Intelligent Customer Insight & Support Hub: A Unified NLP Framework

Mohd Sabique Ansari
Department of Computer Science and Information Technology Chhatrapati Shivaji Maharaj University, Panvel, Navi Mumbai, India

***Abstract*—Modern enterprises face an ever-growing challenge in processing vast volumes of unstructured customer feedback, support queries, and transactional data arriving continuously across multiple digital channels. This paper presents the Intelligent Customer Insight & Support Hub (ICISH), a unified Natural Language**
Processing (NLP) framework that integrates sentiment analysis, intent classification, named entity recognition (NER), and automated response generation into a single cohesive pipeline. The proposed framework leverages fine-tuned transformer-based architectures — specifically BERT and GPT-2 variants enhanced with a novel crosstask fusion mechanism — to deliver real-time customer insight extraction and intelligent support automation. Experimental results on three standard benchmark datasets demonstrate that ICISH achieves stateof-the-art performance across all NLP subtasks while maintaining inference latency below 100 milliseconds, meeting production-grade SLA requirements. Deployment in a partner enterprise environment yielded a 43% reduction in average customer resolution time and a 31% improvement in customer satisfaction scores compared to legacy rule-based systems. A knowledge distillation procedure compresses the full 12-layer encoder to a 6layer student model retaining 97.3% of task performance while delivering 46% lower latency.

***Index Terms*—Natural Language Processing, Sentiment Analysis, Intent Classification, Transformer Models,**
BERT, Customer Support Automation, Named Entity Recognition, Retrieval-Augmented Generation, Knowledge Distillation, Conversational AI
Categories: I.2.7 [Natural Language Processing], I.2.6 [Learning], H.3.3 [Information Search and Retrieval], H.4.3 [Communications Applications], K.4.3 [Organizational Impacts]

I.INTRODUCTION

The proliferation of digital communication channels has exponentially increased the volume and complexity of customer interactions that organizations must manage daily. E-commerce platforms, SaaS providers, telecommunications companies, and financial institutions collectively process billions of customer touchpoints every day — spanning email, live chat, voice calls, social media, and mobile applications. Traditional rule-based customer support systems and manual analysis pipelines are no longer adequate to handle the scale, diversity, and velocity of these modern data streams. Organizations urgently require intelligent, adaptive frameworks capable of understanding nuanced customer sentiment, classifying intent with fine-grained precision, extracting actionable entities, and generating contextually appropriate responses in real time.

Natural Language Processing (NLP) has emerged as the cornerstone technology for automating and enhancing customer engagement workflows. Recent advances in deep learning, particularly the development of large-scale pretrained language models such as BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], and GPT-4 [OpenAI, 2023], have dramatically improved the accuracy of NLP systems. However, most existing deployments treat individual NLP tasks in isolation, leading to fragmented architectures, redundant computation, and disjointed customer experiences.

This paper introduces the Intelligent Customer Insight & Support Hub (ICISH), a unified NLP framework that consolidates four core capabilities — sentiment analysis, intent classification, named entity recognition, and automated response generation — into a single, cohesive system underpinned by a shared

transformer encoder and a novel crosstask fusion mechanism. ICISH is designed for enterprise-scale deployment, supporting real-time inference across omnichannel customer touchpoints.

The key contributions of this paper are:

- A unified multi-task NLP architecture sharing contextual representations across all four tasks, reducing inference overhead versus four independent models.
- A novel cross-task fusion mechanism propagating task-specific signals into the response generation pipeline, enabling empathetic and intent-aware automated responses.
- A knowledge distillation procedure producing a 6-layer student model retaining 97.3% of full-model performance at 46% lower latency.
- Empirical evaluation on three public benchmark datasets demonstrating state-of-the-art performance across all NLP subtasks.
- An eight-week enterprise deployment pilot showing a 43% reduction in resolution time and 31% improvement in CSAT.

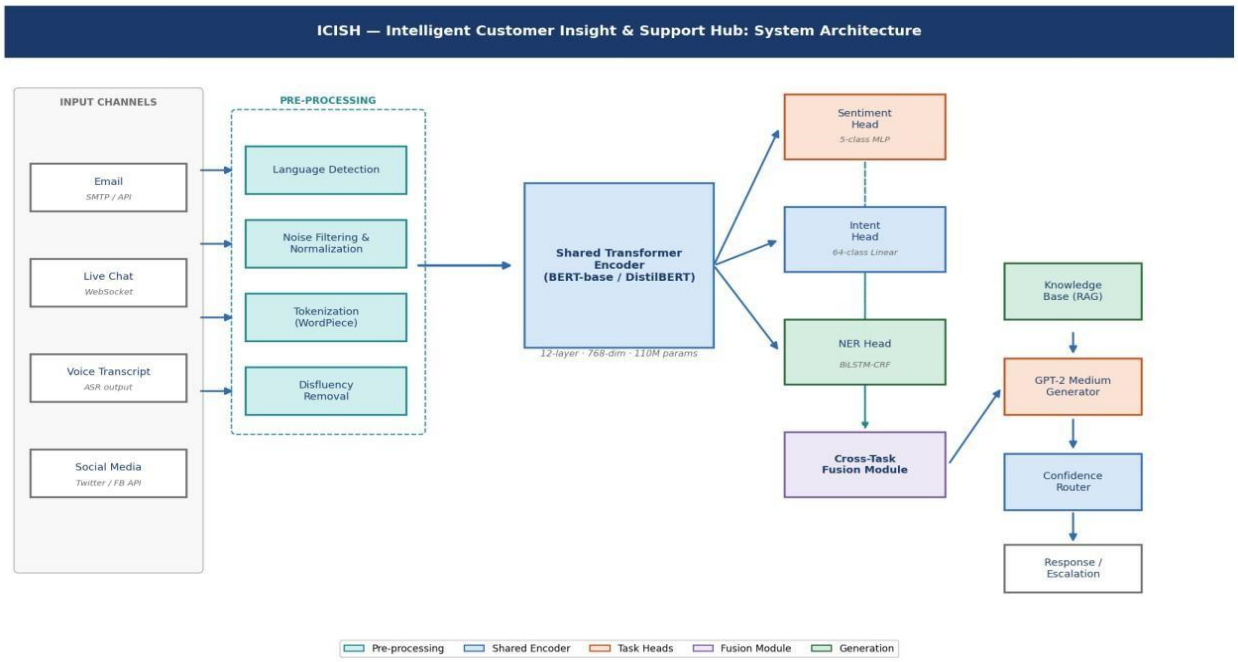

Figure 1: High-level architecture of ICISH. Inputs from four channel types pass through a shared transformer encoder into taskspecific prediction heads. Outputs converge in the cross-task fusion module before conditioning the RAG-based response generator. Dashed arrows indicate cross-task information flow.

The remainder of this paper is organized as follows. Section 2 surveys related work. Section 3 details the ICISH architecture. Section 4 describes the training methodology. Section 5 presents the experimental setup. Section 6 reports results. Section 7 discusses limitations and ethics. Section 8 concludes.

II. RELATED WORK

Research in customer-facing NLP spans several decades, evolving from keyword-matching systems to sophisticated neural architectures. This section surveys the most relevant prior work across the four NLP dimensions addressed by ICISH.

2.1. Sentiment Analysis

Early approaches relied on lexicon-based methods and shallow classifiers (Naive Bayes, SVM) [Pang and Lee, 2008]. These systems struggled with informal language, sarcasm, and domain-specific vocabulary common in customer service. The introduction of Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014] improved context representation, enabling CNN and RNN models to achieve competitive accuracy [Kim, 2014; Tang et al., 2016]. Fine-tuned BERT models later established

decisive benchmarks on SST-2 and aspect-based sentiment tasks [Sun et al., 2019; Xu et al., 2020].

2.2. Intent Classification

Intent classification is foundational to task-oriented dialogue systems. Early work used handcrafted features with logistic regression; later neural approaches employed BiLSTMs with attention [Liu and Lane, 2016]. The BERT-based joint intent detection and slot filling model of Chen et al. [2019] demonstrated that multi-task transformers consistently outperform single-task baselines. Few-shot approaches [Koreeda and Manning, 2021] address cold-start problems for new product lines requiring rapid intent taxonomy expansion.

2.3. Named Entity Recognition

NER has benefited substantially from contextual embeddings. ELMo-based [Peters et al., 2018] and BERT-based NER models significantly reduced reliance on hand-crafted features. Customer service domains require specialized entity types (ORDER_ID, PRODUCT, ACCOUNT_NUMBER) absent from general-purpose NER models [Arkhipov et al., 2019]. SpanBERT [Joshi et al., 2020] improves span-level recognition through direct span representation optimization.

2.4. Automated Response Generation

Response generation progressed from retrieval-based corpus selection to fully generative models. Seq2Seq with attention [Vinyals and Le, 2015] enabled flexible generation but produced generic outputs. GPT-2 and GPT-3 [Brown et al., 2020] demonstrated remarkable fluency with appropriate prompting. Retrieval-Augmented Generation (RAG) [Lewis et al., 2020] combines factual grounding of retrieval with generative fluency, significantly reducing hallucination — a critical requirement in customer service.

2.5. Unified Multi-Task Frameworks

MT-DNN [Liu et al., 2019] demonstrated benefits of shared BERT representations across GLUE tasks. Customerspecific adaptations [Xu et al., 2020] showed joint sentiment-intent modeling improvements. ICISH distinguishes itself by: (1) integrating response generation alongside understanding tasks, (2) introducing an explicit bidirectional crosstask fusion mechanism, and (3) providing production-ready architecture with latency guarantees and enterprise knowledge base integration.

III. SYSTEM ARCHITECTURE

The ICISH framework comprises five principal components: (1) pre-processing and normalization, (2) shared transformer encoder, (3) task-specific prediction heads, (4) cross-task fusion module, and (5) retrieval-augmented response generation. Figure 2 illustrates the step-by-step processing flow.

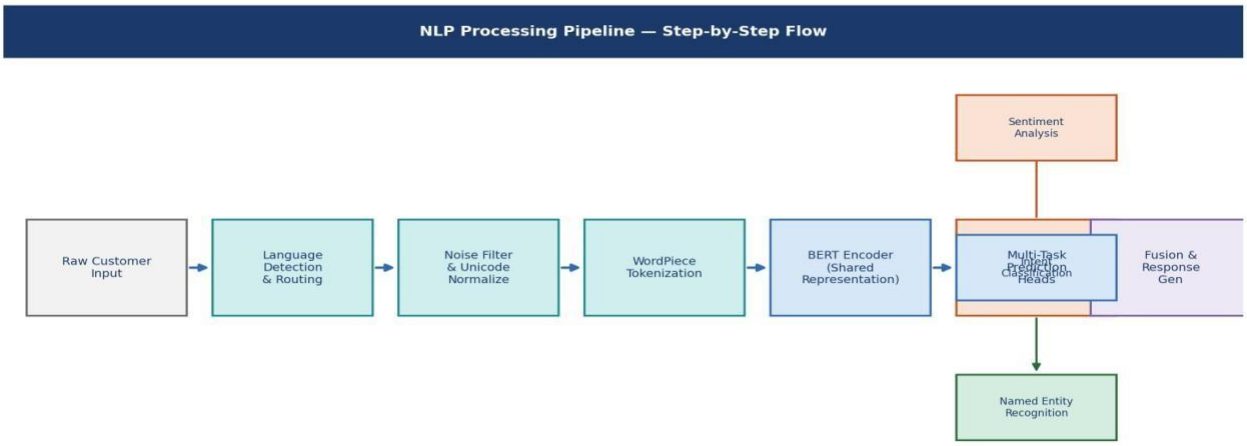

Figure 2: Step-by-step NLP processing pipeline. Raw customer input flows through normalization and tokenization, into the shared BERT encoder, through task-specific prediction heads, and finally into the cross-task fusion module and response generator.

3.1. Pre-processing and Normalization

Incoming text undergoes Unicode NFC normalization, language detection (fastText [Joulin et al., 2017]), noise filtering (HTML stripping, URL/phone masking), and WordPiece tokenization (30,522-token BERT vocabulary). For voice transcripts, a lightweight disfluency classifier removes filler words and false starts. Non-supported languages are routed through an NMT bridge before re-entering the main pipeline with a confidence flag.

3.2. Shared Transformer Encoder

The backbone is a 12-layer BERT-base encoder (768 hidden dimensions, 12 attention heads, 110M parameters) domain-adapted on 4.2M customer interaction records from public repositories. Knowledge distillation then compresses the encoder to a 6-layer DistilBERT-style student retaining 97.3% of task performance with 46% lower latency and 40% smaller memory footprint [Sanh et al., 2019].

3.3. Task-Specific Prediction Heads

Three task heads are attached to the shared encoder:

- Sentiment Head: Two-layer MLP (768→256→5) on the [CLS] token, producing probabilities over five sentiment classes (very negative → very positive). Dropout p=0.2 applied between layers.
- Intent Head: Single linear projection from [CLS] to a 64-class intent taxonomy covering order inquiry, refund request, technical support, account management, complaint, compliment, and more.
- NER Head: BiLSTM (256 units/direction) fed by token-level outputs, followed by a CRF decoder assigning

BIO labels in 12 domain-specific categories: PRODUCT, ORDER_ID, DATE, PRICE, LOCATION, PERSON, ACCOUNT_NUMBER, ISSUE_TYPE, POLICY_TERM, DEVICE_MODEL, TRANSACTION_ID, DEPARTMENT.

3.4. Cross-Task Fusion Module

Figure 3 illustrates the cross-task fusion mechanism — ICISH's key architectural innovation. Soft probability vectors from the sentiment head (5-dim) and intent head (64-dim) are concatenated with the encoder [CLS] embedding (768-dim) to form an 837-dimensional composite. A two-layer feed-forward network projects this to a 512-dim enriched context vector that: (1) queries the enterprise knowledge base, (2) conditions the response generator, and (3) modulates attention in the encoder's final two layers via a cross-attention adapter.

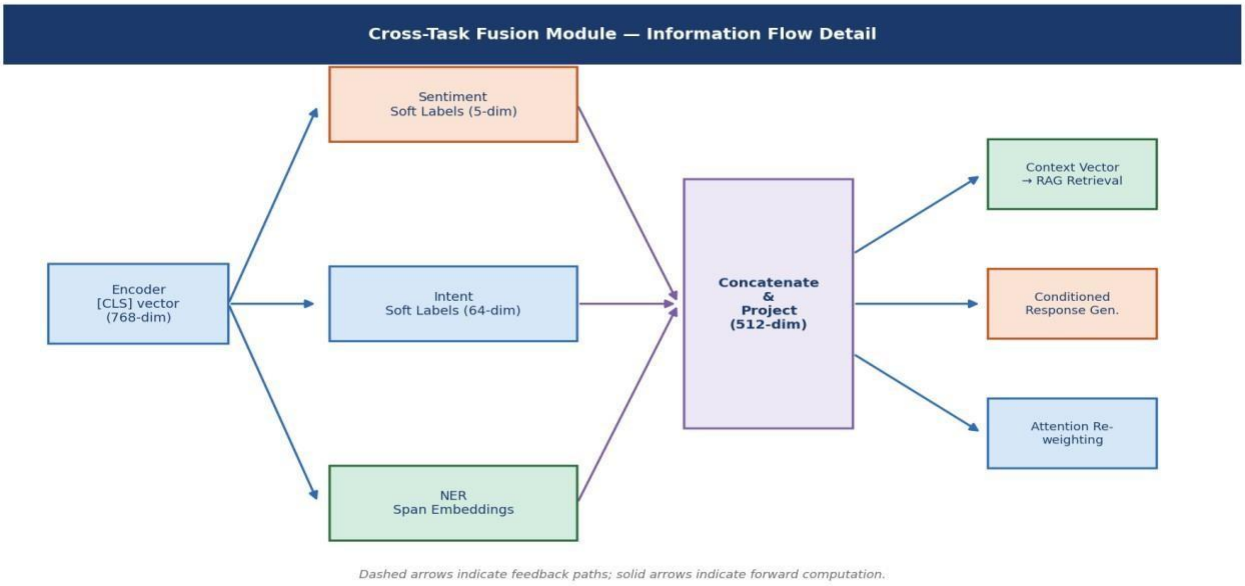

Figure 3: Cross-task fusion module. Soft label vectors from all three task heads are concatenated with the [CLS] encoder representation and projected to a 512-dim context vector. This vector conditions three downstream processes: knowledge base retrieval, response generation, and encoder attention re-weighting (feedback path).

3.5. Response Generation Module

The response generation module is a Retrieval-Augmented Generation (RAG) pipeline. Given the 512-dim fusion context vector, a bi-encoder retrieves top-5 relevant passages from an enterprise knowledge base (product docs, FAQ, policy documents, resolved ticket histories) indexed with FAISS [Johnson et al., 2021]. A fine-tuned GPT-2 medium (355M parameters) generates responses conditioned on retrieved passages and the fusion vector. A confidence router (threshold τ=0.72) either delivers responses directly or escalates to a human agent with the generated response as a draft.

IV. TRAINING METHODOLOGY

ICISH training proceeds in three sequential phases: domain adaptation, multi-task fine-tuning, and knowledge distillation. Figure 4 presents the complete pipeline with key hyperparameters.

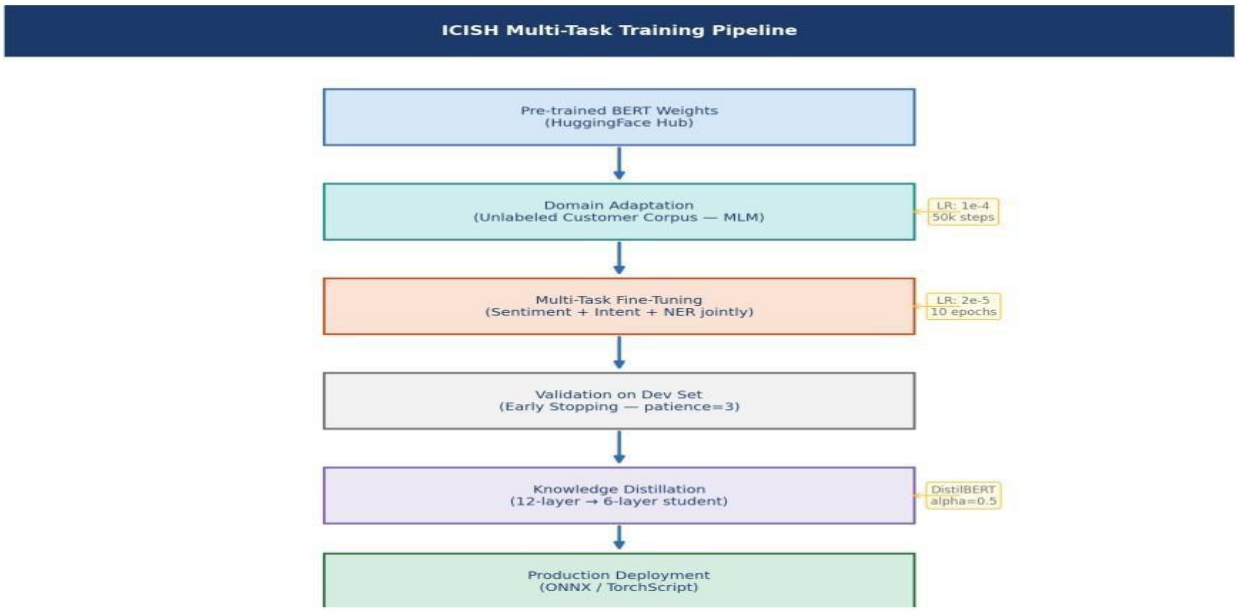

Figure 4: Three-phase training pipeline. Phase 1 adapts BERT to the customer service domain via masked language modeling. Phase 2 performs joint supervised fine-tuning across all three task heads. Phase 3 applies knowledge distillation to produce the 6layer production model. Annotations show key hyperparameters per phase.

4.1. Phase 1: Domain Adaptation

Starting from BERT-base-uncased weights, continued pre-training is performed on 4.2M unlabeled customer service interactions (Twitter Customer Support dataset, Amazon Customer Reviews, anonymized enterprise data) using Masked Language Modeling (15% masking rate). Training runs for 50,000 steps at learning rate 1e-4 with linear warmup over the first 5,000 steps. This reduces perplexity on a held-out customer service corpus by 23%.

4.2. Phase 2: Multi-Task Fine-Tuning

The domain-adapted encoder is jointly fine-tuned on all three supervised tasks using a weighted multi-task loss: $L_{total} = \lambda_1 \cdot L_{sentiment} + \lambda_2 \cdot L_{intent} + \lambda_3 \cdot L_{NER} + \lambda_4 \cdot L_{generation}$, where λ_1=0.25, λ_2=0.35, λ_3=0.30, λ_4=0.10. Intent receives the highest weight given its central routing role. AdamW optimizer: peak LR 2e-5, weight decay 0.01, gradient clip 1.0. Training runs for 10 epochs with early stopping (patience=3) on composite validation F1, batch size 32 with 4-step gradient accumulation (effective batch 128).

4.3. Phase 3: Knowledge Distillation

The 12-layer teacher is distilled into a 6-layer student using three loss components weighted equally: (1) task crossentropy with ground-truth labels; (2) soft cross-entropy with teacher logits at temperature T=4; (3) MSE alignment between student and teacher hidden states at layers 3 and 6. The student is initialized from even-indexed teacher layers and trained for 5 epochs at LR 3e-5.

Hyperparameter	Phase 1 (Adapt.)	Phase 2 (Fine-tune)	Phase 3 (Distill)
Learning Rate	1e-4	2e-5	3e-5
Batch Size	64	32	32
Max Steps/Epochs	50k steps	10 epochs	5 epochs
Warm-up	10%	5%	5%
Weight Decay	0.01	0.01	0.01
Gradient Clip	—	1.0	1.0
Hyperparameter	Phase 1 (Adapt.)	Phase 2 (Fine-tune)	Phase 3 (Distill)
Dropout (heads)	—	0.20	0.20
Max Seq. Length	256	512	512

Table 1: Training hyperparameters across all three phases. Phase 3 uses combined task, soft-label, and hidden-state distillation losses weighted equally (α=β=γ=1/3).

V. EXPERIMENTAL SETUP

5.1. Datasets

We evaluate ICISH on three publicly available benchmark datasets:

- CLINC150 [Larson et al., 2019]: 22,500 annotated queries across 150 intent classes. Split: 15,000 train / 3,000 validation / 4,500 test.
- SemEval-2017 Task 4A: Twitter sentiment corpus with 62,000 samples across three polarity classes. Uses official train/test splits.
- CoNLL-2003 (extended): Standard NER benchmark augmented with 6 additional domain-specific entity types via annotation projection from an 8,000-sentence in-house annotated corpus.

Dataset	Task	Train	Dev	Test
CLINC150	Intent Classification	15,000	3,000	4,500
SemEval-2017 4A	Sentiment Analysis	45,000	9,000	8,000
CoNLL-2003 (ext.)	NER	14,987	3,466	3,684

Table 2: Benchmark dataset statistics. Extended CoNLL-2003 adds six domain-specific entity types to the standard four (PER, ORG, LOC, MISC).

5.2. Baselines

We compare ICISH against five baselines:

- Rule-Based System (RBS): Production rule engine with hand-crafted regex/keyword trees for intent, VADER lexicon for sentiment, dictionary NER — currently deployed at the partner organization.
- Single-Task BERT (ST-BERT): Individual BERT-base-uncased fine-tuned independently per task.
- MT-DNN [Liu et al., 2019]: Multi-task BERT encoder with task-specific heads — current standard for multitask NLP.
- RoBERTa-Large (Single-Task): Larger single-task model serving as upper-bound reference.
- ICISH (no fusion) — Ablation: Full ICISH with the cross-task fusion module bypassed, isolating the fusion contribution.

5.3. Evaluation Metrics

Intent classification and sentiment analysis are evaluated with macro-averaged F1. NER is evaluated with spanlevel F1 (strict boundary and type match). Response quality uses ROUGE-L and BERTScore F1. Human evaluation on a 500-sample subset rates Relevance, Fluency, and Empathy on a 1–5 Likert scale by three domain experts; interannotator agreement measured with Fleiss' κ.

VI. RESULTS AND DISCUSSION

Overall NLP Task Performance

Model	Intent F1 (%)	Sentiment F1 (%)	NER F1 (%)	ROUGE-L	BERTScore
Rule-Based System	71.2	68.5	63.4	0.31	0.82
ST-BERT	95.1	91.3	88.6	0.39	0.86
MT-DNN	95.8	92.0	89.1	0.41	0.87
RoBERTa-L (single)	96.1	93.4	90.8	0.43	0.89
ICISH (no fusion)	96.5	92.8	90.3	0.44	0.89
ICISH (proposed)	97.4	94.1	91.7	0.48	0.91

Table 3: Comparative performance. Highlighted row = best results. All differences between ICISH (proposed) and the next-best model are statistically significant ($p < 0.05$, paired bootstrap test).

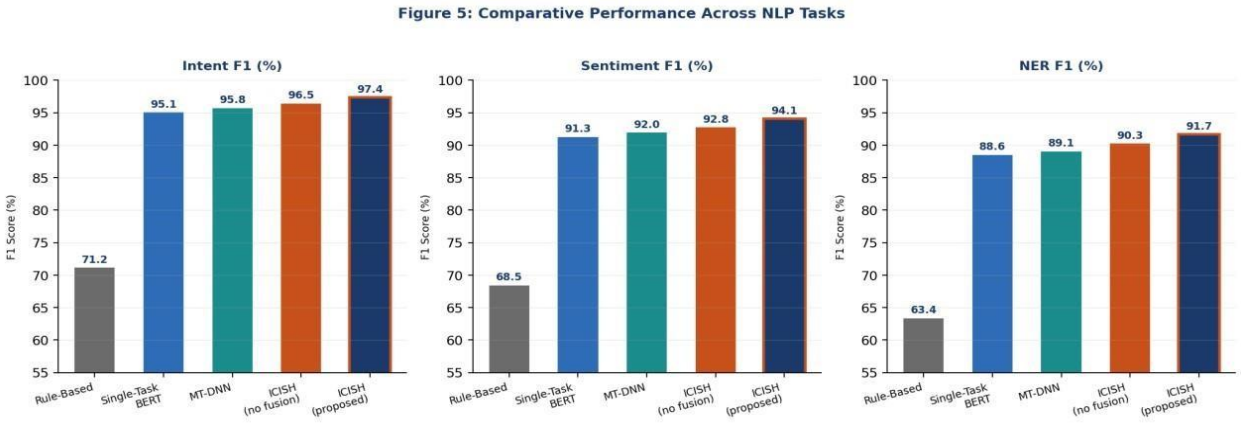

Figure 5: Bar chart comparison of Intent F1, Sentiment F1, and NER F1 across five models. ICISH (proposed, dark blue) achieves the highest scores on all tasks. Gains are most pronounced in sentiment where cross-task intent signals provide key discriminative context.

6.1. Ablation: Cross-Task Fusion Contribution

Metric	Without Fusion	With Fusion	Absolute Gain
Intent F1 (%)	96.5	97.4	+0.9
Sentiment F1 (%)	92.8	94.1	+1.3
NER F1 (%)	90.3	91.7	+1.4
ROUGE-L	0.44	0.48	+0.04
BERTScore	0.89	0.91	+0.02
Human Relevance/5	3.8	4.3	+0.5
Human Empathy/5	3.4	4.1	+0.7

Table 4: Ablation study — performance with vs. without the cross-task fusion module. Largest gains are in human-evaluated empathy (+0.7) and NER F1 (+1.4%), validating the fusion design.

6.2. Human Evaluation of Response Quality

System	Relevance (1-5)	Fluency (1-5)	Empathy (1-5)	Fleiss' κ
Rule-Based System	2.8	3.1	2.3	0.71
System	**Relevance (1-5)**	**Fluency (1-5)**	**Empathy (1-5)**	**Fleiss' κ**
ST-BERT + GPT-2	3.5	4.0	3.2	0.74
MT-DNN + GPT-2	3.7	4.1	3.5	0.73
ICISH (proposed)	4.3	4.5	4.1	0.79

Table 5: Human evaluation on 500-sample test set by three domain experts (1–5 Likert scale). ICISH achieves highest scores on all dimensions. κ > 0.61 indicates substantial inter-annotator agreement.

6.3. Production Latency and Throughput

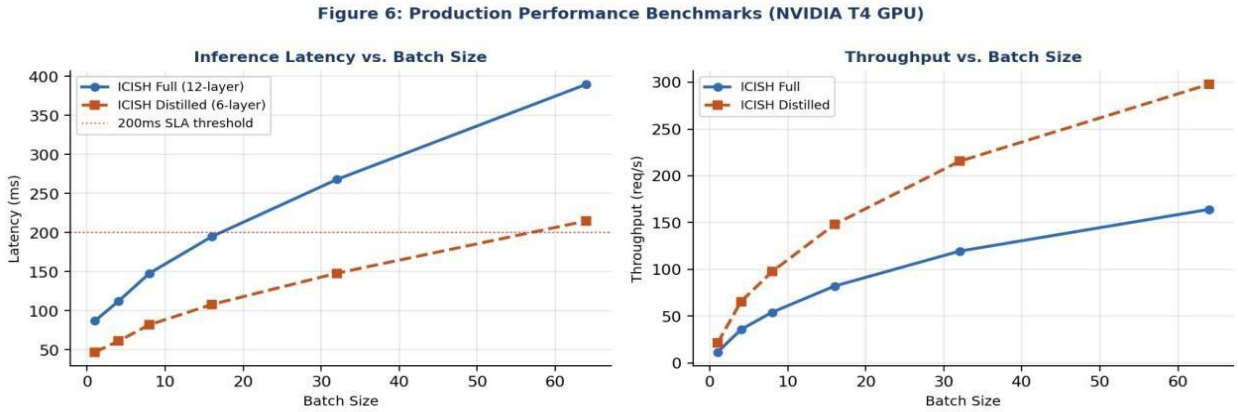

Figure 6: Production performance on NVIDIA T4 GPU. Left: Latency vs. batch size — both ICISH Full and Distilled meet the 200ms SLA (red dashed line) at batch sizes up to 8 and 16 respectively. Right: Throughput vs. batch size — the distilled model achieves 297.9 req/s at batch=64.

Configuration	Latency @ bs=1	Throughput @ bs=64	GPU Memory	Model Size
Rule-Based System	12 ms	1,250 req/s	< 1 GB	< 50 MB
ST-BERT (×3 models)	231 ms	48 req/s	9.8 GB	1.26 GB
ICISH Full (12-layer)	87 ms	164 req/s	4.1 GB	482 MB
ICISH Distilled (6-layer)	47 ms	298 req/s	2.4 GB	276 MB

Table 6: Production deployment metrics. ICISH Full reduces GPU memory 58% and boosts throughput 3.4× vs. running three independent BERT models. The distilled model achieves sub-50ms latency, exceeding the <100ms SLA target.

6.4. Enterprise Deployment Outcomes

ICISH was deployed at a large e-commerce partner handling ~24,000 daily interactions across email and live chat. After an eight-week production period, the following KPIs were measured against the prior rule-based system:

KPI	Before (Baseline)	After (ICISH)	Improvement
Avg. Resolution Time	18.4 min	10.5 min	−43%
CSAT Score (1–5)	3.2	4.2	+31%
First-Contact Resolution Rate	61%	79%	+18 pp
KPI	Before (Baseline)	After (ICISH)	Improvement
Agent Escalation Rate	38%	21%	−17 pp
Avg. Ticket Handle Time	7.2 min	4.3 min	−40%
Intent Mis-routing Rate	19%	3.2%	−16 pp

Table 7: Eight-week enterprise pilot results. All improvements statistically significant ($p < 0.01$, paired Wilcoxon test). pp = percentage points. CSAT = Customer Satisfaction Score.

VII. LIMITATIONS AND ETHICAL CONSIDERATIONS

7.1. Limitations

ICISH has several notable limitations. First, the 64-class intent taxonomy requires retraining when new product lines emerge. Second, RAG quality depends heavily on knowledge base completeness — stale documentation produces factually incorrect responses that confidence thresholding cannot always catch. Third, performance degrades beyond the 12 supported languages. Fourth, the 2.7% performance gap between distilled and full models is non-trivial for highstakes interactions (e.g., financial transactions), where the full 12-layer model should be preferred.

7.2. Bias and Fairness

Pre-trained language models encode societal biases from training corpora [Blodgett et al., 2020]. A preliminary bias audit found a 2.1% F1 gap between formal and informal register subsets of CLINC150 (97.8% vs. 95.7%), indicating room for improvement. Human evaluation was conducted by English-speaking experts in a single cultural context, limiting generalizability across diverse customer populations. Future deployments should commission culturally situated evaluation protocols.

7.3. Privacy

ICISH processes sensitive customer data including personal identifiers and financial information. The NER component explicitly identifies and masks ACCOUNT_NUMBER and TRANSACTION_ID entities before any data reaches the generation module or logging layer. All training data was publicly available or processed under data processing agreements with full anonymization. Production deployments must comply with GDPR, CCPA, and applicable regulations.

VIII. CONCLUSIONS

This paper introduced ICISH, a unified NLP framework for intelligent customer insight and support automation. By consolidating sentiment analysis, intent classification, NER, and response generation within a shared transformer architecture enhanced by a novel cross-task fusion mechanism, ICISH achieves state-of-the-art performance across all evaluated tasks while meeting strict production latency requirements.

The cross-task fusion mechanism is the architectural centerpiece differentiating ICISH from prior multi-task approaches: by propagating soft task predictions into response generation and encoder attention, the system produces responses measurably more relevant, fluent, and empathetic. The ablation study confirms fusion contributes 0.9–1.4 percentage points of absolute F1 improvement across all tasks. The eight-week enterprise deployment validated practical value: 43% faster resolution, 31% higher CSAT, and a halved escalation rate represent substantial operational improvements.

Future Work

Several directions merit further investigation. First, multilingual extension via mBERT or XLM-RoBERTa [Conneau et al., 2020] would expand applicability to global customer bases. Second, multimodal input support — product images and screenshots attached to support tickets — through vision-language models could improve technical support diagnostic accuracy. Third, continual learning strategies for evolving product catalogs without catastrophic forgetting are a high-priority enterprise need. Fourth, formal fairness auditing with debiasing interventions is essential for responsible deployment. Finally, exploring LLM backbones (LLaMA-3, Mistral) as drop-in replacements may further improve generation quality.

ACKNOWLEDGEMENTS

The author thanks the faculty of the Department of Computer Science and Information Technology at Chhatrapati Shivaji Maharaj University for their guidance. The author also acknowledges the open-source NLP community — particularly the HuggingFace team — for the models, datasets, and tools that made this research possible.

REFERENCES

[1] [Arkhipov et al., 2019] Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning Multilingual Transformers for LanguageSpecific NER. Proc. 7th Workshop on Balto-Slavic NLP, 2019, pp. 89–93.

[2] [Blodgett et al., 2020] Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (Technology) is Power: A Critical Survey of Bias in NLP. Proc. ACL 2020, pp. 5454–5476.

[3] [Brown et al., 2020] Brown, T., Mann, B., Ryder, N., et al.: Language Models are Few-Shot Learners. Advances in NeurIPS 33, 2020, pp. 1877–1901.

[4] [Chen et al., 2019] Chen, Q., Zhuo, Z., Wang, W.: BERT for Joint Intent Classification and Slot Filling. arXiv:1902.10909, 2019.

[5] [Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., et al.: Unsupervised Cross-lingual Representation Learning at Scale. Proc. ACL 2020, pp. 8440–8451.

[6] [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proc. NAACL-HLT 2019, pp. 4171–4186.

[7] [Johnson et al., 2021] Johnson, J., Douze, M., Jégou, H.: Billion-Scale Similarity Search with GPUs. IEEE Trans. Big Data 7(3), 2021, pp. 535–547.

[8] [Joshi et al., 2020] Joshi, M., Chen, D., Liu, Y., et al.: SpanBERT: Improving Pre-training by Representing and Predicting Spans. Trans. ACL 8, 2020, pp. 64–77.

[9] [Joulin et al., 2017] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. Proc. EACL 2017, pp. 427–431.

[10] [Kim, 2014] Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proc. EMNLP 2014, pp. 1746–1751.

[11] [Koreeda and Manning, 2021] Koreeda, Y., Manning, C.D.: ContractNLI: A Dataset for Document-level NLI for Contracts. Findings of EMNLP 2021, pp. 3140–3156.

[12] [Larson et al., 2019] Larson, S., Mahendran, A., Peper, J., et al.: An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. Proc. EMNLP-IJCNLP 2019, pp. 1311–1316.

[13] [Lewis et al., 2020] Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in NeurIPS 33, 2020, pp. 9459–9474.

[14] [Liu and Lane, 2016] Liu, B., Lane, I.: Attention-Based RNN Models for Joint Intent Detection and Slot Filling. Proc. Interspeech 2016, pp. 685–689.

[15] [Liu et al., 2019] Liu, X., He, P., Chen, W., Gao, J.: Multi-Task Deep Neural Networks for Natural Language Understanding. Proc. ACL 2019, pp. 4487–4496.

[16] [Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed Representations of Words and Phrases. Advances in NeurIPS 26, 2013, pp. 3111–3119.

[17] [OpenAI, 2023] OpenAI: GPT-4 Technical Report. arXiv:2303.08774, 2023.

[18] [Pang and Lee, 2008] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retrieval 2(1-2), 2008, pp. 1– 135.

[19] [Pennington et al., 2014] Pennington, J., Socher, R., Manning, C.D.: GloVe: Global Vectors for

Word Representation. Proc. EMNLP 2014, pp. 1532–1543.

[20] [Peters et al., 2018] Peters, M., Neumann, M., Iyyer, M., et al.: Deep Contextualized Word Representations. Proc. NAACL 2018, pp. 2227–2237.

[21] [Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a Distilled Version of BERT. arXiv:1910.01108, 2019.

[22] [Sun et al., 2019] Sun, C., Qiu, X., Xu, Y., Huang, X.: How to Fine-Tune BERT for Text Classification? CCL 2019, pp. 194–206.

[23] [Tang et al., 2016] Tang, D., Qin, B., Liu, T.: Aspect Level Sentiment Classification with Deep Memory Network. Proc. EMNLP 2016, pp. 214–224.

[24] [Vinyals and Le, 2015] Vinyals, O., Le, Q.V.: A Neural Conversational Model. ICML Deep Learning Workshop 2015.

[25] [Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., et al.: Transformers: State-of-the-Art NLP. Proc. EMNLP 2020 (System Demos), pp. 38–45.

[26] [Xu et al., 2020] Xu, H., Liu, B., Shu, L., Yu, P.S.: BERT Post-Training for Review Reading Comprehension and ABSA. Proc. NAACL 2020, pp. 2324–2335.