UDC 004.89:368

HYBRID ORCHESTRATION OF MODELS FOR AUTOMATED INSURANCE CORRESPONDENCE PROCESSING IN INFORMATION SYSTEMS

Teliukov D.S.

e-mail: dmytro.teliukov@nure.ua

Scientific supervisor – Ph. D., Associate Professor Imanhulova Z.A.

Kharkiv National University of Radio Electronics, CMIT Dept.

Kharkiv, Ukraine

The paper presents a hybrid architecture for automated insurance correspondence processing. The approach combines deterministic extraction, machine learning routing, local large language models, optional external language refinement, and a mandatory verification layer. Deterministic methods identify stable entities such as policy numbers, dates, amounts, addresses, and insurer reference numbers, while machine learning supports primary intent routing. A local large language model is activated only for low-confidence or attachment-heavy cases. An external model is restricted to stylistic rewriting of a minimized case summary and is not allowed to introduce new facts. The final reply is generated from templates and verified before release.

Insurance correspondence in modern information systems includes customer e-mails, broker messages, insurer notifications, and attachment texts extracted from PDF or OCR pipelines. These documents are heterogeneous in both structure and meaning. Some messages contain clearly identifiable requisites such as a policy number, date, amount, address, or insurer reference number, that is, an internal claim, case, or tracking number assigned by the insurer to register the received request, whereas many customer e-mails are written in free form, contain incomplete data, include scanned attachments, or express the request only implicitly. Under such conditions, the use of only one method, whether deterministic rules, machine learning, or a large language model, does not provide a sufficient balance between quality, latency, explainability, and data protection [1, 2].

In this paper, deterministic methods are understood as rule-based procedures and regular expressions for extracting stable textual patterns from customer e-mails and attachment texts. Machine learning refers to classical statistical models for text classification, for example approaches based on TF-IDF representations and logistic regression. Large language models are neural sequence models capable of semantic interpretation and language generation. The proposed solution combines these components into one orchestrated pipeline rather than treating them as competing end-to-end solutions.

The target architecture follows the local-first processing principle. This means that the initial analysis of customer e-mails, attachment texts, and sensitive requisites is performed inside the local processing contour. In this study, security refers to confidentiality, controlled processing, and restricted propagation of insurance-related customer data. External large language models, if they are used, must not analyse raw source documents and must not become a source of new case facts.

The first layer of the system is ingestion. It receives the e-mail subject, sender, body, attachment text, and auxiliary metadata, then normalizes the input into one case object. The second layer is deterministic extraction, which identifies policy numbers, multiple policy identifiers, dates, amounts, insurer reference numbers used by the insurer for internal request registration and tracking, addresses, and other structured entities. For stable entities, this layer provides high precision, low latency, and clear traceability, which is why it is treated as the primary source of locally confirmed facts [1, 3].

The routing layer combines a rule-based router with a machine learning classifier. The rule-based router checks characteristic patterns related to common scenarios such as address change, payment inquiry, additional contribution, contract consolidation, or insurer autoreply. If these rules are insufficient for a confident decision, a machine learning classifier based on statistical text features performs primary intent routing.

A key element of the architecture is the confidence gate. It decides whether deterministic extraction and primary routing are sufficient for the current case or whether a more expensive semantic layer must be activated. If the intent is identified with high confidence and critical fields are present, the system proceeds directly to structured case construction and response generation without involving a large language model. If the case remains weakly structured, if attachment text contains essential information, or if critical fields are still missing after the initial analysis, the orchestrator activates a local large language model.

The local language model is not the default first step for all correspondence. It is used as a semantic refinement component for low-confidence cases, attachment interpretation, noisy OCR text, and gap filling of critical fields. After the analysis stage, the system builds a structured case representation containing the final intent, extracted fields, missing required fields, a confidence estimate, route metadata, and the whitelist of confirmed facts allowed to appear in the reply. Here, raw attachments mean the original attachment documents or their unfiltered extracted texts, whereas a minimized case summary contains only the intent, confirmed facts, missing fields, and the communication goal [3, 4].

Response generation follows a template-first strategy. A template either confirms receipt of the request, asks for missing required information, or produces an internal note instead of a customer-facing reply. For example, if a case related to address change lacks a policy number, the system must ask for the customer's VSNR; if an additional contribution request lacks the amount, the reply must explicitly request this value. The optional external language refinement layer may improve style and fluency, but it receives only the minimized case summary and must not add new facts.

The essential component of the proposed system is the verification level. A verifier checks the final reply before release. It validates intent consistency, explicit requests for all missing required fields, absence of unsupported new facts, and compliance with scenario-specific business rules. If the verifier detects a violation, the system falls back to the safer template-based response or escalates the case to manual review. Such a mechanism reduces hallucination risks and makes the generative part of the pipeline controllable [2, 5].

The practical value of the proposed approach lies in creating a research prototype that supports justified distribution of responsibilities among deterministic methods, machine learning, local large language models, optional external language refinement, and verification procedures. The hybrid orchestration principle makes it possible to combine high precision on structured requisites, acceptable routing quality, semantic flexibility for difficult cases, and controlled final response generation. Therefore, the proposed architecture can serve as a practical basis for further development of insurance information systems oriented toward automated correspondence processing.

References:

1. Jurafsky D., Martin J.H. Speech and Language Processing. 3rd ed. draft. Stanford University, 2024. URL: https://web.stanford.edu/~jurafsky/slp3/ (accessed: 12.03.2026).

2. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Harlow: Pearson, 2021. 1166 p.

3. Kleppmann M. Designing Data-Intensive Applications. Sebastopol: O’Reilly Media, 2017. 616 p.

4. Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.

5. EIOPA. Consumer Trends Report. Frankfurt am Main: European Insurance and Occupational Pensions Authority, 2024. URL: https://www.eiopa.europa.eu/publications/consumer-trends-report-2024_en (accessed: 12.03.2026).