Prompted Weak Supervision for Sentence Classification in Civic Transcripts

Pak Iong Long
Brown University
Providence, RI
pak_iong_long@brown.edu

Abstract

This project addresses the challenge of training effective classifiers under conditions of limited labeled data by leveraging weak supervision with large language models (LLMs). We design a pipeline that generates sentence-level labels for a binary classification task—identifying whether a sentence discusses financial planning—using prompt-based heuristics and a label model to aggregate noisy supervision signals. Built on the Alfred framework (Yu and Bach, 2023), our pipeline integrates multiple prompted label functions, custom regex matchers, and probabilistic label modeling to reduce labeling noise and improve downstream performance. We train end models on both majority-vote and partial label model-derived probabilistic labels and evaluate them against zero-shot LLM baselines. Our results show that all trained models outperform the zero-shot baseline, with the best configuration achieving a 63.7% F1 score compared to 57.9% from the baseline. This work demonstrates the effectiveness of prompted weak supervision for constructing high-quality training signals without human annotation and highlights Alfred's extensibility for real-world text classification tasks.

1 Introduction

Automatically classifying whether a sentence pertains to financial planning is a nuanced task, especially when annotated data is scarce. Financial planning discussions often rely on implicit context and subtle linguistic cues, making it difficult for traditional keyword-based or rule-based classifiers to distinguish relevant content from general municipal discourse. This complexity is compounded in real-world data sources such as city council transcripts, where topics are diverse and sentence structure is often informal.

The problem is both important and interesting: improving fine-grained understanding of civic financial planning can support better transparency, analytics, and decision-making tools for public institutions and researchers. However, the limited availability of high-quality labeled data creates a bottleneck for supervised learning methods.

Naive approaches—such as basic phrase-matching or zero-shot prompting—fail to capture the subtle semantic variation present in this domain. Existing solutions either depend on large hand-labeled corpora, which are expensive to curate, or rely on zero-shot LLM inference, which yields inconsistent results and cannot be tuned for downstream tasks. Prior work in weak supervision has shown promise in low-label settings (Smith et al., 2024). We extend this by applying weak supervision to sentence-level classification tasks.

In this work, we propose a weak supervision pipeline that integrates prompt-engineered label functions with a probabilistic label model to generate training data for an end classifier. Built on top of the Alfred framework, our approach includes[1]:

- A curated set of prompt templates designed for precision and recall trade-offs.
- Regex-based heuristic labelers to provide complementary signals.
- A partial label model to resolve conflicts and produce soft labels.
- Fine-tuning of RoBERTa classifiers on generated labels, with and without focal loss to account for class imbalance.

Our pipeline improves performance over both majority vote baselines and zero-shot LLM classifiers. While our method effectively boosts accuracy in a weakly supervised setting, it is not without limitations. The design and tuning of prompt templates require domain intuition and iterative refinement. Additionally, performance gains are

[1] https://github.com/BatsResearch/alfred-meeting-bank

modest, and in tasks with higher variance or noisier domains, the label model may propagate ambiguity. Nonetheless, our results suggest that prompting, when integrated thoughtfully into a weak supervision workflow, offers a scalable path forward in low-resource classification tasks.

2 Methodology

Our methodology centers on building a weak supervision pipeline that generates high-quality training labels from unlabeled city council meeting transcripts (Hu et al., 2023). The goal is to train an accurate binary classifier for detecting financial planning discussions—without relying on manually labeled training data.

To achieve this, we implemented a four-stage process:

1. Prompt-Based Label Function Design
2. Heuristic Matching and Vote Aggregation
3. Probabilistic Label Modeling
4. End Model Training on Weak Labels

Each stage is designed to introduce modular, extensible components that balance domain expertise (via heuristic templates) and model-driven generalization (via probabilistic aggregation and fine-tuning).

2.1 Prompt-Based Label Function Design

We began by designing a set of prompt templates to query an LLM for binary classification labels. Each prompt was crafted to maximize precision, recall, or coverage, and returned a Boolean label indicating whether the input sentence discusses financial planning. To increase coverage, we diversified the prompts across three categories:

- Label Prompts that directly elicit a yes/no response.
- Positive Heuristic Prompts designed to match likely positive cases.
- Negative Heuristic Prompts targeting non-planning content.

These templates were iteratively refined by evaluating their individual performance on a labeled development set (800 example sentences), using their precision and recall scores as primary metrics. This ensured that only templates contributing signal were retained. Ultimately, 13 prompt templates were selected (3 label, 5 positive heuristic, 5 negative heuristic).

All prompts were executed using the Mistral 7B Instruct V0.2 model (Jiang et al., 2023), an open-weight instruction-tuned LLM that offers competitive zero and few-shot reasoning capabilities at relatively low inference cost. We selected Mistral for its balance of performance and accessibility, particularly given the need to scale labeling over thousands of sentences. Compared to larger proprietary models (e.g., GPT-4), Mistral demonstrated strong alignment on simple binary tasks with less prompt engineering, making it a practical choice for this domain-specific classification pipeline.

2.2 Heuristic Matching and Vote Aggregation

To complement prompt-based signals, we constructed regex-based matchers extracting the prediction from LLM responses. Each label map, combined with regex logic, acted as a labeling function, outputting one of 0, 1, -1, where -1 indicates abstention due to low confidence or ambiguity.

These labeling functions voted on every sentence in the dataset. We constructed a 13×N voting matrix, where each row represented a label function and each column a sentence. This matrix served as the input to both baseline label aggregation (majority vote) and probabilistic modeling (below).

2.3 Probabilistic Label Modeling

To address label noise and conflicting signals, we trained a partial label model on the voting matrix using the Noisy Partial Labelm Model (NPLM) library(Yu et al., 2022). This model estimates the accuracy and correlation of each labeling function and outputs soft probabilistic labels per sentence. This step is essential to denoise weak labels, particularly when prompt performance varies significantly across contexts.

We used the development set to validate that the partial label model's output correlated well with true labels (F1 = 61.5%), surpassing majority vote aggregation (F1 = 56.9%). Once validated, the model was applied to the 8,000 unlabeled training sentences to produce two sets of training labels: majority vote and probabilistic labels.

2.4 End Model Training

To convert weak labels into a high-performance classifier, we fine-tune RoBERTa-based models (Liu, 2019) on the outputs of our label aggregation

pipeline. We chose RoBERTa because of its strong performance on sentence-level classification tasks, robustness to noisy supervision, and its pre-training on a diverse corpus that makes it well-suited to municipal and policy-related text. Its transformer architecture also enables effective fine-tuning with limited supervision, making it a good fit for our low-label setting.

Using the generated weak labels, we trained four RoBERTa-based classifiers:

Two on majority-vote labels (with and without focal loss),

Two on probabilistic labels (with and without focal loss).

Focal loss was included to handle mild class imbalance. All models were trained for 10 epochs using a batch size of 32 and a learning rate of 1e-5.

This modular architecture allows us to compare the effects of labeling strategy (majority vs. probabilistic) and loss function (standard cross entropy vs. focal) independently, forming the basis for our experiments in Section 3.

3 Experimentation

We empirically evaluate the effectiveness of our prompted weak supervision pipeline in training accurate sentence-level classifiers for identifying financial planning content. Our experiments assess three dimensions: (1) the quality of generated labels, (2) the performance of downstream classifiers trained on these labels, and (3) the contribution of individual components via ablation and robustness analysis.

3.1 Experimental Setup

We use a corpus of 1,366 city council meeting transcripts from the meetingbank dataset (Hu et al., 2023). Each transcript was sentencized because we wanted the classification to be at the sentence level instead of an entire transcript. Only unique sentences were retained, and 800 examples were manually labeled to serve as a development set. The training set consists of 8,000 unlabeled sentences.

We evaluate models using the F1 score and report standard errors from five training runs with different random seeds (results shown in Table 5). All models are based on RoBERTa and are trained with 10 epochs, a batch size of 32, and a learning rate of 1e-5.

We adapted an existing open-source script for RoBERTa fine-tuning (fin), which provided a standardized pipeline for training and evaluation. This ensured consistency across runs and allowed us to focus on the effects of label quality and loss functions rather than implementation details. We modified the script to support both cross-entropy and focal loss.

3.2 Label Quality Evaluation

We first assess the quality of weak labels produced by our voting framework. Table 2 compares label agreement on the development set between:

- Majority vote aggregation.
- Partial label model probabilistic labeling.

The partial label model yields stronger alignment with human-labeled ground truth (F1 = 61.5%) compared to majority voting (F1 = 56.9%), supporting its use in end-model training.

Voting Technique Baseline	F1
Majority Vote	56.9%
Partial Label Model	61.5%

Table 2: Voting Technique Baseline.

3.3 End Model Performance

We train four RoBERTa classifiers on the generated weak labels, evaluating the impact of label type and loss function. Results are summarized in Table 3.

All models outperform a zero-shot LLM baseline (F1 = 57.9%), confirming the value of weak supervision in this context. Models trained on majority vote labels performed marginally better than those trained on probabilistic labels, though differences are small.

Voting method	Focal loss	Without focal loss
Majority vote	63.7%	64.7%
Partial label model	63.7%	63.0%

Table 3: The accuracies of four end models with the respective loss functions and data used to train.

3.4 Baseline Comparison

To contextualize our results, we compare against two baseline strategies:

- Zero-Shot Baseline: The average F1 score (57.9%) across three LLMs, each prompted with a distinct label function.

	Partial Label Model	Zero-Shot Baseline	Model Accuracy
F1 Score	61.5%	57.9%	63.7%
Std Error	$\pm$0.3	$\pm$12.3	$\pm$0.2

Table 1: Performance comparison of the partial label model, zero-shot baseline, and the trained end model in terms of accuracy.

- Aggregated Prompt Baseline: A majority-vote ensemble of label prompts used directly to train a model (F1 = 61.0%).

Our full pipeline surpasses both baselines, demonstrating the added value of integrating prompted weak supervision with probabilistic modeling and fine-tuned classifiers. The direct comparison of our model accuracy against partial label model and zero-shot baseline can be found in Table 1.

Model Baseline	F1
Zero-Shot Baseline	57.9%
Aggregated Baseline	61.0%

Table 4: Zero-shot baseline accuracies.

3.5 Ablation Study

To understand the contribution of individual components, we conduct an ablation study by varying the loss function used during training. The results indicate that when training a model with probabilistic labels generated from a partial label model, incorporating focal loss provides a slight advantage. We attribute this improvement to the class imbalance present in our dataset. However, for this specific task, the difference in performance is not statistically significant.

The integration of labeling, majority voting, and probabilistic modeling produced reliable classification results, demonstrating the effectiveness of the combined methods in identifying financial planning-related content.

3.6 Validation of Results

To validate the consistency of our end model as well as our zero-shot baseline we conducted an experiment using five different random seed values during model training. Our results in Table 5 shows that the average F1 score of the end model over the five random seed values is 63.6% with standard error of 2.0. The average F1 score of the zero-shot baseline where we take the average of three individual end models is 55.6% with standard error of 0.7.

Our experimentation indicate that prompted weak supervision proves to be an effective strategy for generating high-quality labels without the need for manual annotation. End models trained on these labels consistently outperform both zero-shot baselines and those using aggregated prompts. While probabilistic labeling offers a slight improvement in label quality over simple majority voting, downstream task performance remains largely similar. Additionally, the system demonstrates robustness to random initialization and sees modest gains when using tailored loss functions.

4 Future Work

While our results demonstrate the promise of prompted weak supervision for low-resource classification, several directions remain for future exploration. First, prompt design remains a manual and iterative process, integrating automatic prompt generation or reinforcement learning-based prompt selection could further improve label quality and reduce engineering overhead. Second, while we evaluated only one open-weight LLM (Mistral 7B), testing across a wider range of models (including proprietary models like GPT-4 or Claude) could reveal trade-offs between label fidelity, cost, and reproducibility.

Additionally, our pipeline was applied to a single binary classification task in the civic domain. Future work could extend this approach to multi-label or multi-class tasks, or to domains with noisier text (e.g., social media or transcribed speech). Finally, a deeper analysis of labeling function conflicts and label model uncertainty could improve the interpretability and trustworthiness of weak supervision outputs in sensitive or policy-facing applications.

5 Conclusion

This work presents a prompted weak supervision pipeline for sentence-level classification under low-label conditions, applied to the task of identifying financial planning discussions in municipal tran-

scripts. By combining prompt-engineered label functions, regex heuristics, and probabilistic aggregation using a partial label model, we generate high-quality training labels without human annotation. We show that end models trained on these weak labels consistently outperform both zero-shot LLM baselines and direct prompt aggregation strategies.

Our findings demonstrate that integrating prompting as a first-class component in a weak supervision framework is both effective and scalable. The use of open-weight models like Mistral 7B makes this approach reproducible and accessible for practical deployments. While prompt crafting and label model tuning require domain intuition, the overall pipeline offers a viable solution for low-resource classification tasks where manual annotation is impractical.

Looking ahead, this work opens opportunities for broader applications of prompted weak supervision across domains and tasks, and for improving automation in label generation through smarter prompting and model selection.

Acknowledgements

I would like to thank Peilin Yu and Stephen Bach for advising this project.

References

Fine tuning roberta for sentiment analysis.

Yebowen Hu, Tim Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. *arXiv preprint arXiv:2305.17529*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM/JMS Journal of Data Science*, 1(2):1–30.

Peilin Yu and Stephen H Bach. 2023. Alfred: A system for prompted weak supervision. *arXiv preprint arXiv:2305.18623*.

Peilin Yu, Tiffany Ding, and Stephen H. Bach. 2022. Learning from multiple noisy partial labelers. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11072–11095. PMLR.

A Appendix: Supplementary Figures

Random Seed Value	F1
10	67.2%
20	65.0%
30	61.4%
40	56.6%
50	67.8%
avg	**63.6%**
std err	±**2.1**

Table 5: We conducted a random seed experiment ensuring consistency of the end model trained with different random seed values.

Random Seed Value	Label Prompt 0 F1	Label Prompt 1 F1	Label Prompt 2 F1	Average F1
10	54.0%	67.3%	41.9%	54.4%
20	60.5%	72.6%	40.6%	57.9%
30	52.4%	71.6%	38.0%	54.0%
40	57.6%	64.7%	42.6%	55.0%
50	55.0%	71.5%	43.8%	56.8%
avg	55.9%	69.5%	41.4%	**55.6%**
std err	±1.4	±1.5	±1.0	±**0.7**

Table 6: We conducted a random seed experiment ensuring consistency of the baseline with different random seed values.

Type	Topic	Prompt Template
label	government financial keywords	"Instruction: Answer with 'True' or 'False' for the following question." "Sentence: [[text]]" "Question: Does the sentence contain any government related financial keywords?" "Answer: "
label	financial planning keywords	"Instruction: Answer with 'True' or 'False' for the following question." "Sentence: [[text]]" "Question: Identify if the sentence includes key terms related to financial planning such as budget, fiscal policy, financial forecast, or investment planning." "Answer: "
label	financial actions	"Instruction: Answer with 'True' or 'False' for the following question." "Sentence: [[text]]" "Question: Does the sentence reference financial actions or decisions made by a government body?" "Answer: "
positive	contextual analysis	"Instruction: Review the above sentence from a city council transcript. Analyze the context of the discussion in the sentence provided. Does it explicitly relate to managing finances, planning budgets, or allocating resources? Provide only a 'True' or 'False' answer. Answer:"
positive	decision making focus	"Instruction: Review the following sentence from a city council transcript. Evaluate the sentence for discussions on financial decision-making or strategic financial actions. Is there any mention of decision-making related to finance? Provide only a 'True' or 'False' answer. Sentence: [[text]] Answer:"

positive	policy and procedures	"Instruction: Review the following sentence from a city council transcript. Does this sentence discuss policies or procedures that involve financial planning, such as budget approvals or financial audits? Provide only a 'True' or 'False' answer. Sentence: [[text]] Answer:"
positive	outcome based inquiry	"Instruction: Review the following sentence from a city council transcript. Consider the outcomes or goals mentioned in the sentence. Do they involve financial planning or fiscal management? Provide only a 'True' or 'False' answer. Sentence: [[text]] Answer:"
positive	event planning focus	"Instruction: Review the following sentence from a city council transcript. Look for any mention of planning for events or projects that require budgeting or financial resources. Is there any financial planning for events mentioned? Provide only a 'True' or 'False' answer. Sentence: [[text]] Answer:"
negative	public safety and law enforcement	"Instruction: Review the following sentence from a city council transcript. Does the sentence describe any new initiatives related to public safety or law enforcement? Sentence: [[text]] Answer:"
negative	education and local schools	"Instruction: Review the following sentence from a city council transcript. Analyze the sentence for any mention of local education policies, school funding, or educational programs. Is there a discussion about educational matters? Sentence: [[text]] Answer:"
negative	environmental and green initiatives	"Instruction: Review the following sentence from a city council transcript. Check if the sentence covered topics related to environmental protection or green initiatives. Did the sentence discuss environmental concerns? Sentence: [[text]] Answer:"
negative	public utilities and services	"Instruction: Review the following sentence from a city council transcript. Does the sentence discuss issues related to public utilities like water, power, or waste management? Please confirm if such topics were covered. Sentence: [[text]] Answer:"
negative	housing and real estate development	"Instruction: Review the following sentence from a city council transcript.Review the sentence for any discussions on housing policies, real estate development, or zoning regulations.Were housing issues discussed?Sentence: [[text]]Answer:"""

Table 8: Label and heuristic prompts used to produce the labels for end model training.