Assessing GPT Model Uncertainty in Mathematical OCR Tasks via Entropy Analysis

Alexei Kaltchenko

[*]Physics and Computer Science, Wilfrid Laurier University,
75 University Ave W, Waterloo, N2L3C5, Ontario, Canada.

Corresponding author(s). E-mail(s): akaltchenko@wlu.ca
ORCID: 0000-0001-5182-0357;

Abstract

In this work, we propose an entropy-based framework to quantify uncertainty in Generative Pre-trained Transformer (GPT) models when extracting mathematical equations from images of varying resolutions and converting them into mathematical notation. By measuring the conditional entropy of the model's output token sequences, we assess how confidently GPT recognizes and converts mathematical expressions into LaTeX format. Our experimental results, obtained using a custom Python implementation that we developed and made publicly available on GitHub, reveal an inverse relationship between image clarity and model uncertainty: high-resolution images yield lower entropy values and more accurate LaTeX outputs, while low-resolution inputs increase entropy and error rates. These findings demonstrate how entropy analysis, grounded in information-theoretic concepts, can effectively quantify GPT model uncertainty in real-world tasks.

Keywords: GPT Model Uncertainty, Machine Learning, Uncertainty Quantification, Entropy Analysis, Mathematical Equation Extraction

1 Introduction and Literature Review

Uncertainty quantification (UQ) in Large Language Models (LLMs), including Generative Pre-trained Transformer (GPT) models, has become increasingly important as these models are deployed in applications where reliability and trustworthiness are critical. One such application is Optical Character Recognition (OCR), particularly in tasks involving the recognition and extraction of mathematical expressions from images.

1.1 Uncertainty Quantification in Machine Learning and Language Models

Early work in deep learning UQ provided a foundation for understanding and reducing uncertainty in machine learning models by exploring methods like *Bayesian approximation* and *ensemble learning* [2]. These approaches aim to capture both epistemic (knowledge-based) and aleatoric (inherent) uncertainties, enhancing model robustness in tasks spanning *computer vision* and *natural language processing (NLP)*. Beyond model ensembling, *uncertainty optimization* strategies have also emerged, leveraging rough sets and uncertainty theory to select optimal feature subsets [25], thereby reducing ambiguity during classification.

In the context of language models, initial studies showed that larger models begin to learn tasks without explicit supervision [23], suggesting inherent abilities in models like GPT-2. However, the need for models to assess the validity of their own outputs led researchers to investigate *self-evaluation* mechanisms, where LLMs predict the correctness of their answers and estimate the probability of being true [11], thus improving calibration across diverse tasks. Recently, more specialized applications have emerged, including automated sarcasm detection to reduce the spread of misinformation or “fake news” [9], highlighting the breadth of domains that benefit from reliable UQ.

Subsequent efforts enabled GPT-3 models to express uncertainty in *natural language*, bridging the gap between probabilistic confidence and verbalized expressions of certainty [14]. This approach allows models to generate responses alongside confidence levels (e.g., “90% confidence”), providing more transparent and calibrated outputs. Notably, in *closed-source LLMs* where internal parameters remain inaccessible, *black-box* methods to measure uncertainty have gained attention [15, 31]. Such methods often rely on semantic dispersion or consistency checks (e.g., multiple rephrasings) to estimate response quality. Researchers have also studied prompting strategies to mitigate overconfidence in black-box LLMs [30] and developed interpretable models on engineered features to estimate confidence [20].

Beyond standard NLP tasks, distinguishing *epistemic* and *aleatoric* uncertainties remains a critical challenge. Methods that probe LLMs to detect when uncertainty arises from insufficient knowledge help in assessing output reliability [3]. Researchers have proposed model explanations that measure confidence by interpreting explanation consistency [5], while additional efforts demonstrate that derived probabilities can still predict correctness, even when miscalibrated [21]. Comparing probabilistic to verbalized confidence, studies show that the former is generally more accurate, although expressing internal confidence remains challenging [17]. Techniques like *reconfidencing* LLMs address grouping loss to align confidence scores with response accuracy [7].

To comprehensively evaluate UQ methods, some works have introduced datasets focusing on *data uncertainty* arising from irreducible randomness [32], enabling realistic assessments. Conformal prediction methods provide correctness coverage guarantees, converting heuristic uncertainty measures into rigorous prediction sets [28], while supervised approaches leverage labeled datasets to further enhance calibration across tasks [16]. Information-theoretic metrics have also been employed to differentiate epistemic from aleatoric uncertainties [1], helping identify conditions under which model outputs are unreliable due to knowledge gaps.

1.2 Entropy-Based Uncertainty Measures

A significant line of research has explored *entropy* as a measure of uncertainty in LLMs, including GPT models. Entropy, a fundamental concept in *information theory*, quantifies unpredictability within a system. Initially, entropy was employed to measure uncertainties in *geological models* [29] and *gene expression data classification* [26]. In more recent work, researchers introduced *semantic entropy* for *natural language generation* tasks, capturing linguistic invariances and shared meanings to better gauge confidence [13]. Building on this, *Semantic Entropy Probes (SEPs)* approximate entropy directly from hidden states to detect hallucinations in LLMs, reducing computational overhead without sacrificing performance [12].

Further extending entropy-based methods, the concept of *semantic density* was introduced to extract uncertainty information from probability distributions in semantic space, offering a task-agnostic approach with strong performance and robustness across multiple LLM benchmarks [22]. Meanwhile, some researchers have proposed new parametric entropy measures for *fuzzy* or *intuitive* sets [4, 24, 27], showcasing how advanced uncertainty metrics can be adapted to different types of ambiguous data. Additionally, GPT-based compression techniques approximating information distance for few-shot learning demonstrate links to *Kolmogorov complexity* [10], justifying certain pre-training objectives and improving tasks like semantic similarity or zero-shot classification.

Other recent works have leveraged fuzzy set theory for multi-criteria decision making under uncertainty. For instance, *modified TOPSIS* with q-rung picture fuzzy information measure has been applied to green supplier selection [8], while new *uncertainty-based optimization* approaches have been proposed for fuzzy feature subset selection [25]. These methods underscore how entropy-based and fuzzy-logic concepts can handle imprecise data across a wide range of applications.

Moreover, utilizing internal states from LLMs has proven effective for detecting hallucinations. By analyzing semantic information retained within the model, methods like [6] identify hallucinated content without relying solely on output tokens. Such techniques highlight the diversity of *entropy-based*, *information-theoretic*, and *fuzzy-logic* strategies for UQ.

1.3 Meta-Analysis of Existing Research

Table 1 summarizes key literature on UQ, entropy-based methods, and OCR tasks. We include works covering Bayesian ensembles, semantic entropy, fuzzy-logic approaches, and other advanced metrics, illustrating the breadth of methods applied in different domains. Notably, while GPT-based references cover a variety of UQ methods [18, 19], there remains a gap in uniting *GPT-driven image recognition and OCR* for *mathematical expressions* with an *entropy-based* approach.

Table 1 Meta-analysis of representative works in UQ, entropy-based methods, fuzzy approaches, and OCR.

Reference	**Model/Approach**	**Uncertainty Method**	**Task/Application**
Abdar et al. [2]	Bayesian/Ensemble	Epistemic, Aleatoric	General ML (Computer Vision/NLP)
Fkih et al. [9]	ML-based Classifier	Sarcasm Detection	Fake News Prevention
Sinha et al. [25]	Rough Set Feature Sel.	Uncertainty Optimization	Classification, ML
Taruna et al. [27]	Generalized Exp. Entropy	Exponential Entropy	Intuitionistic Vague Sets
Raj et al. [24]	Fuzzy Soft Matrices	Cosine Sim., Entropy	Decision-Making
Arora et al. [4]	Pythagorean Fuzzy Sets	Entropy in MCDM	Decision-Making
OpenAI [14]	GPT-3	Confidence Expressions	NLP Uncertainty
Wellmann et al. [29]	Geological Modeling	Shannon Entropy	Geological Uncertainty
Sun et al. [26]	Feature Selection	Shannon Entropy	Gene Expression
Kuhn et al. [13]	GPT-based LLM	Semantic Entropy	Hallucination Detection
Wang et al. [12]	GPT-based LLM	SEPs	Hallucination Detection
Liu et al. [22]	GPT-based LLM	Semantic Density	UQ Across Benchmarks
Huang et al. [10]	GPT-based Compression	Kolmogorov Complexity	Few-Shot Learning

Several additional works provide deeper insights into black-box calibration [30], interpretable confidence estimation [20], reconfidencing LLMs [7], or distinguishing epistemic from aleatoric components [1]. Specialized datasets for data uncertainty [32], conformal prediction [28], and supervised calibration [16] further refine UQ approaches. Yet few attempts focus on *mathematical OCR* with GPT-based models and *entropy metrics*, leaving a notable gap for our current work.

1.4 Our Contribution

While significant progress has been made in UQ for LLMs—including the entropy-based and fuzzy-logic approaches discussed above—very little work has examined *GPT-based image recognition and mathematical OCR*. To our knowledge, **no existing approach combines GPT-driven image recognition and math OCR with entropy-based uncertainty quantification (UQ)**.

This paper addresses that gap by introducing an *entropy-focused framework* for quantifying GPT model uncertainty in recognizing and converting mathematical expressions from images of varying resolutions. By computing the *conditional Shannon entropy* of output token sequences, our method yields a *quantitative measure* of the model's uncertainty. Empirically, we show that *higher-resolution images* produce lower entropy and fewer errors, whereas *lower-resolution images* increase entropy and recognition errors.

In summary, our key contributions are as follows:

- An **entropy-driven strategy** for GPT-based image recognition and math OCR,
- An **information-theoretic analysis** of GPT model uncertainty and error rates,
- **Freely available Python code**, facilitating replication and laying the groundwork for future extensions.

We provide our Python implementation openly on GitHub, thereby promoting *reproducibility* and further investigation into entropy-based techniques for GPT-driven mathematical OCR.

1.5 Organization of the Paper

Section 2 introduces fundamental concepts from *information theory* as they apply to OCR systems, focusing on mutual information and conditional entropy in a noisy channel setting. Section 3 details how GPT-based OCR can be modeled via entropy measures, including a token-level analysis of uncertainty. Section 4 presents our methodology algorithmically.

In Section 5, we describe our experiments with the OpenAI GPT-4o model [18, 19] on multi-resolution images and discuss the results in terms of entropy and accuracy. Finally, Section 6 addresses limitations and future work, and Section 7 concludes with overall findings and implications for further research.

2 Information-Theoretic Foundations and the Impact of Image Resolution on OCR

In this section, we present a unified information-theoretic framework for OCR tasks and examine how image resolution influences uncertainty in recognized text. We begin by defining essential concepts of entropy and mutual information in the OCR context, then analyze how varying image resolution affects conditional entropy and error rates.

2.1 Entropy and Mutual Information in OCR

Let X be the random variable representing the input image, and Y be the random variable representing the recognized text output from an OCR system. The entropy of X, denoted $H(X)$, measures the information content or uncertainty in the image; the entropy of Y, denoted $H(Y)$, measures the uncertainty in the recognized text. Their joint entropy, $H(X, Y)$, characterizes the combined uncertainty of image-text pairs.

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x), \quad H(Y) = -\sum_{y \in \mathcal{Y}} P(y) \log P(y). \tag{1}$$

The mutual information $I(X; Y)$ quantifies the amount of information in X that is preserved in Y:

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \tag{2}$$

A higher $I(X; Y)$ indicates better retention of the input information by the OCR system (i.e., fewer recognition errors).

2.2 Noise and the Channel Model

In practice, OCR can be viewed as transmitting an image X through a noisy channel—where noise may stem from poor lighting, image blur, or compression artifacts—to produce the recognized text Y. Noise introduces uncertainty, thereby reducing $I(X; Y)$. In an ideal noiseless channel, we would have $I(X; Y) \approx H(X)$, meaning all information from the image is retained in the recognized text.

2.3 Conditional Entropy and Uncertainty

The conditional entropy $H(Y \mid X)$ measures the remaining uncertainty about the recognized text Y after observing the image X. A lower $H(Y|X)$ corresponds to higher confidence in the OCR output, since the image provides more clues to accurately recognize the text:

$$H(Y|X) = H(Y) - I(X;Y). \tag{3}$$

When $H(Y)$ is approximately constant (e.g., when a fixed text is being recognized), any changes in $H(Y|X)$ directly reflect changes in $I(X;Y)$.

2.4 Resolution Effects on Mutual Information and Conditional Entropy

2.4.1 High-Resolution Images

High-resolution images X contain more detailed character features, which increases $I(X;Y)$ and thus decreases $H(Y|X)$. In other words, the OCR system faces less ambiguity because additional visual cues help distinguish characters more reliably.

- **Mutual Information Increases.** Detailed input images yield higher $I(X;Y)$.
- **Conditional Entropy Decreases.** Since $H(Y)$ is fixed, increasing $I(X;Y)$ reduces $H(Y|X)$, leading to fewer recognition errors.

2.4.2 Low-Resolution Images

Low-resolution images lack clear character boundaries, decreasing $I(X;Y)$ and, consequently, increasing $H(Y|X)$. As resolution diminishes, the OCR system becomes more uncertain about character identities.

- **Mutual Information Decreases.** Less-detailed images convey less information about the text.
- **Conditional Entropy Increases.** The loss of character detail elevates uncertainty, leading to higher error rates.

2.4.3 Data Processing Inequality

The Data Processing Inequality states:

$$I(X;Y) \geq I(f(X);Y), \tag{4}$$

where f is any processing of X. Reducing image resolution can be viewed as applying a lossy function f on X, which cannot increase $I(X;Y)$. Hence, lowering resolution inevitably leads to higher uncertainty $H(Y|X)$.

2.4.4 Fano's Inequality

Fano's Inequality provides a bound relating conditional entropy to the probability of recognition errors P_e:

$$H(Y|X) \geq H(P_e) + P_e \log(|\mathcal{Y}| - 1). \tag{5}$$

Lower $H(Y|X)$ thus implies lower P_e. Improving resolution (and thus reducing $H(Y|X)$) decreases the likelihood of misclassification in the OCR process.

2.5 Practical Implications for OCR Systems

- **Optimizing Resolution.** Increasing resolution boosts $I(X;Y)$, reducing $H(Y|X)$ and error rates. Although higher-resolution images demand more storage and processing resources, the gain in OCR accuracy can justify the cost.
- **Balancing Performance.** In settings with limited computational or bandwidth resources, moderate resolutions may be chosen to balance OCR accuracy and system constraints.

In summary, an information-theoretic perspective clarifies how image resolution directly impacts the conditional entropy $H(Y|X)$ in OCR systems. High-resolution images retain more relevant cues for accurate character recognition, decreasing uncertainty and error rates. Conversely, low-resolution inputs increase ambiguity and reduce $I(X;Y)$, thus elevating $H(Y|X)$. Tools such as the Data Processing Inequality and Fano's Inequality provide theoretical grounding for how noise and reduced image detail degrade OCR performance.

3 GPT as the Entropy Model

In this section , we consider using a GPT model for task of recognizing and extracting mathematical expressions from a given image, and saving them in LaTeX format. In such a setting, the interaction with the GPT model is as follows : the GPT model is provided with an input $\boldsymbol{x}$, which includes the image along with a verbal prompt, instructing the model to recognize math expressions in the image and output them in LaTeX format.

The model 's output is an text sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ composed from tokens. Let ϕ represent the GPT model , where $\phi(\boldsymbol{x}, y_{1:i-1}) = P_i(y_i|\boldsymbol{x}, y_1, y_2, \ldots, y_{i-1})$ models the probability distribution from the next token y_i given the input $\boldsymbol{x}$ and the preceding output tokens $y_1, y_2, \ldots, y_{i-1}$. The function $\phi(\boldsymbol{x}, \boldsymbol{y})$ outputs the sequence of next token probability distributions $(P_1, P_2, \ldots, P_n)$ for the entire output sequence .

Thus, for a given input $\boldsymbol{x}$, the GPT model generates the output $\boldsymbol{y}$ along with auxiliary information consisting of the conditional probabilities $P_i(y_i|\boldsymbol{x}, y_{1:i-1})$ by each output token y_i.

3.1 Shannon Entropy of the Output Token Sequence Given the Input

The Shannon entropy measures the average amount of information or uncertainty associated with a random variable. In our context, for each position i in the output token sequence, the entropy H_i of the next-token distribution P_i is given by:

$$H_i = -\sum_{y \in \mathcal{V}} P_i(y|\boldsymbol{x}, y_{1:i-1}) \log P_i(y|\boldsymbol{x}, y_{1:i-1}), \tag{6}$$

where $\mathcal{V}$ is the vocabulary of possible tokens.

The total conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$ of the output token sequence $\boldsymbol{y}$, given the input $\boldsymbol{x}$, is computed by summing the entropies at each position:

$$H(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{n} H_i = -\sum_{i=1}^{n} \sum_{y \in \mathcal{V}} P_i(y|\boldsymbol{x}, y_{1:i-1}) \log P_i(y|\boldsymbol{x}, y_{1:i-1}). \tag{7}$$

This total conditional entropy represents the cumulative uncertainty of the GPT model in generating the output sequence $\boldsymbol{y}$ given the input $\boldsymbol{x}$.

3.2 Normalized Shannon Entropy of the Output Token Sequence

To facilitate comparison across different sequences or models, we compute the normalized Shannon entropy at each position by dividing the entropy H_i by the maximum possible entropy at that position. The maximum entropy occurs when the next-token distribution is uniform over the vocabulary $\mathcal{V}$, in which case, we have $H_i^{\max} = \log |\mathcal{V}|$, where $|\mathcal{V}|$ is the size of the vocabulary.

The normalized entropy h_i at position i is then defined as follows:

$$h_i = \frac{H_i}{H_i^{\max}} = \frac{H_i}{\log |\mathcal{V}|}. \tag{8}$$

Similarly, the normalized total entropy $H(\boldsymbol{y}|\boldsymbol{x})$ of the output sequence $\boldsymbol{y}$, given the input $\boldsymbol{x}$, is defined as:

$$H(\boldsymbol{y}|\boldsymbol{x}) = \frac{H(\boldsymbol{y}|\boldsymbol{x})}{H^{\max}(\boldsymbol{y})} = \frac{\sum_{i=1}^{n} H_i}{n \log |\mathcal{V}|} = \frac{1}{n} \sum_{i=1}^{n} h_i. \tag{9}$$

The normalized entropy $H(\boldsymbol{y}|\boldsymbol{x})$ ranges from 0 to 1, where 0 indicates no uncertainty (the model is certain about its next token at each position), and 1 indicates maximum uncertainty (the model's predictions are uniformly distributed over the vocabulary).

3.3 Interpretation and Relevance

Calculating the conditional Shannon entropy $H(\boldsymbol{y}|\boldsymbol{x})$ of the output token sequence provides insights into the GPT model's confidence and uncertainty during the generation process given the input $\boldsymbol{x}$. Lower entropy values suggest that the model is more confident about its predictions, which is desirable in tasks like mathematical expression recognition where precision is critical.

In the context of extracting mathematical expressions in LaTeX format, *low entropy* indicates that the model confidently predicts the next token given the input, suggesting a high likelihood of correctly recognizing and formatting the mathematical expressions. *High entropy* suggests uncertainty in the model's predictions given the input, which may lead to errors in the recognized expressions or incorrect LaTeX syntax.

By analyzing the entropy values, we assess the model's performance and identify positions in the output sequence, where the model may require additional support or where errors are more likely to occur.

3.4 Practical Computation

When the GPT model generates the output sequence $\boldsymbol{y}$, it provides the conditional probabilities $P_i(y_i|\boldsymbol{x}, y_{1:i-1})$ for each token y_i. We compute the entropy H_i at each position using these probabilities. If the full distribution $P_i(y_i|\boldsymbol{x}, y_{1:i-1})$ is available (i.e. for each value of $y \in \mathcal{V}$), we can compute H_i exactly as per equation (6). In other words, the exact calculation of the entropy requires access to the entire probability distribution over the vocabulary at each position.

In practice, however, the model only outputs the probabilities for a subset of tokens (e.g., the top-k most probable tokens), so we approximate the entropy by considering only the available probabilities as follows:

$$H_i = -\sum_{y \in \tilde{\mathcal{V}}} P_i(y|\boldsymbol{x}, y_{1:i-1}) \log P_i(y|\boldsymbol{x}, y_{1:i-1}), \tag{10}$$

where $\tilde{\mathcal{V}}$ is a subset of the vocabulary of all possible tokens, such as, for example, the top-k most probable tokens.

3.5 Example Calculation

Suppose the GPT model provides the following next-token probabilities (for simplicity, we consider a small subset of the vocabulary):

Token y	$P_i(y\mid\boldsymbol{x}, y_{1:i-1})$
‘\{}‘	0.60
‘{‘	0.20
‘}‘	0.10
Other	Remaining probabilities summing to 0.10

The entropy at position i is calculated as follows:

$$\begin{aligned} H_i &= -\left[0.60 \log 0.60 + 0.20 \log 0.20 + 0.10 \log 0.10 + 0.10 \log 0.10\right] \\ &= -\left[-0.442 + -0.322 + -0.230 + -0.230\right] \\ &= 1.224 \text{ bits} \end{aligned} \tag{11}$$

Assuming a vocabulary size of $|\mathcal{V}| = 10,000$, the maximum entropy at position i is:

$$H_i^{\max} = \log 10,000 = 13.29 \text{ bits} \tag{12}$$

The normalized entropy is:

$$h_i = \frac{H_i}{H_i^{\max}} = \frac{1.224}{13.29} \approx 0.092 \tag{13}$$

A normalized entropy of approximately 0.092 indicates low uncertainty at this position given the input $\boldsymbol{x}$.

3.6 Relationship of $H(\boldsymbol{y}|\boldsymbol{x})$ to $H(Y|X)$

The calculated conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$ in this section provides a practical, empirical measure of the uncertainty the GPT model has in generating the output token sequence $\boldsymbol{y}$ given the input $\boldsymbol{x}$. This entropy relates directly to the theoretical concepts of conditional entropy $H(Y|X)$ discussed in Section 2 as follows:

- Empirical Measurement of $H(Y|X)$: The entropy $H(\boldsymbol{y}|\boldsymbol{x})$ calculated here represents the sum of the uncertainties at each step in the generation of $\boldsymbol{y}$ by the GPT model, given the input $\boldsymbol{x}$. It serves as an empirical approximation of the conditional entropy $H(Y|X)$ in the OCR context, where Y is the output text and X is the input image.
- Connection to OCR Uncertainty: In Section 2, $H(Y|X)$ quantifies the uncertainty in recognizing the text Y given the image X. Similarly, $H(\boldsymbol{y}|\boldsymbol{x})$ reflects the GPT model‘s uncertainty in generating each token y_i given $\boldsymbol{x}$ (which includes the image and the prompt).
- Impact of Image Resolution: The discussion in Section 2 about how image resolution affects $H(Y|X)$ are mirrored in how the quality and clarity of the input image $\boldsymbol{x}$ influence $H(\boldsymbol{y}|\boldsymbol{x})$. A higher-resolution image lead to lower entropy $H(\boldsymbol{y}|\boldsymbol{x})$, indicating that the model has less uncertainty in generating the correct LaTeX code for the mathematical expressions.

4 Algorithm: GPT-Based Mathematical OCR with Entropy Calculation

Input:

- A PDF document or page(s) containing mathematical expressions.
- A chosen image resolution R (e.g., 72, 96, 150, 300 dots per inch (dpi)).

Output:

1. Recognized LaTeX code for the mathematical expressions.
2. Conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$ measuring the model's uncertainty.

Step 1: Image Conversion and Preprocessing

1. Select the PDF page(s) containing mathematical expressions.
2. Convert the page(s) to an image $\boldsymbol{x}$ at resolution R.
3. (Optional) If necessary, apply basic enhancements or denoising to improve clarity.

Step 2: Formulate Prompt for GPT

1. Prepare a system prompt specifying the role of the GPT model (e.g., advanced OCR focusing on math).
2. Include instructions (user prompt) to extract all mathematical expressions from the image and return them in valid LaTeX format.

Step 3: Inference with GPT

1. Pass the image $\boldsymbol{x}$ and prompts to the GPT model.
2. Obtain the token-by-token output $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, where each y_i is a token in the recognized LaTeX code.
3. Retrieve the model-generated probabilities (or log probabilities) for each token $P_i\big(y_i|\boldsymbol{x}, y_{1:i-1}\big)$.

Step 4: Entropy Computation

1. For each output token position i in $\boldsymbol{y}$, compute:

$$H_i = -\sum_{y\in\mathcal{V}} P_i\big(y|\boldsymbol{x}, y_{1:i-1}\big)\, \log P_i\big(y|\boldsymbol{x}, y_{1:i-1}\big). \tag{14}$$

2. Sum these entropies for all positions to obtain total conditional entropy:

$$H(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{n} H_i. \tag{15}$$

3. Compute the normalized entropy by dividing the total entropy by the maximum possible entropy for the sequence, which is $n\log_2|\mathcal{V}|$, where n is the number of tokens and $|\mathcal{V}|$ is the vocabulary size.

Step 5: Postprocessing and Evaluation

1. Collect the recognized LaTeX expressions from $\boldsymbol{y}$.
2. Check formatting or compile the LaTeX (if desired) to verify correctness.
3. Compare recognized expressions to ground truth (if available) to calculate accuracy or error rates.

Step 6: Repeat for Different Image Resolutions

1. Vary R (e.g., 72, 96, 150, 300 dpi).
2. Re-run Steps 1–5 for each resolution to analyze how resolution impacts $H(\boldsymbol{y}|\boldsymbol{x})$.
3. Summarize trends by correlating image resolution with entropy values and observed errors.

End of Algorithm

5 Experimental Results

5.1 Experimental Setup

In this section, we present the results of our experiments using the OpenAI GPT-4o model [18, 19] to recognize and extract mathematical expressions from images at different resolutions.

Data Source.

We obtained a publicly available PDF from arXiv [33], selecting page 3 for its representative mathematical expressions. A standard PDF-to-image conversion tool was used to generate four image sets (JPEG format) at *72*, *96*, *150*, and *300* dots per inch (dpi). This range of resolutions allows us to systematically investigate how image clarity influences both OCR accuracy and the entropy-based uncertainty metrics introduced in Section 2.

Experimental Procedure.

For each resolution, we prompted GPT-4o to extract the depicted math expressions into valid LaTeX. We then computed token-by-token conditional entropy values, following the method in Section 3, and compared the recognized LaTeX output to ground-truth expressions derived from the original PDF. This setup enabled us to evaluate both the model's performance and its uncertainty (via entropy) under varying image qualities. Complete instructions for replicating these steps along with the Python code are provided in our GitHub repository at: `github.com/Alexei-WLU/gpt-ocr-entropy-analysis`.

5.2 Analysis and Discussion

Image Resolution (dots per inch (dpi))	**Number of Output Tokens**	**Total Entropy (bits)**	**Normalized Entropy (bits)**	**Recognition Accuracy**
300	453	7.77	0.0171	Perfect
150	434	12.28	0.028	Near perfect
96	476	18.27	0.038	With some errors
72	473	38.91	0.082	With many errors

Table 2 Comparison of entropy values at different image resolutions.

The experimental results demonstrate how image resolution affects the GPT model's performance in recognizing and converting mathematical expressions into LaTeX format.

- Entropy Trends: As the image resolution decreases, the total conditional entropy $H(\boldsymbol{y}|\boldsymbol{x})$ increases. This indicates that the model's uncertainty in generating the output tokens grows when processing lower-resolution images.

- Normalized Entropy: The normalized entropy per token also increases with decreasing resolution, from 0.0171 bits per token at 300 dpi to 0.0823 bits per token at 72 dpi. This suggests that each token generated from lower-resolution images carries more uncertainty.

- Output Tokens: The number of output tokens remains relatively stable across different resolutions, with slight variations.

- Impact on Accuracy: Higher entropy values are associated with an increased likelihood of errors in the recognized mathematical expressions. Lower-resolution images provide less detailed information, leading to higher uncertainty in the model's predictions and recognition errors, in perfect agreement with the information-theoretic results discussed in previous sections.

6 Limitations and Future Work

Entropy Calculation: The entropy calculations are based on the log probabilities provided by the model. If the model does not output the full probability distribution over the vocabulary, the entropy estimates may be approximations. Extending the approach to retrieve more complete distributions (e.g., using top-k or nucleus sampling) could improve accuracy and reveal further insights into uncertainty at each token step.

Model Limitations: The OpenAI GPT-4o model's performance may be influenced by factors beyond image resolution, such as its training data or inherent capabilities in handling complex mathematical notation. Future research could investigate the performance of alternative LLMs or specialized OCR engines with built-in mathematical parsing to compare uncertainty profiles.

Further Experiments: Here, we tested only four discrete resolutions *as proof of concept.* Future work could involve a broader spectrum of image qualities (e.g., 50–600 dpi) and additional real-world image distortions, such as varying lighting conditions, rotation, or partial occlusion. This would clarify how well entropy-based metrics generalize to practical use cases like scanned textbooks or mobile phone captures.

Error Analysis: A detailed analysis of errors in the generated LaTeX code could provide insights into specific challenges faced by the model at different resolutions. In particular, future work might categorize error types (e.g., missing braces, misrecognized symbols) and develop targeted strategies or prompt engineering solutions to reduce these errors.

Future Scope: Beyond the tasks described above, our entropy-based framework could be extended to:

- *Multi-Modal Inputs:* Investigate uncertainty when combining mathematical expressions with textual explanations or diagrams.
- *Interactive Systems:* Incorporate feedback loops where a user or a secondary system highlights uncertain tokens for re-query or correction.
- *Confidence-Weighted Models:* Develop downstream applications—like automated grading or scientific metadata extraction—where the entropy signals guide reliability thresholds for acceptance or manual review.

These directions open possibilities for improving GPT-based OCR in educational technologies, automated scientific document analysis, and more.

7 Conclusion

In this paper, we investigated the uncertainty of GPT models in extracting mathematical equations from images of varying resolutions and converting them into LaTeX code. By employing concepts of entropy and mutual information, we examined the recognition process and assessed the model's uncertainty in this specialized OCR task. Our study fills a notable gap in existing research by applying entropy-based methods to GPT-driven mathematical OCR.

Our experimental results demonstrate a clear relationship between input image resolution and GPT model uncertainty. Higher-resolution images yield lower entropy values, indicating reduced uncertainty and improved accuracy in the recognized LaTeX code. Conversely, lower-resolution images lead to increased entropy, reflecting heightened uncertainty and a greater likelihood of recognition errors. These findings empirically validate the theoretical framework established earlier, underscoring how image resolution significantly influences GPT-based OCR outcomes.

The key contributions of this work lie in connecting information-theoretic concepts, such as conditional entropy and mutual information, to the practical performance of GPT-based OCR systems. By calculating the conditional Shannon entropy of the output token sequences, we offered a quantitative measure of the model's uncertainty—extending foundational principles of information theory to modern LLMs.

Furthermore, we have developed Python code for these conditional entropy calculations, which is openly available on GitHub. This ensures reproducibility and encourages further exploration of our methods. Our approach highlights the value of entropy analysis in quantifying GPT model uncertainty in mathematical OCR tasks, offering benefits for applications such as scientific publishing, digital libraries, and educational technologies.

Looking ahead, and as also discussed in Section 6, this work can be expanded by integrating additional image distortions, incorporating confidence-aware prompt engineering, and exploring alternative LLMs or specialized math OCR systems. Such efforts will help refine our understanding of GPT-based mathematical OCR under diverse conditions and enhance real-world applicability in a range of scientific and educational settings.

Acknowledgments

We would like to express our gratitude to the anonymous reviewers for their insightful comments and constructive feedback, which have significantly contributed to improving the clarity and rigor of this manuscript.

Declarations

Conflict of Interest

The authors have no relevant financial or non-financial interests to disclose. The authors have no conflict of interest to declare for the content of this article.

Data Availability

The data used in this study consist of images generated from a PDF file freely available on arXiv at https://arxiv.org/abs/2106.13823. We focused on page 3 due to its variety of mathematical expressions. The page was converted into JPEG images at 72, 96, 150, and 300 dpi, but these images are not directly included in the GitHub repository. They can, however, be recreated by following the instructions provided there. Any additional data (e.g., recognized LaTeX outputs or intermediate log probabilities) can be made available upon reasonable request.

Code Availability

The Python code used in this study, developed by the authors, is openly accessible on GitHub at https://github.com/Alexei-WLU/gpt-ocr-entropy-analysis. This repository contains all scripts necessary to reproduce the experimental results presented in this paper, including guidance on generating the images from the arXiv PDF and running the entropy-based uncertainty calculations.

References

[1] Abbasi-Yadkori Y, Kuzborskij I, György A, et al (2024) To believe or not to believe your llm. ArXiv 2406.02543

[2] Abdar M, Pourpanah F, Hussain S, et al (2020) A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion 76:243–297

[3] Ahdritz G, Qin T, Vyas N, et al (2024) Distinguishing the knowable from the unknowable with language models. ArXiv 2402.03563

[4] Arora HD, Naithani A (2023) Empirical evaluation of pythagorean fuzzy entropy measures with application in decision making. International Journal of Information Technology pp 1–10. URL https://api.semanticscholar.org/CorpusID:264582229

[5] Becker E, Soatto S (2024) Cycles of thought: Measuring llm confidence through stable explanations. ArXiv 2406.03441

[6] Chen C, Liu K, Chen Z, et al (2024) Inside: Llms' internal states retain the power of hallucination detection. ArXiv 2402.03744

[7] Chen L, Perez-Lebel A, Suchanek FM, et al (2024) Reconfidencing llms from the grouping loss perspective. ArXiv 2402.04957

[8] Dhumras H, Bajaj RK, Shukla V (2023) On utilizing modified topsis with r-norm q-rung picture fuzzy information measure green supplier selection. International Journal of Information Technology pp 1 – 7. URL https://api.semanticscholar.org/CorpusID:259145883

[9] Fkih F, Rhouma D, Alghofaily HA (2024) A semantic approach for sarcasm identification for preventing fake news spreading on social networks. International Journal of Information Technology URL https://api.semanticscholar.org/CorpusID:272172361

[10] Huang CY, Xie Y, Jiang Z, et al (2023) Approximating human-like few-shot learning with gpt-based compression. ArXiv 2308.06942

[11] Kadavath S, Conerly T, Askell A, et al (2022) Language models (mostly) know what they know. ArXiv 2207.05221

[12] Kossen J, Han J, Razzak M, et al (2024) Semantic entropy probes: Robust and cheap hallucination detection in llms. ArXiv 2406.15927

[13] Kuhn L, Gal Y, Farquhar S (2023) Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. ArXiv 2302.09664

[14] Lin SC, Hilton J, Evans O (2022) Teaching models to express their uncertainty in words. Trans Mach Learn Res 2022

[15] Lin Z, Trivedi S, Sun J (2023) Generating with confidence: Uncertainty quantification for black-box large language models. Trans Mach Learn Res 2024

[16] Liu L, Pan Y, Li X, et al (2024) Uncertainty estimation and quantification for llms: A simple supervised approach. ArXiv 2404.15993

[17] Ni S, Bi K, Yu L, et al (2024) Are large language models more honest in their probabilistic or verbalized confidence? ArXiv 2408.09773

[18] OpenAI (2023) Gpt-4 technical report. arXiv arXiv:2303.08774. URL https://arxiv.org/abs/2303.08774, updated: 2024-11-30

[19] OpenAI (2024) OpenAI Platform Documentation. URL https://platform.openai.com/docs/models, updated: 2024-11-30

[20] Pedapati T, Dhurandhar A, Ghosh S, et al (2024) Large language model confidence estimation via black-box access. ArXiv 2406.04370

[21] Plaut B, Nguyen K, Trinh T (2024) Probabilities of chat llms are miscalibrated but still predict correctness on multiple-choice q&a. ArXiv

[22] Qiu X, Miikkulainen R (2024) Semantic density: Uncertainty quantification in semantic space for large language models. ArXiv 2405.13845

[23] Radford A, Wu J, Child R, et al (2019) Language models are unsupervised multitask learners

[24] Raj M, Tiwari P, Gupta P (2022) Cosine similarity, distance and entropy measures for fuzzy soft matrices. International Journal of Information Technology 14:2219 – 2230. URL https://api.semanticscholar.org/CorpusID:249319368

[25] Sinha AK, Shende P, Namdev N (2022) Uncertainty optimization based feature subset selection model using rough set and uncertainty theory. International Journal of Information Technology 14:2723 – 2739. URL https://api.semanticscholar.org/CorpusID:249653024

[26] Sun L, Zhang X, Qian Y, et al (2019) Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Inf Sci 502:18–41

[27] Taruna, Arora HD, Tiwari P (2021) A new parametric generalized exponential entropy measure on intuitionistic vague sets. International Journal of Information Technology 13:1375 – 1380. URL https://api.semanticscholar.org/CorpusID:234826165

[28] Wang Z, Duan J, Cheng L, et al (2024) Conu: Conformal uncertainty in large language models with correctness coverage guarantees. ArXiv 2407.00499

[29] Wellmann JF, Regenauer-Lieb K (2012) Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models. Tectonophysics 526:207–216

[30] Xiong M, Hu Z, Lu X, et al (2023) Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. ArXiv 2306.13063

[31] Yang A, Chen C, Pitas K (2024) Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries. ArXiv 2405.13907

[32] Yang Y, Yoo H, Lee H (2024) Maqa: Evaluating uncertainty quantification in llms regarding data uncertainty. ArXiv 2408.06816

[33] Zhou S (2021) Quantum data compression and quantum cross entropy. ArXiv 2106.13823. URL https://arxiv.org/abs/2106.13823

Appendix

A Experiment 1: 300 dpi Resolution

Original document: [arXiv:2106.13823] Page-3

Image file: 2106.13823v3-P3-300.jpg

Image resolution: 300 dots per inch (dpi)

Number of output tokens: 453

Total Entropy: 7.77 bits

Normalized Entropy: 0.0171 bits

Recognized Math Equations:

$$Q(N,\rho,\epsilon)=\sum_{x\in\epsilon\text{-strong-typical}}|i_1\rangle\langle i_1|\otimes|i_2\rangle\langle i_2|\otimes\cdots\otimes|i_N\rangle\langle i_N|. \tag{11}$$

$$\operatorname{tr}\left(Q(N,\rho,\epsilon)\rho^{\otimes N}\right)\geq 1-\delta. \tag{12}$$

$$\operatorname{tr}\left(Q(N,\rho,\epsilon)\rho^{\otimes N}\right)=\sum_{x\in\epsilon\text{-strong-typical}}P(i_1)P(i_2)\cdots P(i_N)=\sum_{x\in\epsilon\text{-strong-typical}}P(x). \tag{13}$$

$$\langle l\rangle=\sum_{i=1}^{D}p_il_i=\sum_{i=1}^{D}p_i\log\frac{1}{q_i}=H(p,q), \tag{16}$$

$$H(p,q)\leq\langle l\rangle<H(p,q)+1. \tag{17}$$

$$\sigma=\sum_{i=1}^{D}q_i|a_i\rangle\langle a_i|,\quad \rho=\sum_{i=1}^{D}p_i|b_i\rangle\langle b_i|,\quad \{q_i\}\neq\{p_i\},\quad \{|a_i\rangle\}\neq\{|b_i\rangle\}. \tag{15}$$

B Experiment 2: 150 dpi Resolution

Original document: [arXiv:2106.13823] Page-3

Image file: 2106.13823v3-P3-150.jpg

Image resolution: 150 dpi

Number of output tokens: 434

Total Entropy: 12.28 bits

Normalized Entropy: 0.028 bits;

Recognized Math Equations:

$$Q(N,\rho,\epsilon)=\sum_{|i\rangle\in\epsilon\text{-strong-typical}}|i_1\rangle\langle i_1|\otimes|i_2\rangle\langle i_2|\otimes\cdots\otimes|i_N\rangle\langle i_N|. \tag{11}$$

$$\operatorname{tr}\left(Q(N,\rho,\epsilon)\rho^{\otimes N}\right) \geq 1-\delta. \tag{12}$$

$$\operatorname{tr}\left(Q(N,\rho,\epsilon)\rho^{\otimes N}\right) = \sum_{x\in\epsilon\text{-strong-typical}} P(i_1)P(i_2)\cdots P(i_N) = \sum_{x\in\epsilon\text{-typical}} P(x). \tag{13}$$

$$H(p,q) = \sum_{i=1}^{D} p_i \log\frac{1}{q_i} = H(q,p), \tag{16}$$

$$H(p,q) \leq \langle l\rangle < H(p,q)+1. \tag{17}$$

$$\sigma = \sum_{i=1}^{D} q_i |a_i\rangle\langle a_i|, \quad \rho = \sum_{i=1}^{D} p_i |b_i\rangle\langle b_i|, \quad \{a_i\} \neq \{p_i\}, \quad \{a_i\} \neq \{b_i\}. \tag{15}$$

C Experiment 3: 96 dpi Resolution

Original document: [arXiv:2106.13823] Page-3

Image file: 2106.13823v3-P3-96.jpg

Image resolution: 96 dpi

Number of output tokens: 476

Total Entropy: 18.27 bits

Normalized Entropy: 0.038 bits

Recognized Math Equations:

$$Q(N,\rho,\epsilon) = \sum_{|i\rangle\ \epsilon\text{-strongly typical}} |i\rangle\langle i| \otimes |i_2\rangle\langle i_2| \otimes |i_N\rangle\langle i_N| \tag{11}$$

$$\operatorname{tr}\left(Q(N,\rho,\epsilon)\rho^{\otimes N}\right) \geq 1-\delta \tag{12}$$

$$\operatorname{tr}\left(Q(N,\rho,\epsilon)\rho^{\otimes N}\right) = \sum_{x\ \epsilon\text{-strongly typical}} P(i_1)P(i_2)\cdots P(i_N) = \sum_{x\ \epsilon\text{-strongly typical}} P(x) \tag{13}$$

$$(1-\delta)^2 2^{N(S(\rho)-\epsilon)} \leq T(N,\rho,\epsilon) \leq 2^{N(S(\rho)+\epsilon)} \tag{14}$$

$$\sigma = \sum_{i=1}^{D} q_i |a_i\rangle\langle a_i|, \quad \rho = \sum_{i=1}^{D} p_i |b_i\rangle\langle b_i|, \quad \{q_i\} \neq \{p_i\}, \quad \{|a_i\rangle\} \neq \{|b_i\rangle\} \tag{15}$$

$$l(x) = \left\lceil \log \frac{1}{q_x} \right\rceil = H(p,q) \tag{16}$$

$$H(p,q) \leq l(x) < H(p,q) + 1 \tag{17}$$

D Experiment 4: 72 dpi Resolution

Original document: [arXiv:2106.13823] Page-3

Image file: 2106.13823v3-P3-72.jpg

Image resolution: 72 dpi

Number of output tokens: 473

Total Entropy: 38.91 bits

Normalized Entropy: 0.082 bits

Recognized Math Equations:

$$Q(\mathcal{N},\rho,\epsilon) = \sum_{k:\langle k|\rho|k\rangle \geq \epsilon} |k\rangle\langle k| \otimes |k\rangle\langle k| \otimes |k\rangle\langle k| \tag{11}$$

$$\mathrm{tr}(Q(\mathcal{N},\rho,\epsilon)\rho^{\otimes n}) \geq 1 - \delta \tag{12}$$

$$\mathrm{tr}(Q(\mathcal{N},\rho,\epsilon)\rho^{\otimes n}) = \sum_{x\in\epsilon\text{-strong-typical}} P(x)\cdot P(x)\cdot\ldots\cdot P(x) = \sum_{x\in\epsilon\text{-strong-typical}} P(x) \tag{13}$$

$$(1-\delta)2^{n(S(\rho)-\epsilon)} \leq P(\mathcal{N},\rho,\epsilon) \leq 2^{n(S(\rho)+\epsilon)} \tag{14}$$

$$\langle \ell \rangle = \sum_x P(x) \cdot \left\lceil \log \frac{1}{P(x)} \right\rceil = H(P,q) \tag{16}$$

$$H(q) \leq \langle \ell \rangle \leq H(q) + 1. \tag{17}$$

$$\sigma = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i| = \sum_i p_i |\psi_i\rangle\langle\psi_i| \tag{18}$$