LLM-based Hierarchical Label Annotation for Foodborne Illness Detection on Social Media

Dongyu Zhang*
AML
ByteDance
San Jose, USA
dongyu.zhang@bytedance.com

Ruofan Hu
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
rhu@wpi.edu

Dandan Tao
Vanke School of Public Health
Tsinghua University
Beijing, China
dandantao@mail.tsinghua.edu.cn

Hao Feng
College of Ag&Environ Sciences
North Carolina A&T State University
Greensboro, USA
hfeng@ncat.edu

Elke Rundensteiner
Computer Science/Data Science Program
Worcester Polytechnic Institute
Worcester, USA
rundenst@wpi.edu

***Abstract*—Foodborne illnesses pose a threat to public health, leading to morbidity, mortality, and economic burden annually. Social media, while providing a rich timely source for training AI models for surveillance, requires effective tools for annotation. While Large Language Models (LLMs) have shown promise for generating simple labels, here hierarchical labels composed of entity types like food type and symptom (at individual word level) and the foodborne illness event (at complete post level) are required. For this, we introduce ICL2FID, the first *LLM-based hierarchical labeling* framework designed to annotate social media posts for foodborne illness detection at two levels using only *a few demonstration examples.* To utilize the interconnection between post and word levels, ICL2FID instructs the LLM to leverage information from one level when predicting the other level. To combat model hallucination and cyclic dependencies, a verification step improves evidence propagation between interconnected word and post-level labeling tasks. Strategies for custom selection of demonstration examples are designed reducing biases and increasing representation. We compare ICL2FID against traditional supervised learning and other LLM methods, demonstrating that it not only achieves superior accuracy but does so at a fraction of the cost and time. These findings highlight ICL2FID's potential as a viable alternative for hierarchical label generation in scenarios with limited resources and huge data sets. Code is available at https://github.com/zdy93/ICL2FID.**

***Index Terms*—Learning with Limited Examples, Large Language Model, In-Context Learning, Chain of Thought**

I. Introduction

Detecting Foodborne Illness Signals in Social Media. Foodborne illnesses pose a major public health challenge, affecting millions in the U.S. annually, leading to loss of productivity, increased healthcare costs, and fatalities [1]. Early detection is vital for managing outbreaks and protecting public health, yet data gathering from formal sources like hospitals or the CDC can be dangerously slow [2]. Instead, consumer-generated data from social media has proven invaluable for public health surveillance. Tools using data from platforms like Twitter (X) [1], Yelp, and Google searches have been deployed across major cities including New York, Chicago, and Las Vegas [3]–[7].

Need for Costly Multi-level Labeling. Machine and deep learning models are crucial for identifying potential foodborne illness from social media posts [2], [8]. Effective models must detect a *two-level structure,* meaning, they must not only determine if a *post* indicates an illness but also extract associated *relevant entities* like food items or symptoms. This two-level task requires the models to both classify the *overall relevance of the post* and to pinpoint *specific entities* associated with the illness to help trace the source and spread of outbreaks. Supervised models require high-quality hierarchically labeled data to ensure accurate results. This demands quality labeling work that is challenging, resource-intensive, and extremely costly. Previous studies [9] using crowdsourcing platforms to gather labels instead of domain experts found significant quality disparities between crowdsourced and expert labels, with costs still high — over $500 per 1,000 tweets annotated.

LLMs to-the-Rescue and SOTA Limitations. Recently, Large Language Models (LLMs) have shown strong in-context learning (ICL) capabilities, making predictions based on a task description and a few examples for tasks like math problems [10], [11]. ICL leverage pretrained models without parameter updates, which is advantageous in resource-limited settings such as ours. In fact, using LLMs for text annotation [12], [13] has been proven more cost-effective than human labeling [10], [14]–[16]. LLMs have recently been applied to tasks like text classification [17], [18], information extraction [12], [13], [19], and machine translation [20]. However, most

This work was supported by the AFRI award no. 2020-67021-32459 from the U.S. Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) and the Illinois Agricultural Experiment Station.

*This work was performed while Dongyu Zhang was a Ph.D. student at Worcester Polytechnic Institute.

[1]We henceforth use the term Twitter Corporation instead of X, because the data set we work with had been extracted before renaming of the company.

existing research limits on single-level annotation. Hierarchical information extraction methods in [21], [22], without self-revisions, tend to accumulate errors from model hallucination. Our study fills this gap by expanding the use of LLMs to multi-level data annotation of social media posts with self-revisions, exploring how the overall context and specific entities interact to enhance understanding of the post content, an area previously unexplored.

Problem Definition. This study tackles the complex challenge of automatically annotating social media posts for foodborne illness detection with two-level labels using a limited number of demonstration examples. As illustrated in Figure 1, given an unlabeled data set of social media posts and a small number of labeled hierarchical examples, such as in TWEET-FID [9], the aim is to develop an LLM-based in-context learning framework. This framework should be capable of generating high-quality hierarchical labels, comprised of post-level annotations to identify foodborne illness incidents and associated word-level annotations to classify each word's entity type related to the incident.

Challenges. This annotation task presents several challenges:

• *Interdependency Between the Two Hierarchical Levels.* This task involves annotating both post and word levels. At the word level, an entity must be directly linked to a foodborne illness incident to be considered relevant. For instance, a sentence like "Just watched a documentary on food safety, which opened my eyes to the importance of avoiding food poisoning. #awareness," mentions "food poisoning" and "food safety" without linking to an actual incident. This necessitates that the LLM accurately assesses the relevance of such entities to an actual foodborne illness incident. Conversely, at the post level, the LLM must determine if any entities suggest a foodborne illness incident. This complex task requires precise guidance for the LLM to comprehend and accurately perform the annotations considering their hierarchical interdependencies.

• *Model Hallucination in Labeling Procedure.* LLMs are prone to "hallucination," in that they generate content that does not align with the ground truth [23], [24]. In our situation, most posts do not indicate foodborne illness incidents; plus, relevant entities are rare within the dataset [9]. However, the LLM might incorrectly label irrelevant posts or words as related, resulting in numerous false positives. Such hallucinations, if kept unchecked, may then disproportionally propagate errors to the subsequent label step of the post based on its erroneous entity components, or, vice versa - in both cases further exacerbating the problem.

• *Resource and Token Constraints.* Access to advanced closed-source LLMs via an API or to locally hosted open-source LLMs incurs costs in terms of money or GPU hours, calculated per token for both inputs and outputs [25]. Additionally, LLMs tend to be limited by their context windows [25], [26], which restrict the length and thus scope of task inquiries and the number of examples that can be included in the demonstration context. It can also restrict the length of output the model can generate; potentially impacting the design of the two-level labeling process. It is thus crucial to design the labeling framework carefully to ensure high-quality annotations while efficiently managing token usage and controlling costs.

Proposed Method. To overcome these challenges, we propose a novel labeling framework called ICL2FID: In-context Learning based Annotation for Two-level Foodborne Illness Detection. This framework decomposes the hierarchical labels annotation process into three sequential steps, each with a unique example selection strategy matching the distinct requirement of that step. These strategies ensure representative diversity of demonstration posts and labels at each step, minimizing repetitive exposure and potential biases. First, in the word-level labeling step, the LLM assesses the overall relevance of the post to a foodborne illness incident and then identifies relevant entities. This is accomplished with the *Semantic Similarity* strategy choosing examples closely related to the query post. Second, in the word-level label verification step, the model verifies the relevance of each of the identified entities to the foodborne illness incident. This is achieved by the *Existence Diversity* strategy providing both positive and negative examples to thoroughly verify relevancy. After verification, irrelevant entities are discarded and the remaining are aggregated per post. For the final post-level labeling step, the LLM is guided to scrutinize whether the aggregate of these word-level results is related to the food poisoning incident to determine the post-level label. For this, we have developed the *Augmented Diversity* strategy, which enriches the demonstration examples with both accurate and inaccurate word-level results. This guides the LLM to identify errors from the previously verified word-level results.

Contributions. Our key contributions are as follows:

• We propose ICL2FID, the first LLM-based labeling framework for annotating posts with two-level hierarchical labels for foodborne illness detection. It generates word and post-level labels through a sequence of steps that uniquely utilize the interdependency between these levels, enabling the LLM to leverage information from one level to enhance predictions at the other. This yields improved labeling results on both levels.

• To minimize model hallucination and errors compounding due to dependencies, our Existence Diversity and Augmented Diversity strategies prompt the model to perform detailed diversified analyses, reducing the risk of errors from previously incorrect results and with it tackling the cyclic dependency. The verification step effectively eliminates incorrect entities identified earlier.

• In our evaluation study, ICL2FID surpasses both traditional supervised learning approaches and modern ICL-based LLM methods. This holds even when ICL2FID is given a small number of labeled demonstration posts. Its performance is very close to human annotation in label quality, yet with significantly reduced costs.

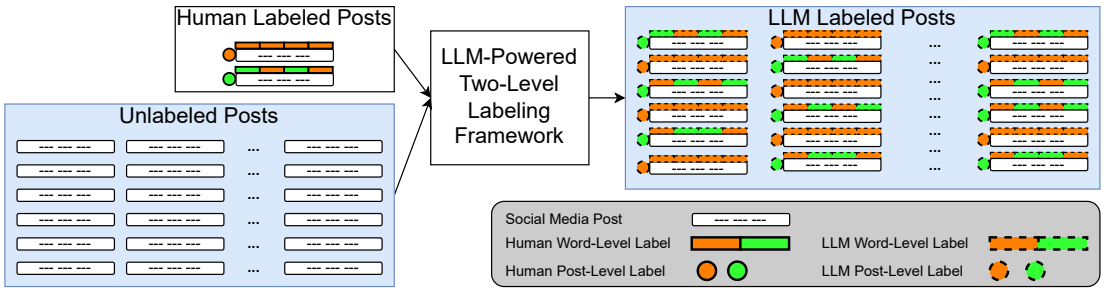


Fig. 1: LLM-based Two-level Foodborne Illness Detection Label Annotation with limited human annotated samples. Given a social media post dataset for foodborne illness detection task, where limited number of posts have human labels for both two levels. Our goal is to develop an in-context LLM-based learning framework to assign labels for unlabeled posts.

II. Related Works

LLMs have advanced natural language processing capabilities [24]. An effective application of pretrained LLMs is *in-context learning (ICL),* where LLMs generate text aligned with given contexts using provided sample inputs and outputs [10], [27]. ICL being a training-free learning framework lowers computational costs otherwise required with tuning models to new tasks [10]. An advancement in ICL, Chain-of-Thought (CoT), introduces an intermediate reasoning step into the demonstrations to enhance LLMs' performance on complex tasks by predicting not only the final answer but also the intermediate reasoning process [11]. ICL and CoT methods have recently been applied to tasks like text classification [17], [18], document information extraction [12], [19], machine translation [20], and relation extraction [13]. Studies on hierarchical information extraction, including intent classification and slot filling cited in [21], [22], develop 'one-way' methods that first extract coarse-grained information, followed by corresponding fine-grained details, without incorporating self-verification or revisions. However, effectively leveraging the interconnections between labels across levels in hierarchical labeling remains an unresolved issue in previous works.

Foodborne Illness Detection Dataset Labeling. Social media data offers timely scalable information crucial for public health. Most previous research focused on identifying a single class label per post to indicate its relevance to foodborne illness events. These studies utilized machine learning to identify relevant posts within specific regions [3], [5], [6], [8], [28], [29]. Beyond that, retrieving more detailed information about these potential events required a manual inspection process. TWEET-FID [9] is the first publicly available social media dataset annotated for foodborne illness detection at both tweet and word levels with labels provided by both experts and crowdsourced workers to support comparative studies TWEET-FID [9] demonstrates that strong performing multi-task classification models can be achieved when provided with high-quality hierarchically-labeled data. Yet, it is well recognized that human annotation of hierarchical labeling of large datasets is prohibitively costly – calling for automated labeling as tackled in our work.

III. Methodology

A. Problem Definition

Let $D = \{\mathbf{x}_i\}_{i=1}^{N}$ denote a dataset of N social media posts. Each post $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,M^i}]$ is a sequence of M^i words. The dataset D splits into a very large unlabeled set $D^u = \{\mathbf{x}_i^u\}_{i=1}^{N^u}$ and a very small human-labeled subset $D^l = \{\mathbf{x}_i^l\}_{i=1}^{N^l}$, with D denoting the union of D^u and D^l. Each $\mathbf{x}_i^l$ in D^l corresponds to a triplet $(\mathbf{x}_i^l, y_i^l, \mathbf{s}_i^l)$, where $y_i^l \in \{0, 1\}$ indicates whether $\mathbf{x}_i^l$ describes a foodborne illness incident (1 for yes, 0 for no), and $\mathbf{s}_i^l = [s_{i,1}^l, s_{i,2}^l, \ldots, s_{i,M^i}^l]$ denotes a sequence of word-level labels, categorizing each word into one of five classes $(c_1, \ldots c_5)$ as detailed in Table I. Due to the high costs of obtaining human annotations, the large majority of posts remain unlabeled, leading to a ratio of $\frac{N^u}{N^l} \gg 1$. Given the labeled set D^l and the unlabeled set D^u, our goal is to develop a hierarchical labeling framework to annotate each post $\mathbf{x}_i^u$ in D^u with its sequence of accurate word-level labels $\hat{\mathbf{s}}_i^u$ and its post-level label $\hat{y}_i^u$.

More precisely, we aim to identify four types of relevant entities (food, location, symptom, keyword) within each post, with all other words classified as *outside* these categories, as detailed in Table I. Each word $x_{i,j}^u$ in post x_i^u is classified into one of five categories $(c_1, \ldots, c_5)$. Only entities directly associated with a foodborne illness incident are considered *relevant.* For instance, the word "apple" in the sentence "I ate an apple and it tastes great!" is irrelevant to a foodborne illness incident. Being not a relevant entity, it should be classified as c_5: *outside* category.

B. Proposed Approach: ICL2FID

1) Overview of Hierarchical Labeling Methodology: We design ICL2FID, a hierarchical framework, ICL2FID, leveraging a pretrained LLM, denoted $\Phi(\theta)$, to annotate social media posts for foodborne illness detection on both post and word levels. The complexity of labeling at both levels requires a detailed input context (prompt) that includes (1) the task description, (2) definitions of the four entity types, (3) examples to illustrate the desired output format, and (4) query text (i.e., post to be labeled). With many LLMs having strict token limits for inputs and outputs, this poses challenges for including comprehensive descriptions and multiple examples

TABLE I: Definition of word level label classes.

Label	Definition
Food c_1	Food item that caused potential foodborne illness incident.
Location c_2	Location where affected person purchased or acquired the food associated with potential foodborne illness.
Symptom c_3	Symptom experienced by affected person as a result of suspected foodborne illness.
Keyword c_4	Other relevant keywords or terms associated with a foodborne illnesses incident, *e.g.*, "food poisoning".
Outside c_5	Words that do not belong to any (relevant) class above. Mentions of entities not related to foodborne illness incident.

into a single input context. For instance, GPT-3.5-turbo allows up to 16,385 tokens for input and 4,096 for output, while GPT-4 allows up to 8,192 tokens for input [25]. Given these restrictions on the lengths, the crafting of an effective yet compact input context that includes both a detailed context and sufficient examples for accurate annotation within these constraints can be problematic. These constraints in part guided our design. That is, the prompting instruction to ask the model to return labeled sentences for the four types of relevant entities and the post-level class (see description in next subsection) experienced repeated failures as the model could not provide an answer due to these token constraints. To overcome this limitation, we structured the ICL2FID annotation process into a series of three interconnected phases:

1) **Word Level Labeling:** $\Phi(\theta)$ identifies the four types of relevant entities in post $\mathbf{x}_i^u$ from the unlabeled dataset D^u, conducting separate a iteration for each entity type. Using the Chain-of-Thought (CoT) method [11], ICL2FID asks the LLM $\Phi(\theta)$ to consider the post's overall relevance to foodborne illness incident while identifying relevant entities within.
2) **Word Level Label Verification:** LLMs $\Phi(\theta)$ tend to suffer from the hallucinations [23], and with that they may overconfidently label irrelevant words as relevant entities [12]. To alleviate the model hallucination problem and its wrongful propagation to the 2nd level of the hierarchy, our next step is for $\Phi(\theta)$ to verify the correctness of each extracted entity's class from the first step. In this verification step, the LLM $\Phi(\theta)$ is instructed to evaluate the identified entity's relevance to the foodborne illness incident in the reasoning. Entities deemed irrelevant are discarded.
3) **Post Level Labeling:** Despite the word-level verification process, some irrelevant entities may persist. With the post $\mathbf{x}_i^u$ and its verified entities as context, LLM $\Phi(\theta)$ is guided to first analyze the word-level result's correctness and then use that verified evidence to determine the post-level label.

The entire annotation methodology ICL2FID is "training free", meaning the model does not update its parameters. In our experiments, we use GPT-3.5-turbo [25] as the backbone LLM [2], that is, neither the LLM nor the embedding model (detailed in the following subsection) is deployed locally. Thus our ICL2FID labeling methodology can be done with limited local computational resources.

ICL2FID's first step produces post-level relevance results in the reasoning step and word-level labeling results. However, due to the model's tendency for hallucination, these results are verified and re-evaluated in subsequent steps. Our method generates and verifies results from distinct perspectives: searching for evidence and analyzing its correctness. These interrelated tasks utilize the interconnection between the two levels of the labels while mitigating hallucination and cyclic dependency problems.

Next, we describe the methodologies for constructing prompts and demonstration examples using specialized retrieval strategies.

2) Step 1: Word Level Labeling - Considering Post-level Relevance Before Making Word-level Prediction.: In the word-level labeling step, the LLM is instructed to identify the four types of *relevant* entities in the post. As we discussed above, due to input and output length restrictions, it is impractical to address all four entity types in a single prompt. Therefore, we generate four separate prompts for each post operating in parallel, each focusing on one entity type.

Figure 2b displays an example of word-level labeling composed of *three parts.* The Part 1 (blue box in Figure 2b) is the task description, which provides a task overview, instructing the model to identify food entities associated with foodborne illness incidents. The opening sentences outline the task, followed by a description of the expected answer format.

Inspired by GPT-NER [12], the LLM's output is designed to simply replicate non-relevant content while using special tokens "^^" to highlight relevant pieces. The following is an example of how best to ask the LLM to point at the extracted relevant *food* entity:

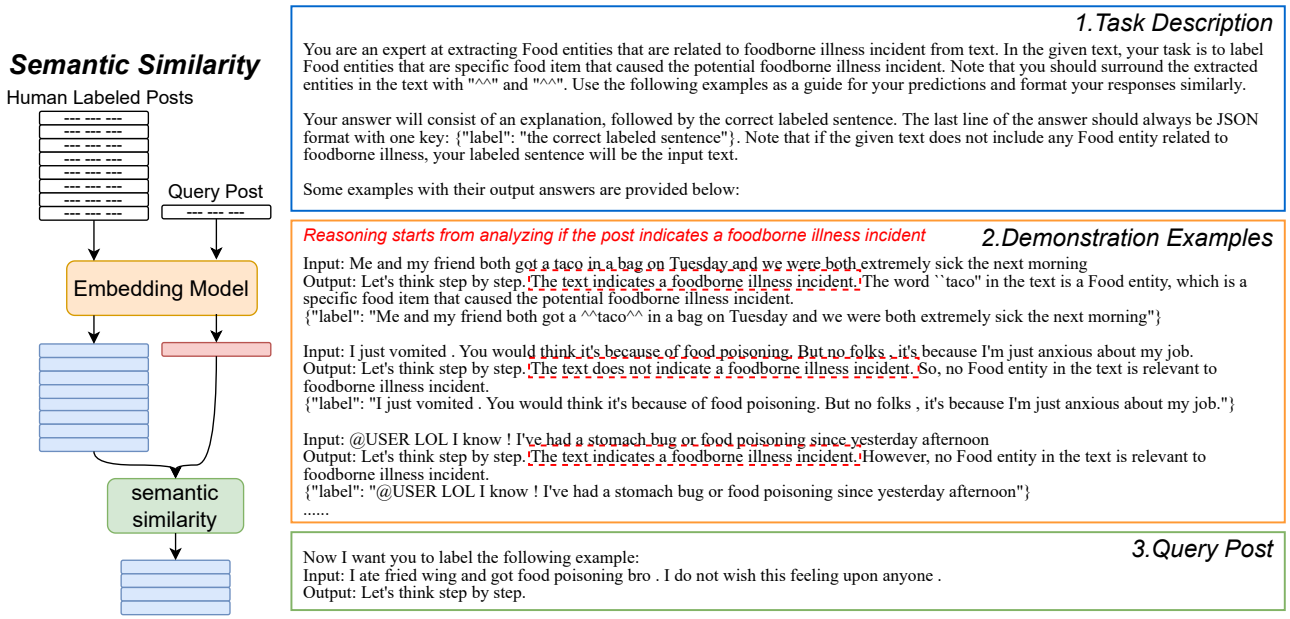
Input: I ate chicken and got diarrhea.

Labels: I ate ^^chicken^^ and got diarrhea.

As discussed in [12], this new format narrows the gap between sequence labeling tasks and generative modeling. Given the sentence "I ate chicken and got diarrhea". The intuitive format of the word-level label sequence is: "O O FOOD O O SYMPTOM", where "O" denotes "outside" relevant entity, "FOOD", "SYMPTOM" denote relevant food and symptom entity, respectively. However, this format requires the model to learn alignment between words and labels, increasing the difficulty of generating the label sequence. In contrast, the new output format design simplifies the model's task to merely mark the locations of the identified entities while replicating the rest of the text [12]. Originally, "@@" and "##" were used to denote extracted entities [12], but due to their common usage in social media, we use "^^" to avoid confusion.

In the task description's third sentence, we define the expected output format. Inspired by the CoT technique [11],

[2]The framework is adaptable and applicable to other advanced LLMs such as GPT-4 [25], Gemini [30] and Llama 2 [26].



(a) (b)

Fig. 2: Word-level labeling step. Left (2a): The Semantic Similarity example retrieval strategy, providing examples most semantically aligned with the query post. Right (2b): An example of word-level labeling prompt composed of three parts: task description, demonstration examples and query post.

we encourage reasoning before answering. As shown in the demonstration examples (Part 2, orange box) in Figure 2b, the model first assesses the post's overall relevance to a foodborne illness incident, then identifies relevant entities within the post. Upon reaching a conclusion, it is asked to format the output according to our specification. The concluding sentence of the task description signals that a few-shot demonstration examples will follow to guide the model's response. Part 2 includes these demonstration examples, and Part 3 (green box in Figure 2b) contains the query post to be labeled.

a) Semantic Similarity: Select Semantic Similar Examples to Enhance Word-level Labeling.: Given we can use only few examples for demonstration, it is critical we select examples that semantically relate to the query post to boost model performance [12], [31]–[34]. To accomplish this, we propose to deploy the *Semantic Similarity (SS)* retrieval strategy in step 1 to select the most representative examples from D^l ensuring quality demonstration (Figure 2a).

Definition 1 (SS). *Given an embedding model $\Phi_E(\theta)$, the labeled (example) set D^l and the unlabeled (query) set defined in III-A, let the semantic content of $x_i^l \in D^l$ and $x_i^u \in D^u$ be: $\mathbf{v}_i^l = \Phi_E(\mathbf{x}_i^l, \theta)$ and $\mathbf{v}_i^u = \Phi_E(\mathbf{x}_i^u, \theta)$, respectively. Then for all $x^u \in X^u$, SS example retrieval strategy with respect to x^u is defined as:*

$$I(x^u) = \text{argsort}_{\downarrow}\{\text{SimilarityScore}(\mathbf{v}^u, \mathbf{v}_i^l)\}_{i \in N^l} \quad (1)$$

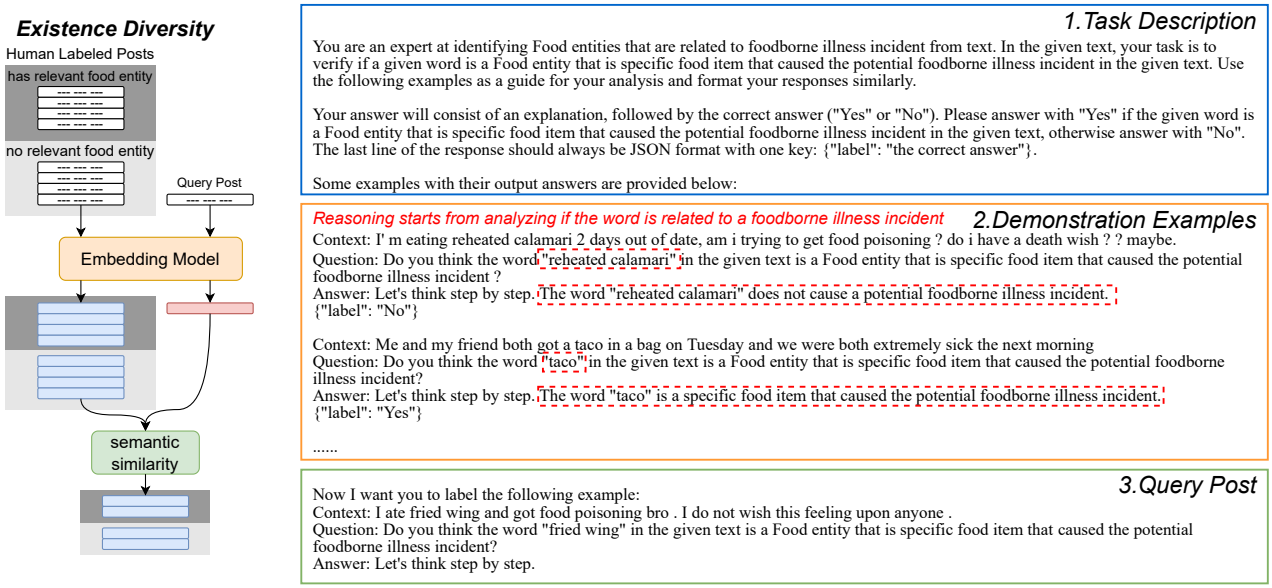
$$S_k(x^u) = \{(x_i^l, y_i^l, s_i^l) \mid i \in I(x)[1:k]\} \quad (2)$$

Remark 1. The SS method ensures that the subset of selected examples from D^l are those whose embedding vectors have the highest similarity score to the embedding vector $\mathbf{v}_q$ of the query post $\mathbf{x}_q$, thereby enriching the model's context for accurate labeling.

3) Step 2: Word Level Label Verification - Analyzing Identified Entity's Relevance Before Determining its Validity.: LLMs often face issues with overconfident prediction or hallucination [23]. This problem is exacerbated in our context of hierarchical labeling, representing *a vicious cycle of interdependencies between the two levels of our labeling hierarchy.* Inaccuracies of entity extraction as an intermediate step in the overall prediction process may adversely affect the next step's outcome.

Observation 1. A post may contain unrelated entities that are indistinguishable from related ones without first resolving the post-level decision. For example, the word "apple" in the sentence "I ate an apple and it tastes great!" might be incorrectly tagged as relevant due to its association with food, despite not being related to a foodborne illness.

The challenge here is that our task demands that the model identifies only entities specifically *related* to foodborne illness incidents. To break this vicious dependency cycle and prevent spurious entities from step 1 from affecting post-level predictions in step 3, we adopt a word-level label verification step to eliminate irrelevant entities extracted in the first step. The insight is that verification can be easier than generation. Intuitively, thinking of the first labeling step as solving an equation, then the verification step can be equated with substituting the solution back into the equation to check for correctness. The latter is an easier task than the former; improving the likelihood that it can be solved by the LLM. This ensures that more relevant entities are passed along as evidence to support



(a) (b)

Fig. 3: Word-level verification step. Left (3a): Existence Diversity example retrieval strategy, providing both positive (correctly extracted entities) and negative (incorrectly extracted entities) examples. Right (3b): An example of a word-level verification prompt composed of three parts: task description, demonstration examples, and the query post.

the final post-level label. In other words, this verification step reassesses the relevance of entities identified in the first step.

Figure 3b illustrates the design of realizing this solution, with our verification prompt structured into three sections similar to the ones in Step 1. Using the CoT method, the model is prompted to analyze whether the identified word is indeed a type of entity related to a foodborne illness incident, followed by a "yes" or "no" response. Entities verified as "yes" are retained, while those marked "no" are excluded. Figure 3b shows demonstration examples (in the orange box) for the food entity verification task retrieved from the labeled set D^l. Each example's question is based on its word-level labels using a standardized template: "Do you think the word ENTITY_WORD in the given text is a Food entity that is the specific food item that caused the potential foodborne illness incident?" Here, ENTITY_WORD is a placeholder. That is, if the post contains a relevant food entity, this entity is inserted as the ENTITY_WORD, and the corresponding response is "Yes". Conversely, if the post lacks a relevant food entity, a random text span from the post is used as ENTITY_WORD, and the answer is "No".

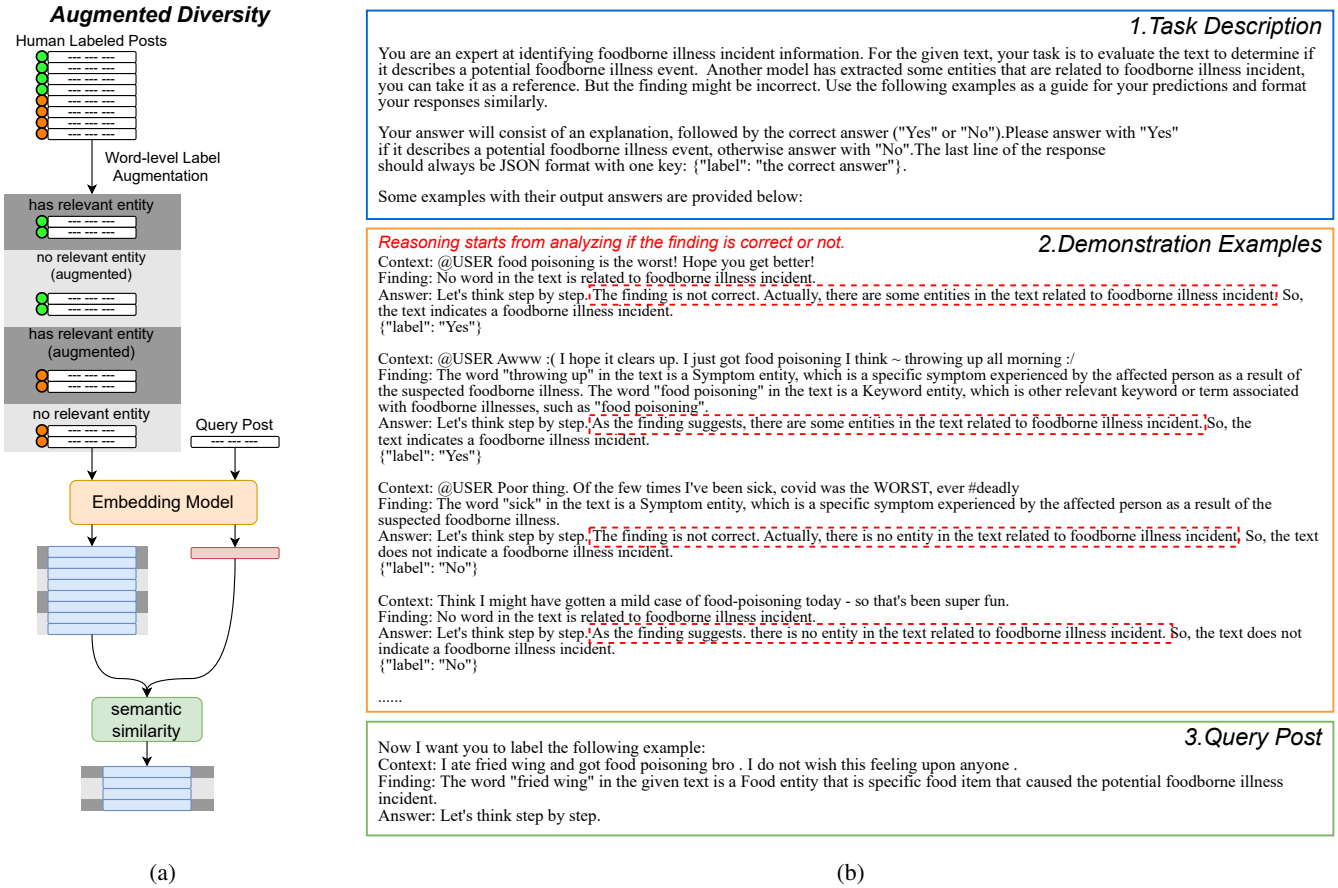
Existence Diversity: Present Both Confirming and Refuting Examples to Encourage Thorough Verification. We face the challenge that leveraging the above SS strategy for the verification step would risk selecting identical posts, potentially biasing the model towards *affirmative responses* if many examples were to mention food entities. To address this challenge, we design a custom example retrieval strategy for the verification step termed *Existence Diversity (ED)*.

Definition 2 (ED). *Given an entity type c_m, the labeled (example) set D^l defined in Section III-A is divided into to two subsets, $D^l_1 = \{\mathbf{x}^l_i \mid \exists s^l_{i,j} \in \mathbf{s}^l_i, s^l_{i,j} = \mathrm{c}_m\}$ and $D^l_0 = \{\mathbf{x}^l_i \mid \forall s^l_{i,j} \in \mathbf{s}^l_i, s^l_{i,j} \neq \mathrm{c}_m\}$. The examples are evenly retrieved from D^l_f and D^u_f using the SS methods.*

Remark 2. This ED design ensures the verification prompt includes a balanced yet diverse set of *both confirming and refuting examples* from D^l, thus supporting entity relevancy verification.

4) Step 3: Post Level Labeling - Determining the Post-Level Label from Scrutiny of Aggregated Word-Level Results: After the verification step, to collate evidence essential for this third step, we *aggregate the confirmed relevant entities* across the four categories into one single set per post, termed *aggregated evidence*. This evidence contains crucial information about the potential foodborne illness incident. However, even after verification, some irrelevant entities may remain. The aggregated evidence is leveraged for the construction of the prompt for the subsequent post-level labeling step. Here, the model is tasked with determining whether the post indicates a foodborne illness event. Figure 4b presents our prompt design for realizing this step. Utilizing the CoT method, the model scrutinizes the aggregated evidence (referred to as "findings" in prompt), assesses whether these entities indicate a foodborne illness, and then decides if the post describes such an incident. This reasoning process guides the model to reach the final post-level conclusion from word-level evidence scrutiny, with the flexibility to revise previously word-level assessments based on further analysis.

a) Augmented Diversity: Introducing Augmented Negative Examples for Comprehensive Post-Level Analysis.: All in-



(a) (b)

Fig. 4: Post-level labeling step. Left (4a): Augmented Diversity example retrieval strategy, flipping some examples' word level label to provide the model with both positive (word-level label result is correct) and negative (word-level label result is incorrect) examples. Right (4b): An example of post-level labeling prompt composed of three parts.

stances in our labeled set D^l are correctly labeled. Yet to train our model to discern between accurate and inaccurate word-level labels, it is crucial to introduce both correct (positive) and incorrect (negative) scenarios in demonstration examples. To provide more complete examples to the LLM, we design an example retrieval strategy named *Augmented Diversity (AD)*.

Definition 3 (AD). *The labeled set D^l from Section III-A is divided based on post-level labels into two groups: $D_1^l = \{\mathbf{x}_i^l | y_i^l = 1\}$ and $D_0^l = \{\mathbf{x}_i^l | y_i^l = 0\}$. Then, 50% of posts from both D_1^l and D_0^l are randomly augmented to become negative examples as follows:*

$$\mathbf{s}_{k,j}^l = c_r,\ c_r \xleftarrow{R} \{c_1, \ldots, c_4\},\ \mathbf{s}_{k,j}^l \xleftarrow{R} \mathbf{s}_i^l,\ \text{if } \mathbf{x}_i^l \in D_0^l \& \mathbf{I}_i^l = 1 \tag{3}$$

$$\mathbf{s}_{i,j}^l = c_5,\ \forall \mathbf{s}_{i,j}^l \in \mathbf{s}_i^l,\ \text{if } \mathbf{x}_i^l \in D_1^l \& \mathbf{I}_i^l = 1 \tag{4}$$

$$\mathbf{s}_{i,j}^l = \mathbf{s}_{i,j}^l,\ \text{otherwise} \tag{5}$$

Here, R denotes random selection, and $\mathbf{I}_i^l$ a binary indicator if the corresponding label $\mathbf{s}_i^l$ is augmented or not. After augmentation, demonstration examples are again selected based on the SS method.

Since we modified 50% of word-level labels, for each query post, roughly 50% of examples are negative cases. Augmented examples are depicted in Figure 4b. The first post incorrectly claims "food poisoning" as irrelevant to a foodborne illness, and the third post mislabels "sick" as a relevant keyword.

Remark 3. This AD strategy augments D^l with counterexamples to show potential errors in word-level findings, encouraging detailed scrutiny of word-level labels before post-level conclusion.

IV. Experimental Study

This section evaluates the effectiveness of our ICL2FID method on the TWEET-FID dataset [9] and compares it with various baseline methods. We also perform an ablation study to highlight the significance of each component within ICL2FID.

Social Media Dataset. We evaluate our method using the public-domain TWEET-FID dataset [9] consisting of 1,362 (33%) relevant and 2,760 (67%) irrelevant tweets related to foodborne illness. Each tweet features both expert labels and crowdsourced annotations per level. For more details on data collection, see [9]. The dataset is stratified and split into

training, validation, and testing sets to maintain a consistent ratio of relevant to irrelevant tweets as determined by expert labels. The training set consists of 3,298 tweets, while the validation and test sets each include 412 tweets.

For our LLM-based approach, the validation set with expert labels serves as the demonstration example set D^l. The training and test sets combined form the unlabeled set D^u, with expert labels used solely for study evaluation. To ensure the model's efficiency and reduce costs, we exclude tweets longer than 42 words — the third quartile of tweet lengths in our demonstration set — leaving 311 tweets for demonstration examples. This maintains label quality while reducing costs, as detailed in Section IV-B. All tweets in the unlabeled set D^u are kept to ensure annotation coverage.

Experimental Setup. We conducted our experiments with GPT-3.5-turbo [25], favored for its cost-effectiveness over GPT-4, and used the Text-embedding-ada-002-v2 [25] as the embedding model. GPT-3.5-turbo was set to a temperature of 0.1 for higher precision, as recommended by [35], with 8 demonstration examples per prompt. Section IV-D provides a detailed discussion on selecting the optimal number of examples. ICL2FID employs AutoLabel [36] for interaction with the LLM and embedding model via OpenAI API. We will release the code and details upon publication.

Performance at both word and post levels was measured using the F1 score and balanced accuracy (B.Acc) [37]. B.Acc is popular for assessing methods in imbalanced datasets. Due to budget constraints, we implemented a single run of all ICL-based methods using gpt-3.5-turbo and employed bootstrap resampling to estimate performance metrics.

Experimental Methodology. Our study evaluates a variety of established supervised learning (SL) models. These models are trained or fine-tuned exclusively on the demonstration example set D^l, matching the knowledge scope of our ICL-based solution. Our ICL2FID also draws demonstration examples solely from D^l. This alignment makes supervised methods and ICL-based approaches comparable. We assess whether our method is a feasible alternative for label collection by comparing its performance and costs against aggregated crowdsourced annotations. We also conduct an ablation study on ICL2FID to assess the impact of cross-level information prompting, the verification step, example retrieval strategies, exclusion of long examples, and the order of labeling steps.

Supervised Learning (SL) Methods. We compare with RoBERTa [38], BERTweet [39], and BiLSTM, each implemented as independent and joint versions. The independent version predicts at one of the two levels, while the joint version simultaneously predicts at both the word and post levels. These SL methods are either trained or fine-tuned on the entire demonstration example set.

Aggregated Human Annotation. We evaluate the quality and cost of ICL2FID's labels against crowdsourced annotations to determine if our method can match or surpass human quality cost-effectively. Label aggregation employs majority voting (MV) for both levels and Bayesian sequence combination (BSC) method for word-level labels due to its suitability for sequential tasks [9], [40].

Variants of ICL2FID. For our ablation study:

1) **ICL2FID-Independent (Ind).** Treats the word and post-level labeling steps as two independent tasks, not using the word-level labeling information to inform the post-level labeling. It omits the instruction for models to consider post-level relevance before making word-level predictions.
2) **ICL2FID w/o Step 2.** Removes the word-level verification step, using initial labeling results directly in the final step.
3) **ICL2FID w/ Extra Verification Step (EVS).** Adds one more word-level/post-level verification step after Step 3.
4) **ICL2FID w/ Semantic Similarity (SS) only.** Uses only the SS strategy for example retrieval, omitting our example selection strategies in Steps 2 and 3. No prompt to assess the accuracy of word-level labels in Step 3.
5) **ICL2FID w/ Random Retrieval (RR) only.** Uses only a random selection strategy for example retrieval for all steps. But still employs augmentation in Step 3.
6) **ICL2FID w/ All Labeled Data (ALD).** Utilizes all labeled data for analysis, including tweets longer than 42 words.
7) **ICL2FID -Reversed Order (RO).** Reverses the order: begins with post-level labeling and verification, then uses post-level results to inform word-level analysis.

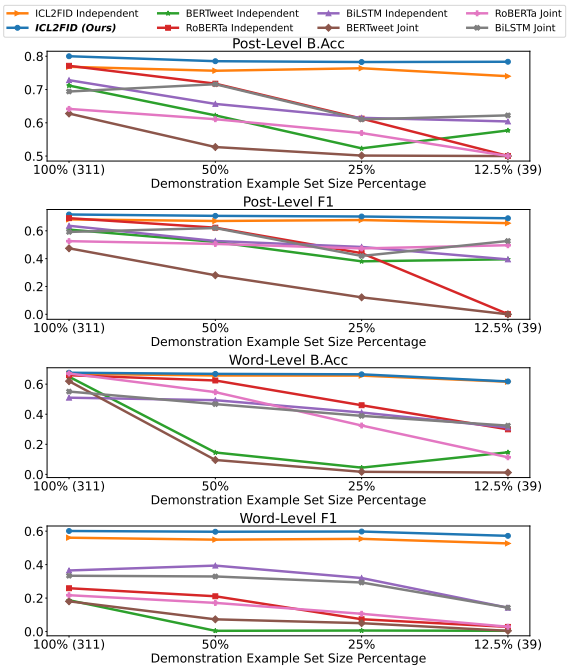


Fig. 5: Performance comparison against SOTA methods under varying sizes of the demonstration example collection from which a subset of examples is selected for each prompt.

A. Results of Comparative Study

Comparison of ICL2FID with SL Methods. Table II shows that ICL2FID surpasses all non-human models in both word

TABLE II: Performance comparison against SOTA methods on Tweet-FID. **Bold** are the highest, Underlined the second highest.

Learning	Model	Method	Word-Level F1	Word-Level B.Acc	Post-Level F1	Post-Level B.Acc
Supervised	BERTweet	Independent	0.1873 ± 0.0051	0.6496 ± 0.0043	0.6077 ± 0.0122	0.7113 ± 0.0076
Supervised	RoBERTa	Independent	0.2588 ± 0.0062	0.6594 ± 0.0051	0.6912 ± 0.0106	0.7715 ± 0.0075
Supervised	BiLSTM	Independent	0.3638 ± 0.0057	0.5101 ± 0.0042	0.6361 ± 0.0105	0.7280 ± 0.0069
Supervised	BERTweet	Joint	0.1808 ± 0.0002	0.6201 ± 0.0014	0.4742 ± 0.0102	0.6278 ± 0.0062
Supervised	RoBERTa	Joint	0.2172 ± 0.0055	0.6709 ± 0.0062	0.5252 ± 0.0118	0.6414 ± 0.0079
Supervised	BiLSTM	Joint	0.3336 ± 0.0068	0.5509 ± 0.0122	0.5911 ± 0.0112	0.6934 ± 0.0076
In Context	gpt-3.5-turbo	ICL2FID-Ind	0.5609 ± 0.0092	0.6693 ± 0.0114	0.6819 ± 0.0097	0.7682 ± 0.0060
In Context	gpt-3.5-turbo	**ICL2FID (Ours)**	**0.6010 ± 0.0088**	**0.6760 ± 0.0110**	0.7171 ± 0.0093	0.8000 ± 0.0058
Human	Crowdsourcing	MV	0.5908 ± 0.0146	0.6701 ± 0.0160	**0.7759 ± 0.0082**	**0.8515 ± 0.0051**
Human	Crowdsourcing	BSC	0.5414 ± 0.0141	0.6711 ± 0.0157	N/A	N/A

and post-level predictions. Unlike SL methods that require parameter updates to tune the pretrained language models, ICL2FID produces high-quality labels without a model learning or fine-tuning step. ICL2FID-Ind also shows superior performance compared to other SL methods, highlighting the capabilities of GPT-3.5-turbo. This advantage stems from the extensive pretraining on a larger dataset, which provides it with robust ICL abilities not present in traditional models.

Comparison of ICL2FID with Human Labelers. ICL2FID's performance nearly matches aggregated human labels and even exceeds word-level labels aggregated via both the BSC and the MV method. Cost-wise, while crowdsourcing labels costs approximately $0.50 per tweet, labeling with GPT-3.5-turbo is significantly cheaper $0.0005 to $0.001 per tweet, factoring in both input and output tokens as detailed in [25]. Additionally, ICL2FID can process labels within hours via the OpenAI API, faster than the days required for human effort. This efficiency and cost-effectiveness make ICL2FID a promising alternative for label generation, especially in resource-limited situations.

TABLE III: Ablation study of ICL2FID on Tweet-FID. **Bold** are the highest, Underlined are the second highest.

Method	Word-Level		Post-Level	
	F1	B.Acc	F1	B.Acc
ICL2FID	**.601 ± .009**	.676 ± .011	**.717 ± .009**	**.800 ± .006**
-Ind	.561 ± .009	.669 ± .011	.682 ± .010	.768 ± .006
w/o Step 2	.503 ± .009	**.685 ± .011**	.705 ± .010	.786 ± .007
w/ EVS	.596 ± .009	.655 ± .011	.629 ± .010	.715 ± .008
w/ SS only	.591 ± .009	.661 ± .011	.694 ± .010	.777 ± .007
w/ RR only	.499 ± .008	.440 ± .010	.710 ± .009	.785 ± .007
w/ ALD	.596 ± .009	.663 ± .011	.706 ± .010	.790 ± .006
-RO	.557 ± .009	.570 ± .012	.682 ± .012	.763 ± .008

B. Ablation Study of ICL2FID

As shown in Table III, ICL2FID-Ind emphasizes the value of linking post and word-level labels. The difference in word-level F1 scores with and without verification (Step 2) highlights its role in improving accuracy and reducing post-level errors. ICL2FID's slightly lower word-level B.Acc is due to filtering relevant entities. Additional verification steps don't improve performance. The SS-only variant underperforms ICL2FID, suggesting our ED and AD strategies reduce bias. The RR-only variant performs worse at the word level, but its post-level results remain comparable to ICL2FID due to augmented negative examples.

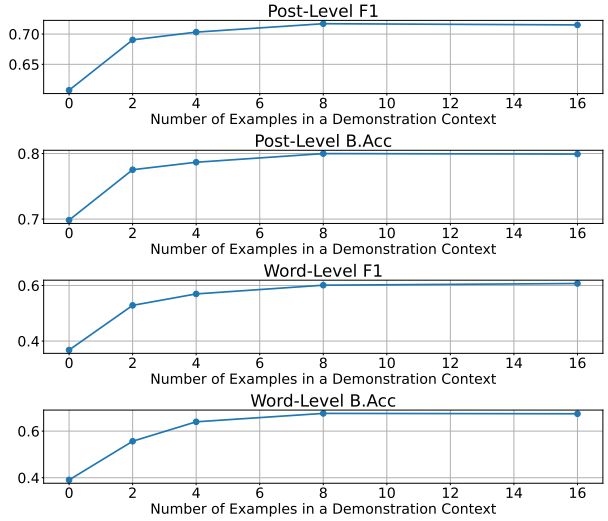


Fig. 6: Performance of ICL2FID under varying number of examples included in the demonstration context of a prompt.

Comparing ICL2FID with its variant using all labeled data shows that excluding long tweets reduces costs and improves performance. The difference in word-level F1 and B.Acc between ICL2FID and its reversed-order variant demonstrates the benefit of using word-level labels to inform post-level inference. In foodborne illness scenarios, multiple relevant entities in a post can guide accurate post-level labeling, even if some are missed. However, incorrect post-level predictions in the reversed-order variant can lead to errors in word-level labeling, causing false positives or missed entities.

C. Effect of Size of Full Demonstration Set

To assess ICL2FID's resilience with limited labeled data, we tested varying demonstration set sizes. Figure 5 shows ICL2FID maintains consistent performance across sets from 311 to 39 tweets, while traditional SL methods decline with smaller training sets. Although ICL2FID-Ind remains robust, it underperforms compared to ICL2FID. These results demonstrate ICL2FID's ability to efficiently generate high-quality labels in resource-limited settings with minimal labeled data.

D. Impact of Demonstration Example Quantity

Figure 6 shows ICL2FID's performance with varying numbers of examples in each prompt. Effectiveness improves on both word-level and post-level tasks as more examples are included, plateauing after 8 examples. Since these examples dominate the input context and adding more would increase input size significantly, 8 examples is a practical choice.

V. Conclusion and Future Work

In this work, we introduced ICL2FID, a novel hierarchical labeling framework leveraging LLMs to annotate posts for detecting foodborne illnesses. Key innovations include guided verification steps that reduce the errors that otherwise would be propagated between the stages of inference across the hierarchy, resulting in a series of LLM-based prompts that break the vicious cycle of two-way dependencies of the label hierarchy. Another critical ingredient of our solution is the example selection strategies customized to the three stages of ICL2FID, which efficiently retrieve a few posts from a minimal labeled set as demonstration examples, aligning with each stage's objective while avoiding repetitive data exposure, minimizing biases, and mitigating error propagation.

By leveraging the intricate relationship between the post- and word-level and countering model hallucination, ICL2FID outperforms existing methods with few selected examples, demonstrating the potential of ICL-based approaches over SL methods. Its labels closely rival crowd-sourced human annotations at a lower cost, proving ICL2FID's efficiency for label collection, especially in resource-limited settings. Future work could extend ICL2FID to other hierarchical labeling tasks and explore further integration of human expertise to enhance LLM performance in complex labeling scenarios.

References

[1] R. L. Scharff, "The economic burden of foodborne illness in the united states," *Food safety economics*, vol. 1, no. 1, pp. 123–142, 2018.

[2] D. Tao, D. Zhang, R. Hu, E. Rundensteiner, and H. Feng, "Crowdsourcing and machine learning approaches for extracting entities indicating potential foodborne outbreaks from social media," *Scientific reports*, vol. 11, no. 1, p. 21678, 2021.

[3] C. Harrison *et al.*, "Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—new york city, 2012–2013," *MMWR*, vol. 63, no. 20, p. 441, 2014.

[4] J. K. Harris *et al.*, "Health department use of social media to identify foodborne illness—chicago, illinois, 2013–2014," *MMWR*, vol. 63, no. 32, p. 681, 2014.

[5] J. P. Schomberg *et al.*, "Supplementing public health inspection via social media," *PloS one*, vol. 11, no. 3, p. e0152117, 2016.

[6] T. Effland, A. Lawson, S. Balter *et al.*, "Discovering foodborne illness in online restaurant reviews," *Journal of the American Medical Informatics Association*, vol. 25, no. 12, pp. 1586–1592, 2018.

[7] A. Sadilek, S. Caty *et al.*, "Machine-learned epidemiology: real-time detection of foodborne illness at scale," *NPJ digital medicine*, vol. 1, no. 1, p. 36, 2018.

[8] A. Sadilek *et al.*, "Deploying nemesis: Preventing foodborne illness by data mining social media," *Ai Magazine*, vol. 38, no. 1, pp. 37–48, 2017.

[9] R. Hu, D. Zhang, D. Tao, T. Hartvigsen, H. Feng, and E. Rundensteiner, "TWEET-FID: An annotated dataset for multiple foodborne illness detection tasks," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6212–6222.

[10] Q. Dong, L. Li *et al.*, "A survey on in-context learning," 2023.

[11] J. Wei, X. Wang *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *NeurIPS*, vol. 35, pp. 24 824–24 837, 2022.

[12] S. Wang, X. Sun *et al.*, "Gpt-ner: Named entity recognition via large language models," 2023.

[13] Z. Wan *et al.*, "GPT-RE: In-context learning for relation extraction using large language models," in *EMNLP*. ACL, Dec. 2023, pp. 3534–3547.

[14] H. Zhang, L. Cao, Y. Yan, S. Madden, and E. A. Rundensteiner, "Continuously adaptive similarity search," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 2601–2616.

[15] D. Hofmann, P. VanNostrand, H. Zhang, Y. Yan, L. Cao, S. Madden, and E. Rundensteiner, "A demonstration of autood: a self-tuning anomaly detection system," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3706–3709, 2022.

[16] H. Zhang, L. Cao, P. VanNostrand, S. Madden, and E. A. Rundensteiner, "Elite: Robust deep anomaly detection with meta gradient," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2174–2182.

[17] X. Sun *et al.*, "Text classification via large language models," in *EMNLP*. Singapore: ACL, Dec. 2023, pp. 8990–9005.

[18] C. Q. Zhu *et al.*, "Hierarchical multi-label classification of online vaccine concerns," 2024.

[19] J. He *et al.*, "Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction," in *ICCV*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2023, pp. 19 428–19 437.

[20] S. Sia and K. Duh, "In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models," in *Proceedings of Machine Translation Summit XIX*. Macau SAR, China: AAMT, Sep. 2023, pp. 173–185.

[21] P. Mirza *et al.*, "Illuminer: Instruction-tuned large language models as few-shot intent classifier and slot filler," 2024.

[22] H. Nguyen *et al.*, "CoF-CoT: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain NLU tasks," in *EMNLP*. Singapore: ACL, Dec. 2023, pp. 12 109–12 119.

[23] L. Huang, W. Yu *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," 2023.

[24] W. X. Zhao, K. Zhou *et al.*, "A survey of large language models," 2023.

[25] OpenAI, "Models," https://platform.openai.com/docs/models, 2023.

[26] H. Touvron, L. Martin, K. Stone, P. Albert, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[27] T. Brown *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[28] J. K. Harris *et al.*, "Health department use of social media to identify foodborne illness—chicago, illinois, 2013–2014," *MMWR*, vol. 63, no. 32, p. 681, 2014.

[29] ——, "Research brief report: using twitter to identify and respond to food poisoning: The food safety stl project," *Journal of Public Health Management and Practice*, vol. 23, no. 6, p. 577, 2017.

[30] M. Reid *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024.

[31] J. Liu *et al.*, "What makes good in-context examples for GPT-3?" in *DeeLIO*. Dublin, Ireland and Online: ACL, May 2022, pp. 100–114.

[32] D. V. Torres, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster, "Prompting palm for translation: Assessing strategies and performance," in *ACL*. Toronto, Canada: ACL, 2023, p. 15406–15427.

[33] H. Zhang, L. Cao, S. Madden, and E. Rundensteiner, "Lancet: labeling complex data at scale," *Proceedings of the VLDB Endowment*, vol. 14, no. 11, 2021.

[34] H. Zhang, B. Yan, L. Cao, S. Madden, and E. Rundensteiner, "Metastore: Analyzing deep learning meta-data at scale," *Proceedings of the VLDB Endowment*, vol. 17, no. 6, pp. 1446–1459, 2024.

[35] Refuel AI, "Guide to large language models in autolabel," https://docs.refuel.ai/autolabel/guide/llms/llms/, 2023.

[36] ——, "Autolabel," https://docs.refuel.ai/autolabel/introduction/, 2023.

[37] R. Wang *et al.*, "Learning to adapt classifier for imbalanced semi-supervised learning," 2022.

[38] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019.

[39] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *EMNLP*, Q. Liu and D. Schlangen, Eds. Online: ACL, Oct. 2020, pp. 9–14.

[40] E. Simpson *et al.*, "A bayesian approach for sequence tagging with crowds," in *EMNLP-IJCNLP*. Hong Kong, China: ACL, nov 2019, pp. 1093–1104.