# Trustworthy Classification for Complex Social Surveys: A Memory-Enhanced Hierarchical Framework with Calibrated Uncertainty

**Zeqiang Wang[1], Rebecca Oldroyd[2], Yuqi Wang[3], Jiageng Wu[4], Jie Yang[5], Wei Wang[6], Nishanth R. Sastry[1], Jon Johnson[2], Suparna De[1]***

[1]University of Surrey
[2]University College London (UCL)
[3]Shanghai Jiao Tong University
[4]Harvard Medical School and Brigham and Women's Hospital
[5]Harvard University
[6]Xi'an Jiaotong–Liverpool University
zeqiang.wang@surrey.ac.uk, s.de@surrey.ac.uk

## Abstract

Automated classification of complex social survey questionnaires is crucial for large-scale social science research but faces significant reliability challenges due to intricate hierarchical label structures, severe class imbalance, semantic ambiguity, and incomplete data coverage. Conventional classification methods often struggle with these combined complexities, yielding results that lack trustworthiness. We introduce HOCM, a framework designed for trustworthy classification in complex, real-world taxonomies. It features two synergistic components: (1) memory-enhanced contrastive learning, tailored to learn robust representations from noisy, imbalanced data by leveraging quality-aware category memory banks; and (2) hierarchical uncertainty calibration, which enforces taxonomic consistency while providing reliable confidence estimates and identifying inputs falling outside well-represented known categories. Our evaluation on a large-scale, real-world social survey dataset—a challenging exemplar of our target problem class—demonstrates that HOCM maintains strong accuracy on known classes while effectively identifying uncertain cases, significantly boosting accuracy on confident predictions. Furthermore, it adeptly detects low-resource/unknown categories. HOCM provides a more reliable automated classification tool, enabling efficient expert review and enhancing the trustworthiness of analysis in domains with complex, hierarchical data.

## Introduction

Timely and effective social policy relies on analyzing data from diverse sources, such as longitudinal population surveys (LPS) and national censuses. A fundamental challenge for researchers and data archives (e.g., the Consortium of European Social Science Archives, CESSDA) is harmonizing this data, which requires mapping thousands of survey questions to standardized classification taxonomies. This process of creating structured metadata is essential for data discoverability and reuse, yet it is traditionally a manual, slow, and expensive endeavor that struggles to keep pace with the volume of new data.

Automating this classification is therefore a critical step towards enhancing the scalability and rigor of computational social science. However, real-world survey data presents a suite of concurrent challenges that render standard classification models unreliable. These challenges, which motivate our work, include:

- **Complex Hierarchy and Severe Class Imbalance:** Social science research employs fine-grained, multi-level classification ontologies. Data distribution across these categories is typically highly imbalanced, with many specific sub-categories being severely under-represented.
- **Semantic Ambiguity and Category Overlap:** Questionnaire items frequently span multiple thematic categories (e.g., work-related stress affecting sleep intersects both Employment and Health), necessitating reliable confidence estimates.
- **Incomplete Data Coverage and Openness:** Real-world annotation seldom achieves exhaustive coverage, creating an open-set problem where models must handle both under-represented knowns and truly out-of-scope content.

Traditional text classification models falter when confronted with these combined challenges. Many are designed for flat label spaces, struggle with imbalance, and lack mechanisms to identify inputs outside their knowledge boundary. While motivated by social survey analysis, these challenges represent a general class of hierarchical classification problems characterized by severe imbalance, semantic ambiguity, and open-set dynamics—applicable across domains from bioinformatics to product taxonomy management.

To address these broadly relevant challenges, we propose **HOCM** (Hierarchical Open-set Classification with Memory-enhanced learning), a framework designed to provide trustworthy classification through two synergistic components (Figure 1). Our main contributions are:

- A unified framework designed to synergistically address the concurrent challenges of hierarchical structure, severe data imbalance, and open-set dynamics in a single, integrated model.

*Corresponding author.

- A quality-aware memory-contrastive learning approach with novel prototype management mechanisms. Its core innovation lies in combating representation instability caused by noisy and low-resource classes, a critical problem that standard contrastive methods often fail to address.
- A taxonomy-aware uncertainty calibration method that uniquely applies tail distribution modeling at every node in the hierarchy and enforces structural consistency through recursive normalization, providing more reliable outlier detection than flat or non-hierarchical approaches.

We perform extensive empirical validation on a complex, real-world dataset, demonstrating HOCM's ability to improve classification reliability and effectively guide expert review.

## Related Work

### Hierarchical Text Classification

HTC methods leverage label hierarchies, with early work including Hierarchical Attention Networks (Yang et al. 2016). Subsequent research improved hierarchical predictions through hierarchical constraints, label imbalance techniques like batch calibration (Zhou et al. 2023), and contrastive learning (Zheng, Chen, and Huang 2020). However, many HTC methods focus primarily on closed-set accuracy without addressing robust uncertainty estimation or explicit open-set rejection—both critical for real-world deployment on noisy survey data.

### Open-Set Recognition and Uncertainty Quantification

OSR addresses classifying known categories while rejecting unknown ones. Foundational approaches include OpenMax (Bendale and Boult 2016) using Extreme Value Theory, dedicated detectors (Shu, Xu, and Liu 2017), distance-based methods (Zhou, Liu, and Qiu 2022), adversarial training (Chen et al. 2024), and baselines like MSP (Hendrycks and Gimpel 2022) enhanced with ODIN (Liang, Li, and Srikant 2018). Recent work focuses on calibration and representation learning for OSR/OOD detection (Wei et al. 2022; Ming et al. 2022; Dai et al. 2023).

Most text OSR methods target flat classification without exploiting hierarchical structures. They assume clean, balanced data, potentially struggling with severe imbalance and semantic ambiguity in survey data.

Uncertainty quantification ensures model confidence aligns with correctness (Guo et al. 2017). While Bayesian networks (Novello, Dalmau, and Andeol 2024) and conformal prediction (Kaur et al. 2022) offer principled frameworks, integrating reliable uncertainty estimates within hierarchical open-set text classification remains challenging.

We note that alternative frameworks for uncertainty quantification, such as Conformal Prediction (CP), offer formal statistical guarantees on prediction sets. While powerful, applying CP effectively to deep hierarchical classification settings with severe class imbalance and open-set assumptions presents significant practical and methodological challenges that are an active area of research. Our approach, by adapting tail distribution modeling to the hierarchy, provides a pragmatic and effective empirical solution tailored to our specific problem context, as demonstrated by the strong performance gains in our reliability-aware evaluation.

### Memory Mechanisms in NLP

Memory mechanisms range from early memory networks (Weston, Chopra, and Bordes 2014; Sukhbaatar et al. 2015) to recent retrieval augmentation (Liu et al. 2024) and dynamic generation (Jain et al. 2024).

HOCM's memory differs significantly—it maintains stable, high-quality prototypes through selective storage and aggregation. Quality-aware filtering (using interquartile range) and reliability-guided sampling mitigate class imbalance and noisy data effects, fostering robust and discriminative features for both closed-set classification and open-set detection.

### Positioning HOCM

While previous works address HTC, OSR, UQ, or memory mechanisms separately, HOCM's novelty lies in the synergistic integration and problem-specific adaptation of these components to holistically address the concurrent challenges of survey classification. Specifically: (1) Our Quality-Aware Memory Management (with IQR filtering and reliability-guided sampling) is uniquely designed to build stable class prototypes from the noisy and imbalanced data characteristic of surveys, a critical issue unaddressed by standard methods. (2) Our Hierarchical Uncertainty Calibration significantly extends flat open-set methods (like OpenMax) by modeling and propagating uncertainty estimates through the entire label hierarchy with recursive normalization. This is a crucial step for providing coherent and granular outlier detection in a taxonomic structure, and our ablation study shows it substantially outperforms a flat approach.

## Methodology

We introduce **HOCM**, a framework designed for reliable hierarchical open-set classification. It integrates two core components: (1) memory-enhanced contrastive learning for robust representation, and (2) hierarchical uncertainty calibration for trustworthy prediction. The entire process is summarized in Algorithm 1.

The framework begins by using a pre-trained encoder (e.g., BERT) to obtain an initial embedding $z$ for an input text $x$. This embedding is then refined through our proposed modules.

### Memory-Enhanced Contrastive Learning

We introduce **HOCM**, a framework designed for reliable hierarchical open-set classification. It synergistically integrates two novel components: memory-enhanced contrastive learning and hierarchical uncertainty calibration. The overall process is depicted in Figure 1 and summarized in Algorithm 1.

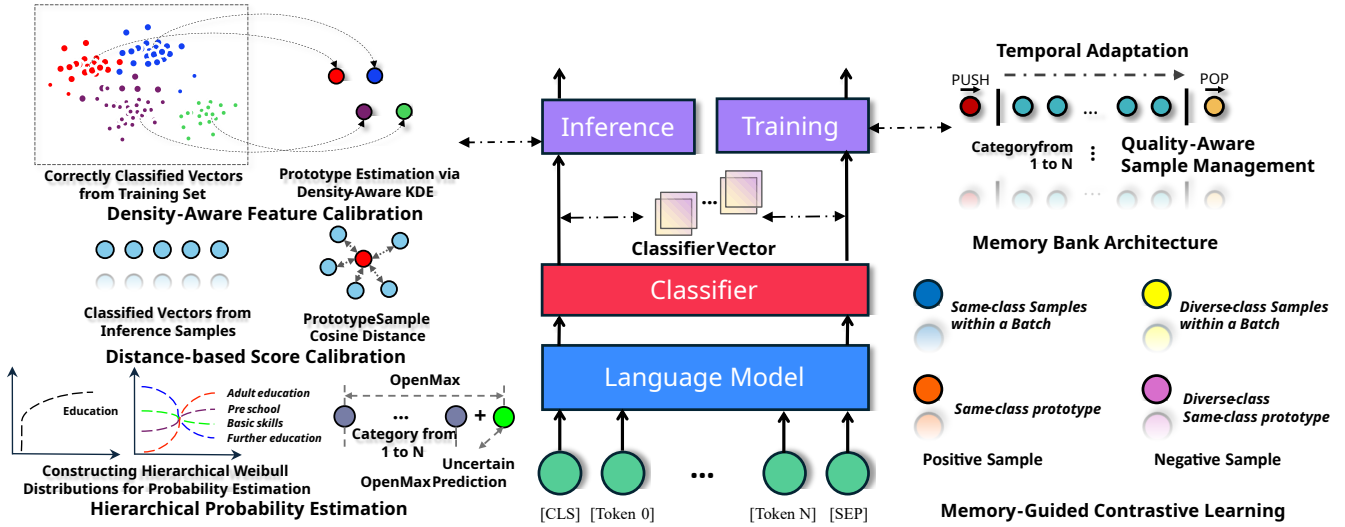


Figure 1: Overview of our proposed framework, HOCM. A pre-trained language model extracts initial features. These are refined by the **Memory-Enhanced Contrastive Learning** module, which uses quality-aware memory banks and adaptive sampling to learn robust category representations, especially under data imbalance. The subsequent **Hierarchical Uncertainty Calibration** module computes uncertainty scores based on distance to prototypes at all levels of the known hierarchy using tail distribution modeling (e.g., Weibull fitting). It then combines these scores and enforces hierarchical consistency via recursive normalization to produce reliable closed-set predictions and effectively detect open-set inputs (under-represented knowns and true unknowns), flagging uncertain cases for review.

**Memory-Enhanced Contrastive Learning** To directly tackle the representation instability caused by severe class imbalance and noisy data, this module learns robust features using category-specific memory banks. For each known class $c \in \mathcal{C}_{known}$, we maintain a memory bank $\mathbf{M}_c$ that stores high-quality instance embeddings.

**Quality-Aware Memory Management** Instead of naively storing all sample embeddings, we employ a dynamic management strategy. A composite quality score—assessing semantic similarity, statistical typicality (via Mahalanobis distance), and recency—guides the inclusion of new embeddings. An Interquartile Range (IQR) based filter on similarity scores further prunes low-quality outliers. This ensures the stability of the category prototype $\mathbf{p}_c$, which is computed as a quality-weighted average of the embeddings in $\mathbf{M}_c$.

**Reliability-Guided Contrastive Learning** These robust prototypes are integrated into a supervised contrastive learning objective (InfoNCE). Our sampling strategy is reliability-guided: for classes with mature prototypes (determined by a reliability score based on memory size and variance), we use the prototype $\mathbf{p}_c$ as the positive sample, providing a stable learning target. For negative sampling, we strategically select prototypes of "sibling" categories (those sharing the same immediate parent), forcing the model to learn fine-grained distinctions crucial for navigating the complex hierarchy.

## Hierarchical Uncertainty Calibration

Building on these robust representations, this component provides trustworthy predictions by explicitly modeling uncertainty across the entire label hierarchy. This is essential for identifying ambiguous or out-of-scope inputs.

**Per-Node Tail Distribution Modeling** Inspired by OpenMax, we model the distribution of distances from correctly classified training samples to their corresponding prototypes. Critically, this is done for every node in the hierarchy (both leaf and internal) by fitting a separate Weibull tail distribution to its distance distribution. This captures outlier likelihood at multiple levels of granularity.

**Recursive Normalization and Open-Set Estimation** During inference, for a new input embedding $z$, we calculate its distance to every prototype $\mathbf{p}_c$ and use the corresponding Weibull model to estimate its outlier probability $w_c$. The initial activation scores $s_c$ are then revised to:

$$s'_c = s_c \cdot (1 - w_c)$$

These revised scores are propagated through the hierarchy via a recursive normalization process, which enforces the structural constraint that a parent's probability must equal the sum of its children's probabilities. The final probability of the input being "unknown," $P_{unknown}(x)$, is the residual probability mass not assigned to any known class after this structurally consistent normalization. A prediction is flagged as "uncertain" if $P_{unknown}(x)$ or the confidence in the top prediction falls below pre-defined thresholds, guiding expert review.

Algorithm 1: HOCM: Training and Inference

**Input:** Training set $\mathcal{D}$, label hierarchy $\mathcal{H}$, input text $x$
**Initialize:** Encoder $f_\theta$, Memory Banks $\{\mathbf{M}_c\}_{c \in \mathcal{C}_{known}}$
**Training Phase (for each batch):**
**for** each sample $(x_i, y_i)$ in batch **do**
    $z_i \leftarrow f_\theta(x_i)$
    Update memory bank $\mathbf{M}_{y_i}$ with $z_i$ using quality-aware filtering (IQR-based)
    Update prototype $\mathbf{p}_{y_i}$ from $\mathbf{M}_{y_i}$
    Assess prototype reliability $R_{y_i}$
    Select positive sample $x_p$ (prototype $\mathbf{p}_{y_i}$ if $R_{y_i}$ is high, else in-batch)
    Select hard negative samples $x_n$ (sibling prototypes)
    Compute contrastive loss $L_{InfoNCE}$ and update $f_\theta$
**end for**
**After Training:**
**for** each node $c \in \mathcal{C}_{known}$ in hierarchy $\mathcal{H}$ **do**
    Compute distances of training samples to prototype $\mathbf{p}_c$
    Fit Weibull tail distribution $W_c$ to the distances
**end for**
**Inference Phase (for a new input $x$):**
$z \leftarrow f_\theta(x)$
Get initial activation scores $\{s_c\}$ from classifier head
**for** each node $c \in \mathcal{C}_{known}$ **do**
    $d(z, \mathbf{p}_c) \leftarrow$ distance from $z$ to prototype $\mathbf{p}_c$
    $w_c \leftarrow$ outlier probability from $W_c(d(z, \mathbf{p}_c))$
    $s'_c \leftarrow s_c \cdot (1 - w_c)$
**end for**
$\{P_{final}(c|x)\} \leftarrow$ RecursiveNormalize($\{s'_c\}$, $\mathcal{H}$)
$P_{unknown}(x) \leftarrow 1 - \sum_{c \in \mathcal{C}_{leaves}} P_{final}(c|x)$
**Return** prediction or "Uncertain" flag based on thresholds

# Experiments

## Experimental Setup

**Dataset.** We evaluate HOCM on a real-world social survey questionnaire dataset, reflecting the complexities encountered in practical social science research. As summarized in Table 1, the dataset comprises 35,168 labeled samples derived from various survey instruments. We split the data into training (20,564), validation (6,862), and test (6,860) sets, focusing on 76 high-resource categories (defined as having $\geq$50 training samples) for standard supervised learning and evaluation.

To rigorously assess open-set recognition capabilities, we utilize an additional external test set containing 882 samples belonging to 21 low-resource categories (those with $< 50$ training samples). These low-resource categories were completely held out during the training phase and serve as our proxy for unknowns, encompassing both genuinely out-of-scope topics and known concepts with insufficient representation in the training data. Detailed per-category statistics are available in the Supplementary Materials.

**Dataset Characteristics.** The dataset exhibits properties typical of large-scale social surveys:

- **Significant Class Imbalance:** Sample counts per category vary drastically (ranging from 2 to 2,668 across the full dataset), presenting a challenge for standard classifiers.
- **Complex Hierarchy:** The label ontology is deeply hierarchical, featuring 16 top-level categories (e.g., *Education, Health, Employment*) branching into 120 fine-grained subcategories (e.g., *Primary Schooling*, *Higher Education*, *Mental Health Services Use* within their respective parents). This reflects the multi-faceted nature of social phenomena.
- **Rich Contextual Input:** Each sample is constructed by concatenating the questionnaire title, the full question text, and all available response options (if any), providing rich textual context for classification. For example:

  *My Five-Year-Old Son/Daughter: How many real meals does your child have per day? [Category: Health Behaviour]*

The defined parent-child relationships within the hierarchy are strictly maintained for evaluation purposes. The low-resource categories, deliberately excluded from training, constitute a realistic benchmark for evaluating the model's ability to handle inputs beyond its well-represented training knowledge.

| Split | Samples | Categories | Avg. Length | Type |
|---|---|---|---|---|
| Training | 20,564 | 76 | 165.6 | HR |
| Validation | 6,862 | 76 | 165.4 | HR |
| Test (In-Dist.) | 6,860 | 76 | 167.4 | HR |
| Test (External) | 882 | 21 | 182.4 | LR |

Table 1: Dataset Overview for Hierarchical Classification and Open-Set Evaluation. High-resource categories ($\geq$50 training samples) form the core training/test sets. Low-resource categories ($<$50 training samples) form the external open-set evaluation set. HR(High-resource), LR(Low-resource).

**Implementation Details.** Our implementation builds upon the YATO toolkit (Wang et al. 2023), extending its capabilities to include hierarchical classification structures and incorporating our proposed memory-enhanced contrastive learning module (Section ) and hierarchical uncertainty calibration (Section ).

## Baselines and Evaluation Framework

We evaluate HOCM under three distinct scenarios reflecting practical deployment needs:

**Evaluation Scenarios and Metrics.**

1. **Closed-Set Classification:** Assesses performance on the 76 *known* high-resource categories using the standard test set (Test In-Dist.). Metrics: standard Accuracy (Acc), Micro-F1 (Mi-F1) robust to imbalance, and Macro-F1 (Ma-F1) sensitive to minority class performance.
2. **Reliability-Aware Classification:** Evaluates HOCM's ability to identify and handle uncertainty on the standard test set. Metrics:

- **Unreliability Rate (Unrel):** Percentage of test samples flagged as uncertain by HOCM (i.e., assigned low confidence or high unknown probability based on thresholds tuned on the validation set).
- **Corrected Accuracy (Corr-Acc) / Corrected Macro-F1 (Corr-Ma-F1):** Performance metrics calculated only on the subset of samples *not* flagged as uncertain. These reflect the expected performance when uncertain cases are routed for expert review, measuring the trustworthiness of high-confidence predictions.

3. **Open-Set Detection:** Measures the ability to identify samples from the 21 *unknown* low-resource categories (Test External) when presented alongside known samples from Test In-Dist. Metrics:
   - **Manual Verification Rate (MVR):** The proportion of samples flagged by the system as 'unknown' or uncertain that truly belong to one of the 21 low-resource categories. This application-centric metric measures the *precision* of the flagging mechanism, indicating the efficiency of potential expert review efforts targeted at novel or under-represented content.
   - **Macro-MVR:** The category-averaged MVR across the 21 unknown classes, providing a balanced view of detection performance for different types of unknown content.

**Metric Rationale.** This suite of metrics provides a comprehensive assessment. Standard metrics evaluate core classification capability. Reliability-aware metrics (Unrel, Corr-Acc, Corr-Ma-F1) directly quantify the system's trustworthiness and its ability to manage uncertainty effectively, crucial for dependable application in social science. Open-set metrics (MVR, Ma-MVR) measure the practical efficiency of identifying novel or under-represented content, vital for discovery and quality control.

**Baseline Methods.** We compare HOCM against representative methods:

- **Traditional/Neural:** TF-IDF feature extraction coupled with XGBoost; a hybrid character-level CNN and word-level BiLSTM model (CharCNN+BiLSTM) using GloVe embeddings (Pennington, Socher, and Manning 2014), similar to architectures like Yang and Zhang (2018).
- **Pre-trained Language Models (PLMs):** Standard fine-tuning of widely adopted PLMs: BERT (Kenton and Toutanova 2019), RoBERTa (Liu 2019), and DeBERTa (He et al. 2020). We also include the efficient Phi-3.5-mini (Abdin et al. 2024) fine-tuned with LoRA for parameter-efficient adaptation.
- **Open-Set Recognition (OSR) Methods:** We implement established OSR methods using a BERT backbone for fair comparison: OpenMax (Bendale and Boult 2016), DOC (Shu, Xu, and Liu 2017), and LOS (Chen et al. 2023). These represent common approaches for handling unknown inputs in classification.

These baselines cover standard practices and dedicated OSR techniques relevant to our task, providing a robust benchmark for HOCM's performance.

## Main Results

We present the quantitative results across the three evaluation scenarios. Tables 2, 3, and 4 summarize the key findings.

**Closed-Set Classification Performance.** Table 2 presents the performance on known categories. HOCM integrated with PLMs (e.g., BERT+HOCM) achieves strong results (84.9% Acc, 82.1% Ma-F1), slightly below the raw fine-tuned BERT (85.3% Acc) but substantially outperforming traditional methods and the base CharCNN+BiLSTM. This minor difference is an expected trade-off. HOCM's memory-enhanced contrastive learning is explicitly designed to prioritize robust and stable representations for reliable uncertainty estimation, rather than solely maximizing accuracy on the closed-set training distribution, which can lead to overfitting on noisy instances. The significant gains in reliability-aware performance (e.g., Corr-Acc rising to 92.1%) validate this design choice. Notably, HOCM consistently improves over its non-PLM base model (CharCNN+BiLSTM+HOCM vs. CharCNN+BiLSTM), demonstrating the benefits of its components.

| Method | Acc (%) | Mi-F1 (%) | Ma-F1 (%) |
|---|---|---|---|
| TF-IDF + XGBoost | 76.2 | 75.0 | 74.5 |
| CharCNN + BiLSTM | 77.1 | 76.5 | 76.0 |
| BERT | 85.3 | 83.7 | 81.6 |
| DeBERTa | 83.9 | 82.3 | 81.1 |
| RoBERTa | 83.4 | 81.5 | 81.2 |
| Phi-3.5-mini | 83.9 | 82.0 | 81.5 |
| CharCNN + BiLSTM + HOCM | 80.7 | 79.1 | 78.5 |
| BERT + HOCM | **84.9** | **83.0** | **82.1** |
| RoBERTa + HOCM | 83.8 | 82.3 | 81.7 |
| DeBERTa + HOCM | 84.2 | 82.5 | 81.8 |

Table 2: Standard Classification Results on the 76 high-resource known categories (Test In-Dist.).

**Reliability-Aware Performance.** Table 3 highlights HOCM's effectiveness in identifying uncertain predictions. For example, BERT+HOCM flags 14.3% of the in-distribution test samples as unreliable (Unrel). When these flagged samples are set aside (simulating referral for expert review), the accuracy on the remaining high-confidence predictions (Corr-Acc) increases significantly from 84.9% to 92.1%, with Corr-Ma-F1 similarly rising to 90.2%. This demonstrates that HOCM effectively separates low-confidence predictions. When these flagged instances are set aside, the accuracy on the remaining high-confidence subset increases substantially (from 84.9% to 92.1%), indicating that the uncertainty scores are a useful proxy for identifying likely misclassifications. The consistent Unrel rates across different PLM backbones (13.7%–15.3%) indicate stable uncertainty estimation.

**Open-Set Detection Performance.** Table 4 evaluates the detection of samples from the 21 unseen low-resource categories. HOCM achieves performance competitive with or slightly exceeding dedicated OSR baselines (e.g.,

| Method | Corr-Acc (%) | Corr-Ma-F1 (%) | Unrel (%) |
|---|---|---|---|
| BERT + HOCM | **92.1** | **90.2** | 14.3 |
| RoBERTa + HOCM | 91.8 | 90.9 | 15.3 |
| DeBERTa + HOCM | 91.5 | 89.6 | 13.7 |

Table 3: Reliability-aware Classification Results on Test In-Dist. Corr-Acc/Corr-Ma-F1 are computed on samples *not* flagged as uncertain (Unrel). Higher Corr-Acc/F1 with a reasonable Unrel is desirable.

BERT+OpenMax, DOC, LOS). With BERT+HOCM attaining an MVR of 78.9% (and Ma-MVR of 78.3%) and further improvements observed with stronger backbones like DeBERTa (MVR up to 79.5%), HOCM effectively identifies inputs that fall outside the well-represented training categories—crucial for directing expert attention to challenging instances.

| Method | MVR (%) | Ma-MVR (%) |
|---|---|---|
| BERT + OpenMax | 77.9 | 77.3 |
| BERT + DOC | 78.4 | 77.8 |
| BERT + LOS | 78.8 | 78.2 |
| BERT + HOCM | **78.9** | **78.3** |
| RoBERTa + HOCM | 79.2 | 78.6 |
| DeBERTa + HOCM | **79.5** | **79.1** |

Table 4: Open-Set Detection Results (MVR/Ma-MVR) on the external test set (21 low-resource categories). A higher MVR indicates better precision in identifying true unknowns/low-resource samples.

**Synthesis.** Overall, the quantitative results indicate that HOCM achieves a compelling balance. It maintains strong classification accuracy on known categories while significantly enhancing prediction reliability through effective uncertainty estimation. Concurrently, it demonstrates competitive performance in detecting under-represented or out-of-scope inputs, making it a well-rounded solution for real-world survey classification.

## Ablation Studies

We perform ablation studies to evaluate key components of HOCM using the BERT backbone. Table 5 reports results with three simplified metrics: closed-set accuracy (Acc), reliability measured by corrected accuracy (Corr-Acc), and open-set detection via manual verification rate (MVR).

The results demonstrate the effectiveness of HOCM's components. Removing the entire memory-enhanced contrastive learning module (ID 4 vs ID 1) significantly degrades both reliability (Corr-Acc drops by 3.1%) and open-set detection (MVR drops by 4.4%), despite a slight increase in standard accuracy, highlighting the importance of robust representations. Ablating specific parts of the memory management, namely the quality filter (ID 2) and reliability sampling (ID 3), leads to noticeable but smaller performance drops, confirming their contribution to stability and effectiveness.

| ID | Acc (%) | Corr-Acc (%) | MVR (%) |
|---|---|---|---|
| 1: Full HOCM | **84.9** | **92.1** | **78.9** |
| 2: *w/o Quality Filter* | 84.5 | 91.5 | 78.0 |
| 3: *w/o Reliab. Sampling* | 84.6 | 91.6 | 78.2 |
| 4: *w/o Memory CL (CE Loss)* | 85.1 | 89.0 | 74.5 |
| 5: *w/ Flat OpenMax* | 84.8 | 91.0 | 76.5 |
| 6: *w/o UC (Softmax)* | 85.1 | - | 72.5 |
| 7: BERT Baseline | 85.3 | - | 68.0 |

Table 5: Ablation study on BERT. Variants: (1) Full HOCM (Memory CL + Hierarchical UC); (2) HOCM w/o Quality Filter; (3) HOCM w/o Reliability Sampling; (4) BERT + Hierarchical UC (No Memory CL); (5) BERT + Memory CL + Flat OpenMax (No Hierarchical UC); (6) BERT + Memory CL (No UC); (7) BERT Baseline (Standard CE, Softmax). Metrics: Accuracy (Acc), Corrected Accuracy (Corr-Acc), and Manual Verification Rate (MVR).

Similarly, the hierarchical uncertainty calibration is crucial. Removing it entirely (ID 6 vs ID 1) eliminates the reliability assessment (Corr-Acc) and substantially reduces OSR performance (MVR drops by 6.4%), even with the strong representations from Memory CL. Replacing the hierarchical mechanism with a standard flat OpenMax (ID 5 vs ID 1) maintains high accuracy but results in lower reliability (Corr-Acc -1.1%) and notably worse open-set detection (MVR -2.4%), underscoring the benefit derived from explicitly leveraging the hierarchy in uncertainty estimation. The full HOCM framework (ID 1) achieves the best balance, particularly excelling in reliability and open-set detection compared to both the baseline (ID 7) and the ablated variants.

## Qualitative Analysis

To gain deeper insights into the effectiveness of our HOCM framework, we conduct extensive qualitative analyses through expert annotation evaluation, focusing on how the learned representations capture both class semantics and uncertainty patterns.

## Expert Annotation Analysis

We conducted a comprehensive expert annotation process to validate our model's performance and uncertainty detection capabilities. This analysis consisted of two phases: (1) expert verification of prediction-ground truth disagreements, and (2) in-depth analysis of samples flagged as low-confidence by our model. For each question, experts verified prediction and ground truth correctness, identified potential multi-topic cases, and documented reasons for classification ambiguity. Table 6 provides examples for dominant patterns.

Our analysis revealed a strong alignment between model uncertainty detection and expert-identified challenging cases. Questions receiving low confidence scores consistently corresponded to cases where experts identified legitimate ambiguity or need for additional context. This alignment was particularly evident in cases requiring expert dis-

| Pattern | Example Question | Expert Note | Remark |
|---|---|---|---|
| **Semantic Ambiguity** | *"How would you assess this pupil's general knowledge compared with other pupils of the same age?"* | "Uncertain if 'general knowledge' fits a cognitive skill label." | Distinguishing Education, Cognitive Skills, and Personality/Temperament is challenging. |
| **Insufficient Context** | *"What do you do if you get it while walking?"* | "Text alone is unclear – 'it' is undefined." | Lacks context, making the question ambiguous. |
| | *"Has this investigation resulted in extra help for the child?"* | "Likely about Learning difficulties or Education, but insufficient info for certainty." | Additional context needed to classify properly. |
| **Category Overlap** | *"To what extent has your child's chronic health condition affected their school attendance and social relationships with classmates?"* | "Spans Physical Health, Education, and Social Support." | Overlapping domains complicate single-label assignment. |
| | *"How satisfied are you with the care you and your partner received during labour?"* | "Involves Health services utilisation and Childbirth." | Multiple domains make single-label classification difficult. |

Table 6: Summary of Patterns Identified in Expert Annotations

cussion for topic determination, questions needing multiple topic labels, and samples exhibiting the key annotation patterns discussed above.

These findings have important practical implications for deploying such systems in real-world scenarios. Organizations implementing similar classification systems should focus on maintaining clear documentation of topic definitions while developing robust protocols for handling multi-topic questions. Additionally, establishing clear procedures for escalating uncertain cases to expert reviewers and continuously refining annotation schemes based on feedback is crucial for system effectiveness. The strong correlation between algorithmic uncertainty estimation and expert judgment validates our quality-aware classification approach, while providing practical guidance for implementing robust quality control measures in large-scale survey classification systems.

## Computational Cost

To assess practical viability, we measured the computational overhead of **HOCM**. On an NVIDIA A100 GPU, compared to a standard BERT baseline, the BERT + HOCM model introduced a modest overhead: approximately 15% additional training time per epoch (due to memory bank updates and contrastive calculations) and a 5% increase in per-sample inference time (due to distance calculations and hierarchical calibration). This indicates that HOCM's significant gains in reliability are achieved with acceptable computational cost, making it suitable for real-world deployment.

## Conclusion

Automated classification of complex social survey questionnaires is essential yet challenging due to intricate hierarchies, class imbalance, ambiguity, and the presence of under-represented or out-of-scope categories, hindering reliable social science analysis. This paper introduced HOCM, a framework designed to address these specific hurdles. HOCM integrates (1) memory-enhanced contrastive learning to derive robust representations from noisy, imbalanced survey data, and (2) hierarchical uncertainty estimation to enforce taxonomic consistency and identify inputs that are likely ambiguous or out-of-scope. This allows routing low-confidence predictions for expert review, thereby increasing the trustworthiness of the final classified dataset.

Experiments demonstrate HOCM's effectiveness: it achieves strong accuracy on known categories (e.g., 84.9% Acc with BERT), successfully flags uncertain instances (14.3% Unrel) leading to substantially improved reliability on confident predictions (92.1% Corr-Acc), and competitively detects under-represented/unknown categories (78.9% MVR). Expert validation confirms that HOCM's uncertainty flags predominantly identify genuine data complexities (semantic ambiguity, category overlap), validating its practical utility.

By delivering more trustworthy classifications and efficiently guiding expert review towards problematic cases, HOCM offers a valuable tool for enhancing the quality, scalability, and rigor of computational social science research based on survey data. Future work includes integrating expert feedback for adaptive learning and addressing evolving concepts in longitudinal analysis.

## References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1563–1572.

Chen, J.; Zhang, R.; Chen, J.; and Hu, C. 2024. Open-Set

Semi-Supervised Text Classification via Adversarial Disagreement Maximization. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2170–2180. Bangkok, Thailand: Association for Computational Linguistics.

Chen, J.; Zhang, R.; Chen, J.; Hu, C.; and Mao, Y. 2023. Open-Set Semi-Supervised Text Classification with Latent Outlier Softening. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 226–236.

Dai, Y.; Lang, H.; Zeng, K.; Huang, F.; and Li, Y. 2023. Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5292–5305.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Hendrycks, D.; and Gimpel, K. 2022. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Jain, V.; Rungta, M.; Zhuang, Y.; Yu, Y.; Wang, Z.; Gao, M.; Skolnick, J.; and Zhang, C. 2024. HiGen: Hierarchy-Aware Sequence Generation for Hierarchical Text Classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1354–1368.

Kaur, R.; Jha, S.; Roy, A.; Park, S.; Dobriban, E.; Sokolsky, O.; and Lee, I. 2022. iDECODe: In-distribution equivariance for conformal out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7104–7114.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

Liu, W.; Tang, Z.; Li, J.; Chen, K.; and Zhang, M. 2024. Memlong: Memory-augmented retrieval for long text modeling. *arXiv preprint arXiv:2408.16967*.

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

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2022. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *The Eleventh International Conference on Learning Representations*.

Novello, P.; Dalmau, J.; and Andeol, L. 2024. Out-of-Distribution Detection Should Use Conformal Prediction (and Vice-versa?). *arXiv preprint arXiv:2403.11532*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Shu, L.; Xu, H.; and Liu, B. 2017. DOC: Deep Open Classification of Text Documents. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2911–2916. Copenhagen, Denmark: Association for Computational Linguistics.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Wang, Z.; Wang, Y.; Wu, J.; Teng, Z.; and Yang, J. 2023. YATO: Yet Another deep learning based Text analysis Open toolkit. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 131–139. Singapore: Association for Computational Linguistics.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, 23631–23644. PMLR.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Yang, J.; and Zhang, Y. 2018. NCRF++: An Open-source Neural Sequence Labeling Toolkit. In *Proceedings of ACL 2018, System Demonstrations*, 74–79.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zheng, Y.; Chen, G.; and Huang, M. 2020. Out-of-Domain Detection for Natural Language Understanding in Dialog Systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 1198–1209.

Zhou, H.; Wan, X.; Proleev, L.; Mincu, D.; Chen, J.; Heller, K. A.; and Roy, S. 2023. Batch Calibration: Rethinking Calibration for In-Context Learning and Prompt Engineering. In *The Twelfth International Conference on Learning Representations*.

Zhou, Y.; Liu, P.; and Qiu, X. 2022. KNN-Contrastive Learning for Out-of-Domain Intent Classification. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5129–5141. Dublin, Ireland: Association for Computational Linguistics.