82	33.10	34.32	**52.92**	42.38	37.55	34.70	53.28
	HS	30.48	23.20	30.77	41.57	33.02	26.78	31.17	44.52	40.45	32.09	32.63	49.34	31.62	26.55	31.37	44.93
	AS	33.20	24.48	30.65	42.17	32.84	27.77	31.89	46.75	40.42	30.96	**48.59**	32.59	31.03	27.53	32.09	47.04
	CLAWS	**60.86**	**44.27**	**40.77**	**60.66**	**54.32**	**42.12**	**38.53**	**58.06**	**49.35**	**34.41**	35.35	52.00	**50.88**	**41.36**	**37.73**	**57.22**
OREAL	PPL	46.52	27.81	31.78	45.60	41.68	26.38	31.25	44.27	55.96	28.11	32.64	47.36	36.90	24.83	31.03	43.62
	WE	49.57	27.39	32.80	48.26	44.87	27.32	32.87	48.82	66.65	32.63	33.48	51.28	36.37	24.79	32.79	48.44
	LE	**55.39**	36.15	34.46	53.11	**49.80**	**35.95**	34.43	53.30	63.53	**33.86**	**34.02**	**53.06**	41.06	31.86	33.29	50.47
	HS	51.95	29.46	32.60	47.90	48.58	28.28	33.20	49.60	**68.10**	31.63	33.30	49.22	**41.65**	28.36	33.12	49.62
	AS	45.56	28.24	31.83	45.56	47.92	29.11	32.70	48.25	65.19	32.74	33.20	49.56	40.18	26.41	32.74	48.48
	CLAWS	54.19	**40.18**	**38.15**	**59.46**	43.83	34.77	**35.57**	**54.78**	59.95	32.74	33.81	51.55	35.70	**31.93**	**35.51**	**54.41**
Qwen-2.5	PPL	25.66	23.30	31.76	42.62	26.40	21.39	31.71	43.31	28.29	25.29	32.00	44.52	24.88	20.34	32.09	44.79
	WE	30.79	29.40	34.71	52.50	22.04	26.04	33.80	51.15	33.23	29.08	33.12	49.24	20.50	23.61	33.12	49.51
	LE	**50.81**	**45.29**	39.50	**59.80**	45.86	40.18	36.15	55.81	**43.20**	**39.01**	**36.40**	**55.83**	45.70	**38.64**	35.23	54.09
	HS	30.67	28.31	36.25	54.53	47.57	31.98	34.81	52.98	20.37	24.17	34.57	52.83	**48.52**	32.54	34.78	52.77
	AS	30.75	26.96	32.05	45.53	38.61	31.55	32.93	48.78	37.32	33.71	33.92	51.42	33.72	28.55	32.86	48.35
	CLAWS	50.35	43.37	**39.88**	59.32	**52.77**	**41.39**	**37.45**	**57.59**	39.05	32.31	33.08	49.31	47.90	36.04	**35.86**	**54.94**

Results & Analysis

For Creativity Detection

- **Imbalanced Dataset**
- Evaluation strategies
  - **XGBoost, MLP, TabM** (for PPL, WE, LE, HS, AS, CLAWS)
- **Most baselines failed to detect three classes.**

Table 3. Results for Creativity detection. Bold values indicate the best performance, underlined values denote the second best, and gray-shaded cells correspond to cases where the model detected only two out of the three classes.

Dataset		TEST				AMC				AIME				A(J)HSME			
Strategy	Method	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC
XGBOOST	PPL	48.70	41.51	41.49	59.98	46.82	37.65	36.90	54.52	44.53	36.78	37.82	55.39	44.21	34.63	35.26	53.28
	WE	49.96	39.15	43.47	63.00	49.93	36.57	38.63	57.69	41.13	32.17	36.10	52.48	45.39	32.74	36.82	55.51
	LE	38.22	32.42	33.83	50.94	40.59	31.06	34.20	51.68	35.19	28.98	33.38	49.97	37.81	28.36	34.17	50.89
	HS	46.98	39.95	41.80	57.98	48.68	37.94	39.70	58.48	41.31	33.40	**50.87**	34.65	44.80	33.46	36.41	54.44
	AS	42.60	35.24	37.91	55.60	42.92	32.76	35.51	53.54	38.85	30.92	33.74	49.56	40.00	29.96	34.70	52.49
	CLAWS	**52.35**	**43.33**	**47.72**	**65.98**	**50.30**	**38.98**	**40.66**	**61.11**	**45.18**	**38.95**	39.75	**55.86**	**47.54**	**36.02**	**39.41**	**59.45**
MLP	PPL	54.27	**46.34**	47.90	67.59	50.01	38.01	**43.58**	61.96	**47.58**	36.13	**41.91**	**62.02**	44.92	37.00	38.23	57.72
	WE	45.67	39.58	38.31	54.97	43.32	32.95	36.49	53.87	40.26	30.78	34.22	51.26	42.11	33.31	34.35	50.53
	LE	29.48	23.10	28.27	41.43	36.53	26.33	33.96	49.47	31.62	25.13	31.19	47.00	32.30	23.38	35.14	50.73
	HS	**54.58**	42.77	46.98	65.86	50.29	36.64	41.90	60.61	43.11	32.96	36.01	52.38	46.24	36.21	37.80	56.01
	AS	43.80	38.05	40.62	59.13	43.04	32.73	35.52	52.88	39.07	30.08	32.98	47.71	39.86	30.64	34.43	51.33
	CLAWS	53.78	45.32	**50.44**	**67.98**	**52.51**	**39.65**	43.06	**63.20**	42.76	**36.59**	39.32	56.49	**53.75**	**41.05**	**42.29**	**62.61**
TabM	PPL	**54.53**	42.74	**47.13**	**65.44**	**50.83**	38.45	**43.07**	60.61	**48.30**	37.01	**43.37**	**60.80**	45.58	34.28	38.54	57.10
	WE	48.15	37.73	40.35	58.70	46.62	33.89	36.59	55.06	38.66	30.12	34.82	51.49	45.58	31.17	34.81	52.06
	LE	33.24	26.04	32.55	50.24	41.43	28.20	34.61	52.39	31.50	25.36	33.98	50.30	38.81	25.93	35.89	52.41
	HS	51.47	41.39	45.50	63.64	50.08	36.79	41.58	**61.00**	43.75	34.48	36.73	53.74	46.01	33.00	37.82	55.41
	AS	45.76	35.86	38.65	56.76	43.53	31.74	36.03	54.26	37.84	29.68	33.06	49.16	41.37	28.96	35.60	54.28
	CLAWS	51.93	**42.92**	45.82	63.83	48.69	**38.90**	39.23	59.61	45.14	**38.28**	38.15	55.26	**47.31**	**35.97**	**39.23**	**59.28**

Results & Analysis

For Creativity Detection

- Balanced dataset
- **CLAWS** outperformed all models on the all dataset across all four metrics, achieving **superior creativity detection performance compared with five white-box baselines**.

Table 4. Results for Creativity detection on the balanced dataset. Bold values indicate the best performance, underlined values denote the second best, and gray-shaded cells correspond to cases where the model detected only two out of the three classes.

Dataset		TEST				AMC				AIME				A(J)HSME			
Model	Method	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC
Deepseek	PPL	35.22	35.22	34.40	52.03	**37.57**	**37.57**	35.04	53.16	**42.57**	**42.57**	**37.65**	**57.34**	36.21	36.21	34.83	52.30
	WE	30.54	30.54	35.34	53.44	28.46	28.46	34.20	51.60	25.80	25.80	33.15	49.40	28.70	28.70	34.35	51.95
	LE	31.89	31.89	32.98	48.91	32.78	32.78	33.35	49.95	29.32	29.32	32.65	47.62	33.06	33.06	33.37	49.94
	HS	27.36	27.36	32.21	45.16	31.40	31.40	33.25	49.58	32.09	32.09	33.14	49.21	32.88	32.88	33.58	50.53
	AS	31.48	31.48	33.81	49.69	32.11	32.11	33.31	49.91	30.56	30.56	33.56	50.40	33.29	33.29	33.55	50.47
	CLAWS	**46.30**	**46.30**	**41.34**	**62.03**	35.90	35.90	**35.87**	**54.62**	36.93	36.93	36.66	55.95	**36.43**	**36.43**	**35.97**	**54.89**
Mathstral	PPL	29.11	29.11	32.86	48.70	32.37	32.37	33.39	49.88	27.67	27.67	32.46	47.79	31.55	31.55	33.04	49.29
	WE	38.76	38.76	36.14	54.71	36.23	36.23	35.05	53.11	**34.21**	**34.21**	33.81	50.74	34.19	34.19	34.41	52.05
	LE	30.60	30.60	32.83	48.54	32.05	32.05	32.99	49.08	26.00	26.00	31.56	44.85	30.59	30.59	32.62	48.09
	HS	27.80	27.80	33.82	49.35	26.85	26.85	33.12	48.96	19.96	19.96	31.70	45.22	25.67	25.67	32.73	48.38
	AS	24.86	24.86	31.88	44.16	28.19	28.19	32.16	46.26	28.62	28.62	32.91	48.53	29.09	29.09	32.16	46.82
	CLAWS	**42.50**	**42.50**	**40.40**	**60.71**	**38.13**	**38.13**	**37.08**	**56.45**	31.86	31.86	**34.23**	**51.84**	**38.04**	**38.04**	**37.05**	**56.43**
OpenMath2	PPL	29.78	29.78	33.01	49.23	27.45	27.45	32.15	46.56	25.40	25.40	31.73	44.40	23.79	23.79	31.37	43.75
	WE	33.85	33.85	34.18	51.55	33.45	33.45	34.29	51.89	31.01	31.01	32.73	48.51	29.53	29.53	33.14	49.50
	LE	36.34	36.34	34.53	52.32	**40.00**	**40.00**	36.16	55.15	31.18	31.18	33.42	50.00	**38.06**	**38.06**	35.53	53.93
	HS	25.49	25.49	31.53	44.33	28.34	28.34	32.25	46.91	36.30	36.30	34.92	52.61	28.43	28.43	32.30	47.38
	AS	23.92	23.92	31.19	43.04	29.84	29.84	32.75	48.20	**38.59**	**38.59**	**35.52**	**54.10**	32.32	32.32	33.77	50.30
	CLAWS	**41.90**	**41.90**	**38.92**	**58.51**	37.66	37.66	**36.93**	**56.36**	24.86	24.86	33.22	49.63	33.47	33.47	**35.60**	**54.23**
OREAL	PPL	29.02	29.02	32.41	47.47	23.55	23.55	31.56	44.09	31.64	31.64	33.65	50.00	23.87	23.87	31.48	44.25
	WE	25.69	25.69	32.14	46.91	27.60	27.60	33.08	49.37	30.21	30.21	33.34	50.00	27.38	27.38	33.00	49.07
	LE	**33.34**	**33.34**	**33.86**	**50.84**	**34.64**	**34.64**	34.96	53.16	29.33	29.33	33.37	49.47	35.79	35.79	35.33	53.88
	HS	30.03	30.03	32.87	48.60	26.77	26.77	31.91	45.89	33.93	33.93	33.76	50.53	27.64	27.64	32.10	46.58
	AS	26.10	26.10	33.10	47.75	31.65	31.65	34.15	51.58	25.07	25.07	33.61	50.53	30.21	30.21	33.59	49.22
	CLAWS	25.27	25.27	32.99	48.31	34.08	34.08	**35.16**	**53.48**	**34.85**	**34.85**	**34.69**	**52.66**	**37.49**	**37.49**	**35.52**	**54.04**
Qwen-2.5	PPL	27.04	27.04	34.14	50.00	27.75	27.75	33.72	49.52	25.76	25.76	33.34	49.53	35.59	31.09	33.92	50.47
	WE	34.62	34.62	34.91	52.96	32.83	32.83	34.20	51.39	31.12	31.12	33.01	49.06	29.08	28.56	33.01	49.11
	LE	**45.25**	**45.25**	**39.53**	**59.24**	40.54	40.54	36.60	55.60	**39.56**	**39.56**	**36.05**	**54.56**	**41.31**	**39.10**	**35.66**	**54.32**
	HS	27.84	27.84	34.67	51.72	30.39	30.39	35.66	53.58	18.97	18.97	33.48	50.31	36.80	31.55	35.08	53.11
	AS	26.88	26.88	32.66	47.17	31.59	31.59	34.13	51.27	32.32	32.32	34.20	51.73	37.59	34.97	34.28	51.52
	CLAWS	31.34	31.34	33.39	49.63	**40.88**	**40.88**	**38.04**	**57.63**	23.76	23.76	31.66	45.28	38.27	36.63	35.56	53.95

Results & Analysis

For Hallucination Detection

- **Two-class Dataset** (Non-hallucinated / hallucinated)
- Evaluation strategies
  - **Threshold (for PPL, WE, LE, HS, AS)**
  - **Prototype(for CLAWS).**
- **CLAWS consistently achieved the best performance across all evaluation metrics and models.**

Table 5. Results for Hallucination detection. Bold values indicate the best performance, underlined values denote the second best, and gray-shaded cells correspond to cases where the model detected only single out of the two classes.

		TEST				AMC				AIME				A(J)HSME			
Model	**Method**	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC	$F1_w$	$F1_m$	AP_m	AUROC
Deepseek	PPL	37.63	40.32	45.13	53.06	52.18	45.38	62.33	51.95	39.75	42.77	36.16	55.60	55.31	46.20	65.42	52.25
	WE	26.44	30.34	43.56	50.00	46.70	38.04	61.39	50.00	16.77	25.06	33.45	50.00	50.42	39.16	64.37	50.00
	LE	27.66	31.46	43.81	50.50	46.82	38.19	61.40	50.03	17.22	25.35	33.36	49.81	50.37	39.12	64.33	49.92
	HS	26.41	30.31	43.52	49.92	47.11	38.56	61.48	50.20	17.64	25.74	33.54	50.19	50.61	39.51	64.34	49.93
	AS	26.60	30.45	43.43	49.72	46.68	38.02	61.37	49.97	16.77	25.06	33.45	50.00	50.59	39.43	64.38	50.03
	CLAWS	**67.46**	**67.24**	**55.73**	**67.78**	**61.77**	**59.79**	**66.64**	**59.84**	**61.95**	**58.17**	**38.45**	**58.68**	**64.93**	**61.67**	**70.38**	**61.62**
Mathstral	PPL	17.09	25.27	33.82	50.00	29.62	31.76	46.58	49.90	9.44	19.45	23.98	50.05	33.84	33.58	50.37	49.95
	WE	17.09	25.27	33.82	50.00	29.71	31.84	46.62	49.96	9.61	19.57	24.00	50.11	33.89	33.63	50.42	50.05
	LE	17.36	25.47	33.82	50.00	29.65	31.79	46.62	49.97	10.12	19.94	24.06	50.27	34.24	33.98	50.50	50.21
	HS	17.03	25.08	33.48	49.24	30.11	32.18	46.36	49.44	9.26	19.33	23.96	50.00	33.96	33.70	50.39	50.00
	AS	17.09	25.27	33.82	50.00	29.66	31.80	46.64	50.00	9.26	19.33	23.96	50.00	33.77	33.51	50.39	50.00
	CLAWS	**72.99**	**69.59**	**49.42**	**69.30**	**63.97**	**63.90**	**55.69**	**64.04**	**65.62**	**50.26**	**24.19**	**50.59**	**61.05**	**61.04**	**57.10**	**61.04**
OpenMath2	PPL	31.78	34.68	44.16	49.74	45.53	40.17	58.19	49.51	34.68	36.15	45.52	47.20	49.49	41.63	61.65	48.76
	WE	27.19	30.70	44.29	50.00	43.09	36.88	58.42	50.00	29.91	31.91	46.86	50.00	47.74	38.36	62.23	50.00
	LE	28.14	31.55	44.39	50.20	43.18	36.99	58.41	49.97	30.38	32.35	46.89	50.06	48.09	38.81	62.32	50.21
	HS	27.55	31.01	44.27	49.95	43.40	37.27	58.43	50.01	31.30	33.21	46.95	50.17	47.68	38.37	62.12	49.78
	AS	27.46	30.94	44.32	50.05	43.30	37.16	58.39	49.92	30.09	32.08	46.90	50.08	48.14	38.90	62.30	50.15
	CLAWS	**64.91**	**64.70**	**54.19**	**65.05**	**62.53**	**61.65**	**65.19**	**61.82**	**58.10**	**57.50**	**52.32**	**58.47**	**63.88**	**61.81**	**68.70**	**61.96**
OREAL	PPL	16.96	25.09	32.61	49.23	21.37	27.87	37.65	49.67	6.79	16.08	18.33	49.13	28.09	31.19	44.51	49.32
	WE	23.20	29.92	33.39	50.99	22.04	28.48	37.93	50.26	10.95	18.75	18.58	49.96	28.10	31.27	44.88	50.09
	LE	20.75	27.91	32.84	49.75	23.17	29.34	37.87	50.14	10.03	18.33	18.77	50.60	30.50	33.45	45.22	50.75
	HS	16.33	24.78	32.95	50.00	20.69	27.37	37.73	49.83	6.01	15.80	18.60	50.05	27.74	30.93	44.81	49.94
	AS	16.47	24.78	32.71	49.45	20.99	27.59	37.71	49.79	6.17	15.85	18.55	49.87	28.04	31.11	44.32	48.94
	CLAWS	**58.13**	**56.36**	**38.36**	**59.87**	**53.10**	**53.06**	**41.17**	**56.34**	**64.02**	**49.22**	**18.98**	**51.22**	**53.96**	**54.41**	**48.36**	**56.42**
Qwen-2.5	PPL	41.30	35.91	56.88	48.72	41.30	35.91	56.88	48.72	33.74	33.64	50.05	49.81	76.90	46.45	84.45	49.38
	WE	41.98	36.51	57.50	50.00	41.98	36.51	57.50	50.00	33.70	33.61	50.19	50.09	**77.69**	46.20	84.66	50.18
	LE	47.87	43.32	58.86	52.72	47.87	43.32	58.86	52.72	38.32	38.24	**51.16**	**51.99**	77.58	46.13	84.62	50.05
	HS	42.12	36.67	57.51	50.02	42.12	36.67	57.51	50.02	33.49	33.40	50.14	50.00	77.53	45.82	84.60	49.97
	AS	41.92	36.45	57.44	49.87	41.92	36.45	57.44	49.87	33.49	33.40	50.14	50.00	77.44	46.04	84.58	49.89
	CLAWS	**54.67**	**53.30**	**59.21**	**53.35**	**72.13**	**55.12**	**80.75**	**54.82**	**47.08**	**47.10**	49.11	47.82	74.68	**52.33**	**85.25**	**52.44**

Results & Analysis

Visualization of Baselines vs. CLAWS

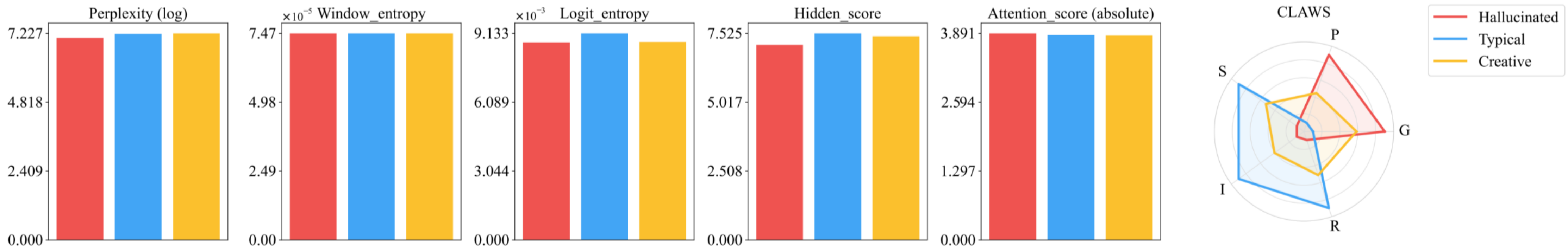


Figure 4. Visualization of class-wise average scores for each method (Qwen2.5-math-7B-inst)

- Baselines show almost no separation among classes
- **CLAWS clearly distinguishes Hallucinated, Typical, and Creative solutions.**

Results & Analysis

NEURAL INFORMATION PROCESSING SYSTEMS

Results on Runtime Consumption

- **CLAWS recorded the highest efficiency among all baselines**
  - **Baselines require re-feeding** outputs into the model
- **CLAWS** directly utilizes attention weights generated during output token generation **(no re-feeding required)**

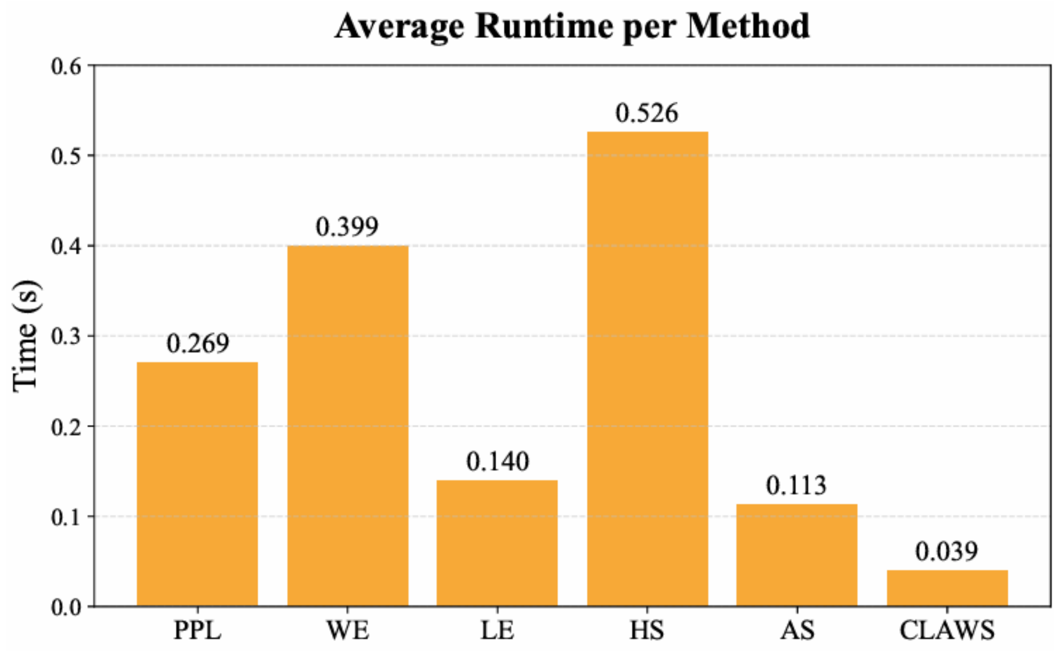


Figure 5. Average runtime per methods

Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

Contribution

I. Introduced **CLAWS, a novel white-box, attention-based method** for **creativity** and hallucination **detection**

II. Proposed **an automated framework for classifying** generated solutions into **Hallucinated, Creative, and Typical, without human intervention**

Thank you for watching :D

If you have any questions, feel free to contact us at

ktkpv94@hanyang.ac.kr , jeh0826@hanyang.ac.kr