THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Using Large Language Models to Automate Annotation and Part-of-Math Tagging of Math Equations

Authors: Ruocheng Shan & Abdou Youssef

The George Washington University

Part-of-Math (POM) Tagging and Annotation

Definition of POM tagging and math annotation:

- Identifying and labeling different components within math equations
  - Such as variables, operators, functions and constants
- Determining their roles and relationships within the equation

Applications of POM tagging:

- Math UIs
- Generating metadata to enrich math-IR systems, and improve their performance
- Create Math datasets for training/finetuning/testing specialized math-AI models

Research Overview

Traditional approaches to math annotation/POM tagging are:

- Manual or semi-manual
- Relying on crafted rules
- Having limited datasets

Objectives of this research

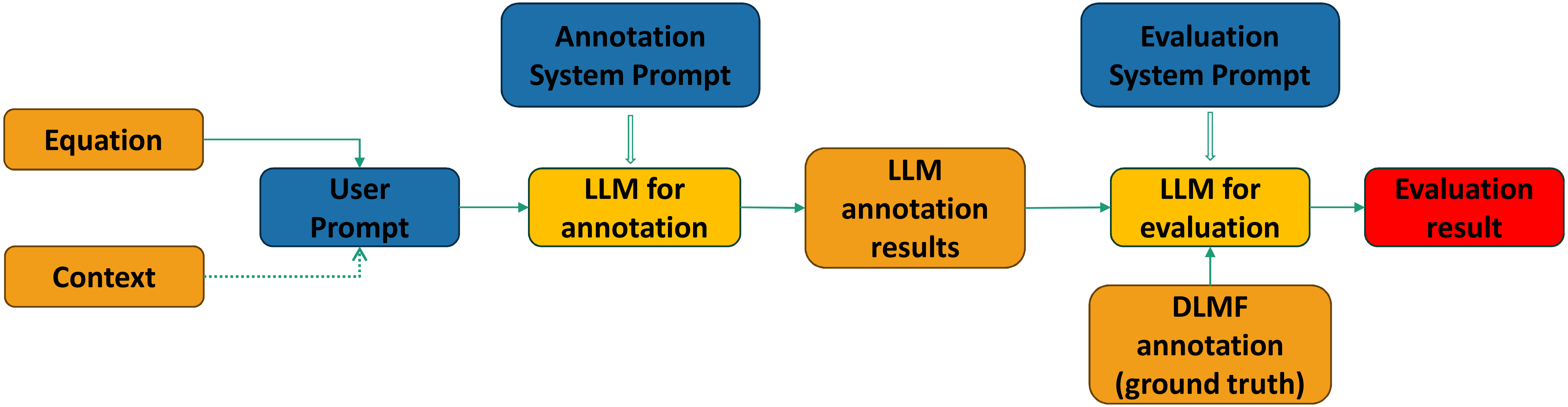
1. Explore the effectiveness of Large Language Models (LLMs) in automating annotation and POM tagging of math equations
   - To reduce human involvement, and improve annotation accuracy

   focus on GPT-3.5 Turbo

2. Investigate the impact of different levels of context on the accuracy of automated annotation

3. Explore the possibility of evaluating the annotation accuracy using LLMs

Methodology Overview



Dataset

DLMF

Digital Library of Mathematical Functions (**DLMF**) from National Institute of Standards and Technology (**NIST**)

https://dlmf.nist.gov/

Youssef & Miller Dataset

CICM'20

A Contextual and Labeled Math-Dataset

"Simple XML" files

Marked-up sentences

https://github.com/abdouyoussef/math-dlmf-dataset

Our Dataset

Extract **equations** from the XML files

Locating **different levels of content** of each equation

Total **7529** equations

The levels of context are:

- **no-context**
- **local-level context**
- **mid-sized context**
- **semi-global context**

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Example of Context Levels

Equation: $$f(z) = f(z_0) + \sum_{n=0}^{\infty} f_n (z - z_0)^{\mu+n}$$

Local (sentence) level context:

"Suppose that $f(z) = f(z_0) + \sum_{n=0}^{\infty} f_n(z - z_0)^{\mu+n}$, where $\mu > 0$, $f_0 \neq 0$, and the series converges in a neighborhood of z_0."

Mid-sized (paragraph) level context:

"Suppose that $f(z) = f(z_0) + \sum_{n=0}^{\infty} f_n(z - z_0)^{\mu+n}$, where $\mu > 0$, $f_0 \neq 0$, and the series converges in a neighborhood of z_0. (For example, when μ is an integer, $f(z) - f(z_0)$ has a zero of order μ at z_0.) Let $w_0 = f(z_0)$. Then (1.10.12) has a solution $z = F(w)$, where $F(w) = z_0 + \sum_{n=1}^{\infty} F_n(w - w_0)^{n/\mu}$ in a neighborhood of w_0 , nF_n being the residue of $1/(f(z) - f(z_0))^{n/\mu}$ at $z = z_0$."

Semi-global context (doc-level notation) :

x,y	real variables.
z	complex variable.
j,k,ℓ	integers.
m,n	nonnegative integers, unless specified otherwise.
⟨f,g⟩	inner, or scalar, product for real or complex vectors or functions.
…	……..

Prompt Engineering for Annotation

System Prompt

Background
You are a mathematician who is also experienced in computer science, especially natural language processing. You understand the essence of deep learning at the data format for training models. You are also familiar with the LaTex representation of math equations.
Task and instruction
Given an equation in LaTex format, segment the equation into math terms/chunks, and give each term a math explanation as annotation. You will also be given some basic contextual information about the input mathematical equation. Use the information as a supporting reference to the equation.
Notices and rules
Give your response in a key-value pair format where each key is a math term/chunk, and each value is an explanation of each term/chunk. Do not escape any math tokens. Return the result in a JSON format.

Background is an introduction and identification of LLM's role

Task and instruction is the specific job needed to be done by the LLM.

Notices and rules is additional dos and don't for the LLM to follow.

User Prompt

The equation is [*equation*],
and some contextual information about the equation: [*contexts*]

Template to input equation and contexts

LLM-Annotation of the All the DLMF Equations

- We got all the (7529) equations of the DLMF annotated by GPT
  - Using different levels of context

- Evaluated the GPT annotations in two ways
  - By hands
  - Using LLM

Annotation examples (No-context)

Equation	$z^n - 1 = (z-1)(z^{n-1} + z^{n-2} + \cdots + z + 1) = 0$
DLMF Annotation	z: variable, n: non-negative variable
LLM Annotation	z^n: The nth power of z, 1: The number 1, $(z-1)$: The difference between z and 1, $(z^{n-1} + z^{n-2} + \cdots + z + 1)$: The sum of terms from z to the power of n-1 to 1

Equation	$\mathrm{Si}\left(z\right) = z \sum_{n=0}^{\infty} \left(\mathsf{j}_n\left(\frac{1}{2}z\right)\right)^2$
DLMF Annotation	$\mathrm{Si}\left(z\right)$: sine integral, $\mathsf{j}_n\left(z\right)$: spherical Bessel function of the first kind, z: complex variable, n: non-negative integer
LLM Annotation	$\mathrm{Si}\left(z\right)$: Sine Integral function of z, z: Variable z, $\sum_{n=0}^{\infty}$: Summation from $n=0$ to infinity, $\mathsf{j}_n\left(\frac{1}{2}z\right)$: Bessel function of the first kind of order n with argument $1/2z$, ^{2}: Squared

Equation	$F(s) = \mathscr{L}f\ (s) = \int_0^{\infty} e^{-st} f(t) \mathrm{d}t$
DLMF Annotation	$\mathscr{L}f\ (s)$: Laplace transform, $\mathrm{d}x$: differential of x, e: base of natural logarithm, $\int$: integral
LLM Annotation	$F(s)$: The Laplace transform of the function f(t) with respect to the variable s, $\mathscr{L}f\ (s)$: Another way to represent the Laplace transform of the function f(t) with respect to the variable s, $\int_0^{\infty}$: The integral from 0 to infinity, e^{-st}: The exponential function with base e raised to the power of -st, $f(t)$: The function f evaluated at t, representing the input function being transformed, $\mathrm{d}t$: The differential element indicating that the integration is with respect to the variable t

Comparison between DLMF-Annotations and LLM-Annotations

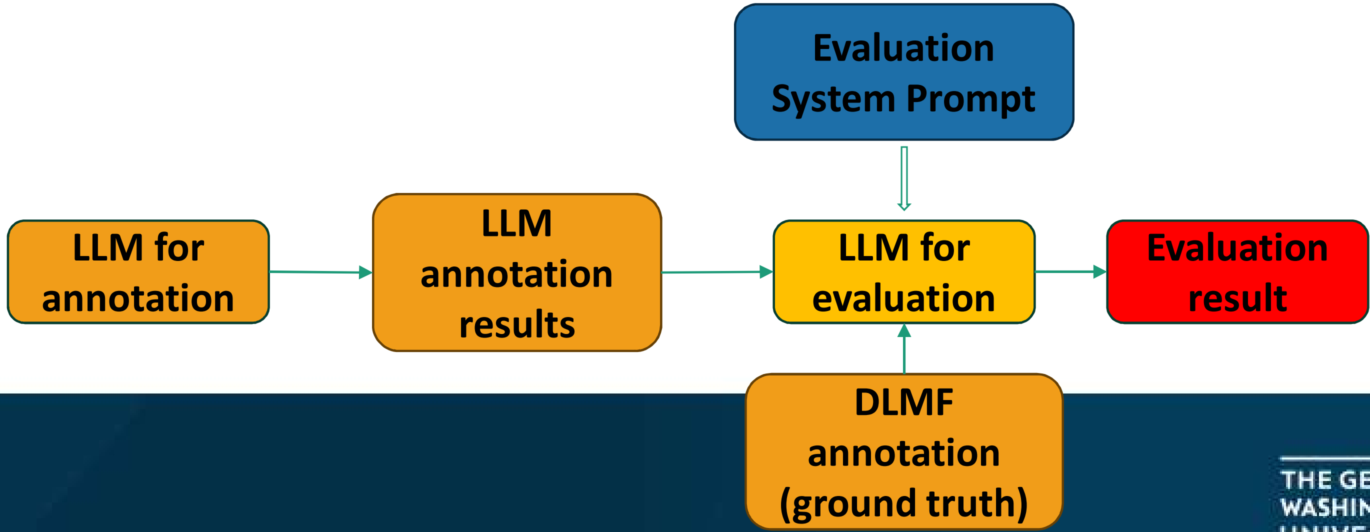
- **The numbers of annotated terms are different**
  - The DLMF annotations are not comprehensive: some terms are labeled, some are not
  - While the LLM tends to label (nearly) all math tokens/symbols
- **The granularity of math chunks varies**
  - Some labels are given to a single math term like “z”, while sometimes given to “z^n”
- **The annotation wordings are different**

- **More importantly: Can we trust the LLM-annotations?** **Subject to investigation**
- **Need to evaluate the LLM accuracy (e.g., compare the LLM annotations to the DLMF’s)**
- **But, the comparison is nuanced and non-trivial (as seen above)**
- **Idea: Use LLMs to do the comparisons and categorize the relationship between the two annotations**

- **Question: Can we trust the LLMs to do the comparison, i.e., to evaluate the accuracy of the LLM-annotations?**

LLM-based Evaluation

- Using a **separate LLM session** for evaluation (same LLM, different sessions)
- **Evaluation task** is viewed as **classification** of **annotation-pairs** (ground-truth annotation, LLM-generated ann.)
- First as **binary** classification of *consistent* vs. *non-consistent*
- Then as **multi-class** classification that is more refined, more informative, better aligned with reality



LLM-based Evaluation: Binary Classification – Prompt

System Prompt

Background
You are a mathematician who is also experienced in computer science, especially natural language processing. You understand the essence of deep learning at the data format for training models. You are also familiar with the Latex representation of math equations.

Task and instruction
There are two versions of math term annotations of a given equation, which also include the explanations of those math terms. Please determine whether the two versions are consistent or not.
If they are consistent, return "Yes". If not, return "No".

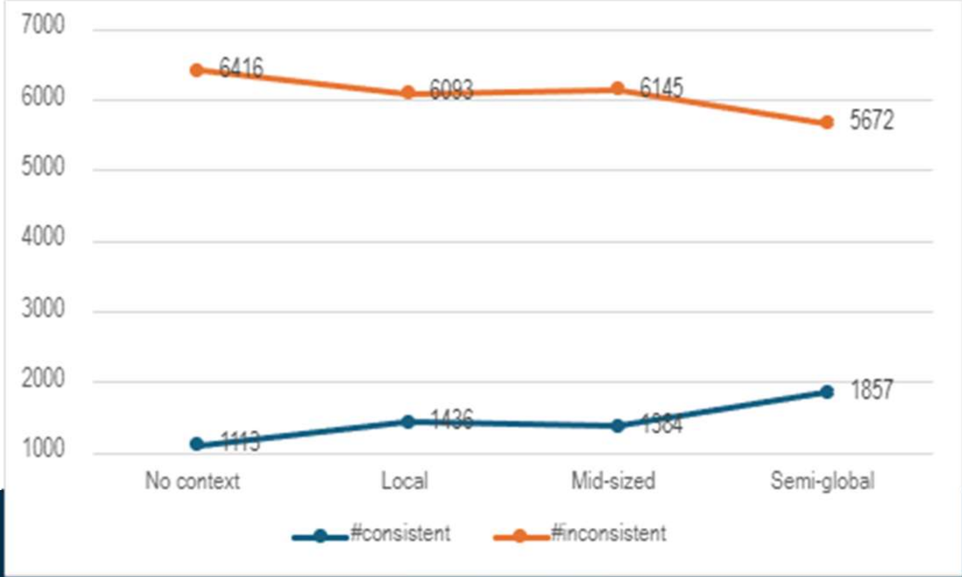
Notices and rules
Note that the granularity of the math term segments could be different but still be consistent. Note also that the format of two versions could also be different, so pay attention to the content instead of the format.

User Prompt

The first version of segmentation/annotation is [*version1*],
the second version of segmentation/annotation is [*version2*]

LLM-based Evaluation: Binary Classification Results

Context level	#consistent	#inconsistent	Consistency rate
No context	1113	6416	14.8%
Local context	1436	6093	19.1%
Mid-sized context	1384	6145	18.4%
Semi-global context	1857	5672	24.5%



- Overall consistency rate is low
- Consistency rate increases with context
    - Best when semi-global context was provided
- Binary classification is not enough to fully describe the relationship between two annotations

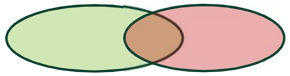
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

LLM-based Evaluation: Multi-Class Classification (More Refined)

System Prompt

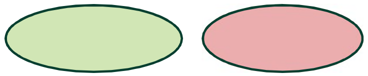
Background
You are a mathematician who is also experienced in computer science, especially natural language processing. You understand the essence of deep learning at the data format for training models. You are also familiar with the Latex representation of math equations.
Task and instruction
There are two versions of segmentations/annotations of a given equation, which also include the explanations of the math terms.
Please analyze and classify the two versions into one of the following classes: consistent, contradictory, indeterminate, independent/incomparable, mixed, subset-of, and superset-of.
Notices and rules
Simply return the classification as the provided label. Do not include explanations of your answer.
Follow the detailed explanations or instructions of the classes below.
consistent: Note that the granularity of segments could be different but still be consistent. Note that the format of two versions could also be different, pay attention to the content instead of the format.
contradictory: One version does not agree or is the opposite of another.
indeterminate: Do not have enough information to make the comparison possible.
independent/incomparable: The two versions talk about two totally different things.
mixed: The two versions are partially agreed or disagreed with one another.
subset-of: The first version is a subset of the second version.
superset-of: The first version is a superset of the second version.

- **Consistent**: the two versions (of annotation) agree with each other semantically (even if different words) $A_{dlmf} \equiv A_{llm}$
- **Subset-of**: 1st version is a subset of 2nd version $A_{dlmf} \subset A_{llm}$
- **Superset-of**: 1st version is a superset of 2nd version $A_{dlmf} \supset A_{llm}$
- **Mixed**: the two versions agree partially and differ partially



$A_{dlmf} \neq A_{llm}$, $A_{dlmf} \cap A_{llm} \neq \emptyset$

- **Contradictory**: one version conflicts with the other version
- **Independent**: the 2 versions are about two totally different things



$A_{dlmf} \cap A_{llm} = \emptyset$

- **Indeterminate**: the 2 versions do not have enough information to make the comparison possible

LLM-based Evaluation: Multi-Class Classification Results 1/2

Context Level / Result class	No context	Local level	Mid-sized	Semi-global
#consistent	1095	1111	1148	2028
#subset-of	2454	2943	2938	3809
#superset-of	1	3	2	9
#mixed	3748	3279	3240	1551
#contradictory	2	1	0	0
#independent	229	192	201	133
#indeterminate	0	0	0	0

favorable outcomes (#consistent, #subset-of)

unfavorable outcomes (#superset-of, #mixed, #contradictory, #independent)

LLM-based Evaluation: Multi-Class Classification Results 2/2

- Observably, the class of *consistent* and *subset-of* are both "good" results, so we group them into "*favorable outcomes*".
- The other classes are grouped into "*unfavorable outcomes*"

Context level / Result class	No context	Local level	Mid-sized	Semi-global
#favorable outcomes	3549	4054	4086	5837
#unfavorabl outcomes	3980	3475	3443	1692
favorable outcomes rate	47.1%	53.8%	54.3%	77.5%

How Good are LLMs as Evaluators?

- An LLM-as-evaluator is fundamentally a classifier of pairs (of annotations): consistent, contradictory, mixed, …
- Its classification accuracy may not be 100%
- To assess its classification (evaluation) accuracy, we need a labeled test set (ground truth)
  - Each instance is a pair of annotations, and a classification label of the relationship between the two annotations
- No such labeled dataset exits

- Creating such a dataset would be too time-consuming

- So we opted for a **statistical approximation** of the classification (evaluation) accuracy
  - We sampled 100 random instances, class-proportionally
  - Humanly labeled the pairs of annotations for all the 100 instances in the sample
  - Computed the classification (evaluation) accuracy based on that 100-instance sample

LLM-based Evaluation – Human verification of binary-classification

LLM-assigned / Human-assigned	Consistent	Inconsistent
Consistent	49	**10**
Inconsistent	**1**	40
Accuracy	98%	80%

Note: the context used here is the *semi-global context*

- Among 50 annotation-pairs labeled as "consistent" by LLM, 49 were found "consistent" by human
- Among 50 annotation-pairs labeled as "inconsistent" by LLM, 40 were found "inconsistent" by human

LLM-based Evaluation – Human verification of multi-class classification

Result class \ #Samples by	Evaluation by LLM	Evaluation by Human
#consistent	18	18
#subset-of	30	30
#superset-of	3	0
#mixed	44	10
#contradictory	2	0
#independent	3	2
#indeterminate	0	0

Note: the context used here is the *semi-global context*

- LLM is very **accurate** when assigning "*favorable outcome*" labels (*consistent* and *subset-of*)
- LLM is **less accurate** when assigning "*unfavorable outcome*" labels.
- High True-Positive rate, somewhat lower True-Negative rate

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Future work

Comparative Analysis of Different LLMs:

- Investigate the effectiveness of various LLMs (e.g., GPT-4, Gemini, Llama, …) in math annotation and POM tagging
- Explore specialized LLMs designed for scientific and mathematical contexts to understand their impact on performance
- Explore the option of finetuning LLMs for better math annotation and POM tagging

Integration of More Context:

- Incorporate external context, such as content dictionaries and knowledge graphs, to enhance LLM understanding and annotation of math equations

Questions?

Thank you!

An example of wrongly labeled “mixed” by the LLM

Equation:

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n.$$

DLMF Annotations:

!: factorial,
z: variable,
n: nonnegative integer

LLM Annotations:

$f(z)$: function f with input variable z,
$\sum_{\{n=0\}}^{\infty}$: sum from $n = 0$ to infinity,
$\frac{f^{\{(n)\}}(z_0)}{n!}$: nth derivative of f at z0 divided by n factorial,
$(z - z_0)^n$: $(z - z_0)$ raised to the power of n

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

An Example of the “mixed” Class of 2 Annotations

An example of “mixed” annotation-pairs

Equation	$v = \ln\left(\ln\left(\frac{1}{x}\right)\right) - 2 + \ln\pi$
DLMF Annotation	v: expansion variable (locally), π: the ratio of the circumference of a circle to its diameter, $\ln(z)$: principal branch of logarithm function, x: real variable
LLM Annotation (no context)	v: velocity, $\ln$: natural logarithm, $\frac{1}{x}$: inverse of x, π: π

- The “v” term has different annotation between DLMF and LLM
- Other terms agreed between the DLMF and the LLM

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

An Example of the “mixed” Class of an 2 Annotations

An example of “mixed” annotation-pairs

Equation	$v = \ln\left(\ln\left(\frac{1}{x}\right)\right) - 2 + \ln \pi$
DLMF Annotation	v: expansion variable (locally), π: the ratio of the circumference of a circle to its diameter, $\ln(z)$: principal branch of logarithm function, x: real variable
LLM Annotation (no context)	v: velocity, $\ln$: natural logarithm, $\frac{1}{x}$: inverse of x, π: π
LLM Annotation (semi-global context)	v: the variable representing the result of the equation, $\ln$: Natural logarithm function, $\frac{1}{x}$: Reciprocal of x, $\ln\left(\frac{1}{x}\right)$: Natural logarithm of the reciprocal of x, $\ln\left(\ln\left(\frac{1}{x}\right)\right)$: Natural logarithm of the natural logarithm of the reciprocal of x, -2: Constant value of negative two, $\ln \pi$: Natural logarithm of the mathematical constant π

- The “v” term has different annotation between DLMF and LLM
- Other terms agreed between the DLMF and the LLM

LLM-based Evaluation: Multi-Class Classification Results 1/3

Result class \ Context Level	No context	Local level	Mid-sized	Semi-global
#consistent	1095	1111	1148	2028
#contradictory	2	1	0	0
#indeterminate	0	0	0	0
#independent	229	192	201	133
#mixed	3748	3279	3240	1551
#subset-of	2454	2943	2938	3809
#superset-of	1	3	2	9

Favorable outcome

Unfavorable outcome

Unfavorable outcome

Unfavorable outcome

Favorable outcome

Unfavorable outcome

Part-of-Math (POM) Tagging and Annotation

Definition of POM tagging and math annotation:

- Identifying and labeling different components within math equations
  - Such as variables, operators, functions and constants
- Determining their roles and relationships within the equation

$$z^n - 1 = (z-1)(z^{n-1} + z^{n-2} + \cdots + z + 1) = 0$$

→

z^n: The nth power of z,
1: The number 1,
$(z-1)$: The difference between z and 1,
$(z^{n-1} + z^{n-2} + \cdots + z + 1)$: The sum of terms from z to the power of n-1 to 1

Applications of POM tagging:

- Math UIs
- Generating metadata to enrich math-IR systems, and improve their performance
- Create Math datasets for training/finetuning/testing specialized math-AI models