

Automating Insurance Benefit Interpretation: A Domain-Intelligence Pipeline for Document Extraction, Benefit Classification, and Plan Recommendation

Susil Kumar Sahu

Submitted:02/11/2022

Revised:17/12/2022

Accepted:29/12/2022

Abstract—Large volumes of benefit documentation are processed by insurance organizations, encompassing things, for example, deductibles, copayments, coinsurance, exclusions, service limits and the pharmacy benefits, and a ton of the information is yet to be comprehended consistently before plans can be contrasted or set up within enterprise sales systems. In this systematic review, we compile 28 sources in document image analysis, named entity recognition, pre-trained language models, text categorization, weak supervision, recommender systems and explainable AI to formulate a domain-intelligence pipeline for benefit interpretation. The pipeline includes document ingestion, content extraction, insurance domain entity recognition, benefit classification, semantic normalization, validation, and plan recommendations. A domain taxonomy enables the mapping of carrier terminology to standardized benefit categories, validation rules are applied to disambiguate incomplete or ambiguous values and send them for human review, and the normalized representation is matched to a versioned product catalog to generate ranked candidate plans, including a confidence score, and a sourced linked explanation. The evidence reviewed shows that a form understanding task can be enhanced by a layout-aware pre-training that can boost entity-level F1 score from 78.95% to 84.20%, domain-adapted language models outperform hand-curated training sets by 12.24 points on a biomedical question answering task, and that weak supervision can generate data sets that are as good as the hand-curated training sets within 3.60 points. Provenance and extracted values, normalized representations, product versions and recommended histories are preserved for auditability. The synthesis confirms that document intelligence and domain-specific normalization can help minimize manual interpretation and ensure consistency and scalability in regulated insurance sales operations.

Keywords—*insurance benefit interpretation, document intelligence, named entity recognition, benefit classification, semantic normalization, plan recommendation, explainable artificial intelligence, human-in-the-loop validation, auditability*

I. INTRODUCTION

Insurance carriers share benefit details with consumers via summaries of benefits, plan brochures, and contract schedules for coverage, deductibles, copayments, coinsurance, exclusions, service limits, and pharmacy tiers and the terminology used by the carriers to describe these details. This information is manually read and entered into plans for comparison and configuration in enterprise sales systems, a process that is slow, inconsistent and hard to audit between reviewers. In literature, the value added from AI in the insurance industry has been studied including the handling of unstructured documents [1].

The challenge is exacerbated by the regulated and high-stakes nature of insurance sales, where mistakes in a copayment or an omission regarding an exclusion could lead to financial losses or regulatory penalties. Similar needs have been studied in the clinical setting where machine learning needs to be accurate, validated, and supervised by a human operator before being deployed [2]. Performance claims must be backed by evidence of reliability and uncertain outputs must be escalated and not accepted as a given.

The closer comparison is the area of electronic health record analysis, where a collection of heterogeneous, nonstandardized documents is transformed into a structure that is appropriate for prediction and decision making [3]. The same approach is taken with benefit documents and the literature is arranged in a seven-stage pipeline. The review protocol is described first, then each stage is discussed in turn,

*Solutions Engineer Executive Advisor
Elevance Health, Inc*

finally comparative evidence, limitations and governance are discussed. The history of the reviewed methods is presented in a summary list in Fig. 1.

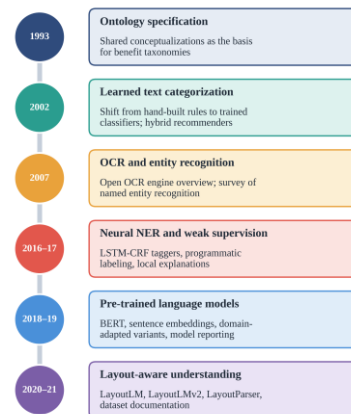


Fig. 1. Chronological evolution of the methods reviewed for each pipeline layer.

II. REVIEW METHODOLOGY AND PIPELINE OVERVIEW

The screening of supplied corpus (28 sources) for relevance to one of seven pipeline stages, coding by technical concept and synthesising the content stage by stage, were conducted in a systematic review process. Coding has been done using an explicit definition of a

shared conceptualization, so that terms like deductible, coinsurance, and out-of-pocket maximum have been coded as formally defined concepts rather than free text [4]. The corpus spans 1993 to 2021, with 6 sources published before 2010, 7 between 2015 and 2018, and 15 between 2019 and 2021. The mapping of literature to pipeline concepts are given in Table I and the final architecture in Fig. 2.

TABLE I. MAPPING OF REVIEWED LITERATURE TO PIPELINE CONCEPTS

Pipeline Layer	Concept Addressed	Refs.
Context	AI value and regulated machine learning	[1]–[3]
Knowledge	Shared conceptualization	[4]
Extraction	OCR and layout analysis	[5], [6]
Extraction	Layout-aware pre-training	[7], [8]
Recognition	Entity recognition	[9], [10]
Recognition	Domain language models	[11]–[14]
Classification	Categorization, weak supervision	[15], [16]
Normalization	Semantic similarity	[17]
Validation	Human–AI interaction	[18]
Recommendation	Paradigms and evaluation	[19]–[22]
Explanation	Local and global XAI	[23]–[26]
Governance	Model and dataset documentation	[27], [28]

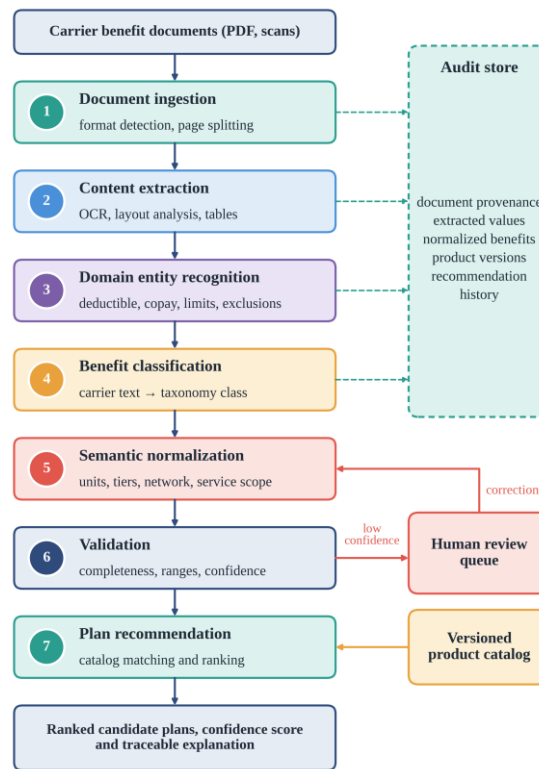


Fig. 2. Architecture of the proposed domain-intelligence pipeline for insurance benefit interpretation.

III. DOCUMENT INGESTION AND CONTENT EXTRACTION

A. Optical Character Recognition and Layout Analysis

After detecting the format, splitting the pages and preparing the image's, scanned pages are recognized as text via optical character recognition (OCR). The open-source Tesseract engine developed from 1984 to 1994 at Hewlett-Packard and released in 2005 was characterized as a series of connected component analysis, line/word recognition, and adaptive classifier [5]. This type of engine provides character-level text, but does not generate any representation of tables, columns or headers.

In addition to character recognition, layout analysis is necessary in order to segment the benefit schedules, which are dominated by tables, multi-column layouts, and footnotes. The unified toolkit for deep learning-based document image analysis, LayoutParser, packages pre-trained layout detection models, OCR interfaces and annotation tools behind a common application programming interface (API), through which table regions and text blocks detected by the deep learning models are passed to downstream document recognition components [6].

B. Layout-Aware Pre-training

Words and page coordinates are treated jointly by layout-aware language models. The large model was an encoder with the pre-trained model of LayoutLM, which added 2D position embeddings to a Bidirectional Encoder Representations from Transformers (BERT) style encoder and obtained a model with 113 million or 343 million parameters, achieving an entity-level F1 score of 78.95 percent on the FUNSD form understanding benchmark [7].

The spatial-aware self-attention mechanism and image features were added for the extension of this design in LayoutLMv2. The attention score α_{ij} of tokens i and j is obtained by computing the query-key projection $W^Q x_j + b_{ij}$ and key-query projection $W^K x_i$, where the bias b_{ij} is a learnable term, and the sequence offset and horizontal/vertical page offset are multiplied together, and the produced attention scores are scaled by the square root of the head dimension d_h , as in (1). In Fig. 3 and Table II, the large variant, with 426 million parameters, achieved an F1 score of 84.20, 96.01 and 97.81 percent on FUNSD, CORD and SROIE respectively, while gaining a score of 5.25 points from FUNSD with a 24 percent increase in parameters, as compared.

$$\alpha_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_h}} + b_{ij} \quad (1)$$

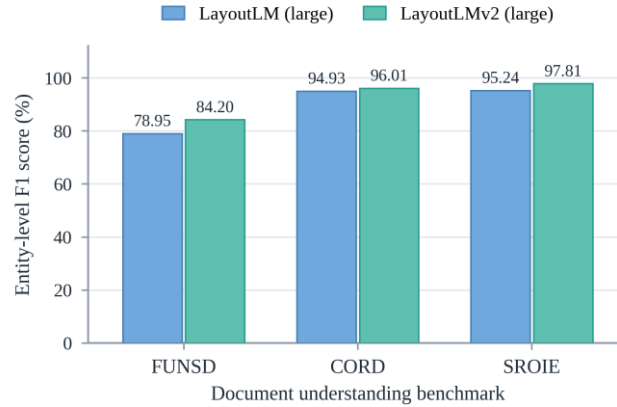


Fig. 3. Entity-level F1 scores of LayoutLM and LayoutLMv2 on three document understanding benchmarks.

TABLE II. COMPARISON OF CONTENT EXTRACTION APPROACHES

Approach	Inputs Used	Pre-training Data	Params . (M)	FUNSD F1 (%)
Tesseract	Character images	Not applicable	n/a	n/a
Layout-Parser	Page images, regions	Pre-trained detectors	n/a	n/a
LayoutLM (large)	Text, 2-D position	11 M pages	343	78.95
LayoutLMv2 (large)	Text, 2-D position, image	11 M pages	426	84.20

n/a = no comparable figure reported. M = million.

IV. ENTITY RECOGNITION AND DOMAIN LANGUAGE MODELS

A. Named Entity Recognition

The insurance-domain entity recognition identifies spans like deductible, copayment, coinsurance, limit of visits, network tiers, and exclusion clauses. Rule-based, feature-engineered, and supervised methods have been all explored in the context of named entity recognition (NER), and the fact that the many aspects of named entities (types, domains, and measure) that make it hard to compare among systems has been highlighted [9]. A general-purpose tagger must then be modified for monetary and percentage entities and temporal entities which are tightly coupled to specific services.

The need for hand-crafted features was removed by neural sequence taggers. A bidirectional long short-

term memory (LSTM) network with conditional random field (CRF) output layer was constructed with no language-specific resources and obtained F1 scores of 90.94 for English, 85.75 for Spanish, 81.74 for Dutch and 78.76 for German, as shown in Fig. 4(a). For an input sequence X, tagged sequence y is scored by adding together the tag-transition scores A and the per-token scores P generated by the network as defined in (2) [10].

$$s(X, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (2)$$

B. Domain-Adapted Language Models

Pre-trained transformers were then provided as the next step. BERT is a model where a subset M of the input tokens is masked, and the tokens in M are masked LMs using the remaining context $X_{-}(\backslash M)$ through

minimization of (3). Two models with 110 million and 340 million parameters were released and the GLUE score was increased to 80.5 percent and the question-answering F1 on SQuAD v1.1 was increased to 93.2 [11].

$$L_{MLM} = -\sum_{i \in M} \log P(x_i | X_{\setminus M}) \quad (3)$$

It is always a good advantage to adapt to corpora that are specialized. Continued pre-training on about two million clinical notes retrieved from the MIMIC-III database resulted in embeddings which were then evaluated on a variety of clinical entity-recognition and inference tasks, with better results than general-domain and biomedical embeddings [12]. This comparison is

facilitated by the fact that documents written by insurers have different wordings and terminology to those of general text, so continued pre-training is supported.

The improvement of the BioBERT model in the biomedical entity recognition task was 0.62 F1 points over the previous best biomedical entity recognition model, and the relation extraction and the question answering tasks gained 2.80 F1 points and 12.24 mean reciprocal rank points, respectively, as seen in Fig. 4(b) [13] from the data of the PubMed abstracts (4.5 billion words) and PubMed Central full-text articles (13.5 billion words) used to pre-train the model.

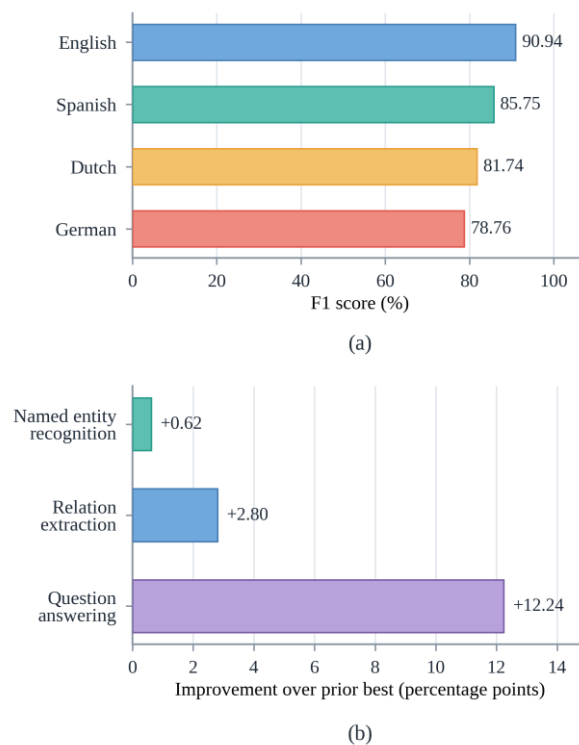


Fig. 4. Reported performance of (a) an LSTM-CRF tagger across four languages and (b) domain-adapted pre-training gains.

The clauses, defined terms, and exclusions in the legal text are similar to those found in policy texts, and LEGAL-BERT had been pre-trained on ~12 GB of English laws, court decisions, and contracts. The pre-training of general BERT, domain pre-training from

scratch and their combinations were compared and competitive results were reported for a lighter variant with only 33 percent of the parameters of the standard model [14]. These models are compared with respect to the characteristics in Table III.

TABLE III. COMPARISON OF PRE-TRAINED LANGUAGE MODELS

Model	Domain	Pre-training Corpus	Reported Benefit
BERT	General	Books and Wikipedia, 3.3 B words	GLUE 80.5%; SQuAD F1 93.2
Clinical BERT	Clinical	MIMIC-III, about 2 M notes	Gains on clinical NER and inference
Bi-oBERT	Biomedical	PubMed and PMC, 18 B words	+0.62 NER, +2.80 RE, +12.24 QA
LE-GAL-BERT	Legal	About 12 GB legal text	Competitive at 33% of size

B = billion; M = million; RE = relation extraction; QA = question answering.

V. BENEFIT CLASSIFICATION AND SEMANTIC NORMALIZATION

A. Taxonomy-Driven Classification

Once extracted each benefit entity is mapped to a class in the domain taxonomy, such as inpatient hospital, specialist visit, generic drug tier, etc. The

problem of text categorization has been formalized as learning a function that maps a document to one of the predefined categories, and manually written rules have been superseded by trained classifiers [15]. The posterior probability of class k is computed in (4) using the softmax function, where w_k is the weight vector of class k , and K the number of classes.

$$P(c_k|h) = \frac{\exp(w_k^T h)}{\sum_{j=1}^K \exp(w_j^T h)} \quad (4)$$

B. Weak Supervision and Semantic Matching

Benefit data is hard to come by due to the need for domain experts to annotate it. Snorkel uses the generation of labels by labeling functions to overcome this limitation: noisy outputs from these functions are combined together by a generative model. Subject matter experts reported building models 2.8 times faster, with an average improvement of 45.5 percent relative to seven hours of hand labeling and an average improvement of 132 percent over previous heuristic approaches, and models were able to achieve performance within 3.60 percent of large hand-curated models, which were used for training [16].

In semantic normalization, the different versions of the member cost sharing carrier phrases are normalized to one standard cost sharing category. Sentence-BERT generates fixed length sentence embeddings where the cosine similarity can be compared directly and the time needed to get the most similar pair of sentences among 10,000 sentences was shortened from about 65 hours using BERT to about 5 seconds [17]. For a phrase embedding u , it is then assigned to a taxonomy entry k , such that the embedding v_k , which maximizes the cosine similarity, as defined in (5), is closest to u . Representative taxonomy entries are listed in Table IV.

$$\hat{c} = \operatorname{argmax}_k \frac{u \cdot v_k}{\|u\| \|v_k\|} \quad (5)$$

TABLE IV. ILLUSTRATIVE BENEFIT TAXONOMY AND NORMALIZATION RULES

Standard Category	Example Carrier Wording	Value Type	Validation Rule
Medical deductible	Annual or calendar-year deductible	Currency, per person or family	Convert to annual amount
Primary care copayment	Office visit, PCP visit	Currency per visit	Append network tier
Specialist coinsurance	Member pays, specialist	Percent	Bound between 0 and 100
Out-of-pocket maximum	OOP max, annual limit	Currency	Not below deductible
Service limit	30 visits per year	Count and period	Normalize period to year
Pharmacy tier	Tier 2, preferred brand	Currency or percent	Separate retail and mail
Exclusion	Not covered	Boolean and clause	Link to source span

Design example for the proposed taxonomy; not drawn from the reviewed sources.

VI. VALIDATION AND HUMAN REVIEW

Validation helps to prevent silent failure. Rules check that required fields are filled, values are reasonable, units and periods are consistent, and that cross-field limits are followed such as not including out of pocket maximum before deductible. Good guidelines for human-AI interaction recommend communicating AI system confidence and correcting

errors efficiently [18]. A composite confidence q is then constructed as the weighted geometric mean of the extraction confidence q_{ext} , the classification confidence q_{cls} and the normalization confidence q_{nm} , with the exponents α , β and γ such that $\alpha + \beta + \gamma = 1$ and shown in (6). Records whose quality parameter " q " is lower than a threshold " τ ", or records for which a rule is violated will be passed to the human for examination, see Fig. 5.

$$q = q_{\text{ext}}^{\alpha} q_{\text{cls}}^{\beta} q_{\text{nm}}^{\gamma}, \alpha + \beta + \gamma = 1 \quad (6)$$

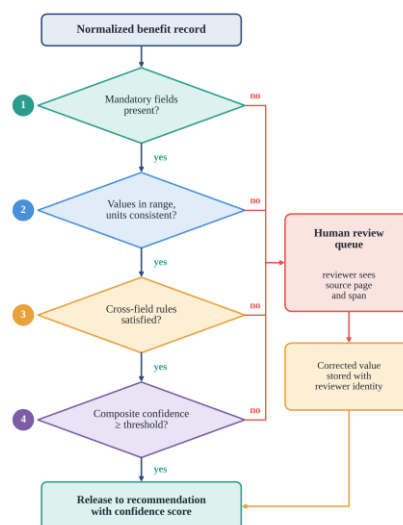


Fig. 5. Validation and human review workflow for normalized benefit records.

VII. PLAN RECOMMENDATION

A. Recommendation Paradigms

A recommender makes a match between the normalized benefit representation and the internal product catalog. Content-based, collaborative, and hybrid recommender systems have been distinguished, and there are limitations such as limited content analysis, overspecialization, and new user and new item problems [19]. Content-based matching and knowledge-driven matching is preferred over collaborative filtering because interaction histories are not always available for new plan configurations in insurance sales.

Individual weaknesses are compensated for by hybridization. The seven hybridization methods that have been found are: weighted, switching, mixed, feature combination, cascade, feature augmentation, and meta-level [20]. The component scores are kept separate for explanation, so a weighted hybrid is chosen for plan ranking. A plan p with J component scores $s_j(p)$ and $\lambda_j \geq 0$ such that $\sum_{j=1}^J \lambda_j = 1$, gives an overall score in (7) and the complete design is shown in Fig. 6.

$$S(p) = \sum_{j=1}^J \lambda_j s_j(p) \quad (7)$$

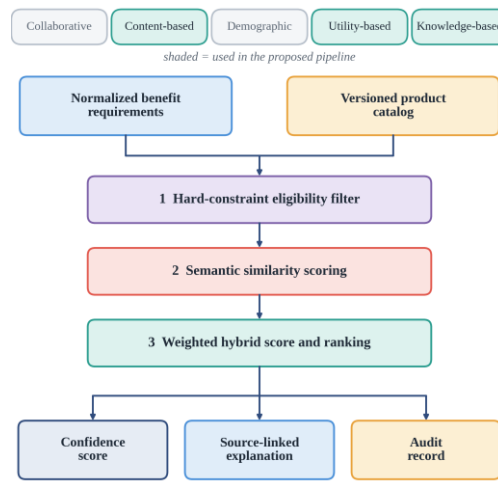


Fig. 6. Hybrid recommendation design combining hard-constraint filtering, semantic similarity, and weighted scoring.

B. Constraint Handling and Evaluation

Explicit constraints are applied prior to ranking in the case of knowledge-based and utility-based recommenders, which is appropriate for mandatory constraints like a maximum deductible or drug tier, as

described in the second edition of the Recommender Systems Handbook [21]. The relevance rel_i of the item at rank i is discounted (based on a logarithmic discount) and divided by the ideal value, as in (8), and metrics such as normalized discounted cumulative gain (NDCG) measure the quality of a ranking.

$$NDCG@k = \frac{\sum_{i=1}^k \frac{rel_i}{\log_2(i+1)}}{IDCG@k} \quad (8)$$

Dense representations are learned from heterogeneous content, and deep learning-based recommenders were surveyed in terms of multilayer perceptrons, autoencoders, convolutional networks, recurrent networks, and attention models. The

paradigms are contrasted against the data requirements, cold start behavior, and explainability criteria in Table V and it is determined that the paradigm best suited for the benefit matching task is hybrid (constraint first).

TABLE V. QUALITATIVE COMPARISON OF RECOMMENDATION PARADIGMS

Paradigm	Data Requirement	Cold-Start Sensitivity	Explainability	Fit to Task
Content-based	Item attributes	Low	High	Good
Collaborative	Interaction history	High	Low	Poor
Knowledge or utility	Rules, constraints	Low	High	Good
Weighted hybrid	Attributes, constraints	Low	High	Best
Deep learning	Large labeled data	Medium	Low	Fair

Ratings are a qualitative synthesis of the reviewed recommender surveys.

VIII. EXPLAINABILITY AND AUDITABILITY

Each recommendation should have a clear and concise explanation. In the neighborhood of x defined by a proximity kernel π_x , the prediction of a model f for the specific instance x is explained by local

$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (9)$$

The motivations for explanation (justification of decisions, control, improvement, and discovery) were mapped and analyzed, along with scope and timing of explanation [24]. Justification and control reign supreme in regulated sales because every recommended plan is subject to audit and review by customers, auditors and regulators.

Taxonomies and challenges of XAI were unified under responsible AI, transparent models were separated from post hoc XAI techniques, and explainability was correlated with fairness, accountability and privacy [25]. Therefore, explanations of the benefits recommendations need to be faithful, understandable to sales personnel and kept as a part of the audit record.

IX. DISCUSSION

A design principle is created through comparative evidence. In the context of high-stakes decisions, interpretable models were advocated over explained black-box models as post hoc explanations may not be faithful to the model being explained [26]. Errors can be detected by rules and review at the black-box components (perception and tagging), while ranking,

surrogate methods which involve fitting an interpretable model g from a class G to the local neighborhood, with a local fidelity loss L being traded off against its complexity $\Omega(g)$ as defined in (9) [23]. For this proposed pipeline, the source document spans and the rule's results are interpreted features.

eligibility and thresholds remain transparent. The best gains were found in domain adaptation and supervision (12.24 points on QA and 132 percent on weak supervision) and the least were in entity recognition (0.62 points), thus suggesting that the most effective approach is to focus on labeled data and constraints instead of architecture.

Documentation completes transparency. Model cards are used to record each extraction, classification, and ranking component and its version and operating threshold, and are provided on the model cards reporting intended use, evaluation conditions, and performance across relevant subgroups [27]. Evaluation should be broken down into categories of carrier, document format and benefit category, as benefits can be masked with aggregate accuracy if there are system failures on specific document formats.

The same care is needed for datasets. Datasets are documented with datasheets that record the motivation, composition, and how the data was collected and applied, which can be used to trace provenance from any normalized value to the source corpus and annotation process [28]. Therefore, the mechanisms are compared in Table VI, wherein the provenance,

extracted values, normalized representations, product versions and recommendation histories are stored by the auditability layer. A weakness of the present synthesis is that there are no insurance specific

benchmarks reported in the literature, and all of the performance figures reported come from neighboring areas.

TABLE VI. EXPLANATION AND DOCUMENTATION MECHANISMS IN THE PIPELINE

Mechanism	Scope	Role in the Pipeline
Source-span highlighting	Local	Shows page and span of each extracted value
Rule trace	Local	Lists validation and eligibility rules that fired
Score break-down	Local	Reports each weighted term of the plan score
Local surrogate model	Local	Explains learned classifier decisions
Model card	Global	Records scope, metrics, version, threshold
Datasheet	Global	Records corpus provenance and annotation

X. CONCLUSION

To create a structured, auditable plan recommendation from insurance benefit documents, a seven-stage domain-intelligence pipeline was synthesized, based on 28 sources. It was demonstrated that all were mutually compatible with layout-aware extraction, domain-adapted language models, taxonomy-driven classification, weak supervision, semantic normalization, rule-based validation, and weighted hybrid matching, and the evidence reported for adjacent domains built the case for the technical feasibility of each stage. Numerically, English sequence tagging scores of $F1 > 90$ and receipt understanding scores of $F1 > 90$ were reported, with domain adaptation systems contributing up to 12.24 points, and weak supervision resulting in a 2.8-fold increase in model construction speed. The key contributions are: a stage-wise mapping of the literature to pipeline concepts, a taxonomy-driven normalization design, a confidence-based routing rule for human review, and an auditability layer to which provenance is preserved throughout the entire pipeline from source page to recommendation.

It was found that the ways by which automation is acceptable are the uncertain records, the traceable explanations, and the versioning of the provenance when they are reviewed by humans. Prospective research efforts will focus on insurance-specific annotated corpora, reference tasks for benefit extraction, confidence score calibration, and reviewer

workload and recommendation accuracy in a real sales process.

REFERENCES

- [1] M. Riikkinen, H. Saarijärvi, P. Sarlin, and I. Lähteenmäki, "Using artificial intelligence to create value in insurance," *Int. J. Bank Marketing*, vol. 36, no. 6, pp. 1145–1168, 2018, doi: 10.1108/IJBM-01-2017-0015.
- [2] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England J. Med.*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/NEJMr1814259.
- [3] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/JBHI.2017.2767063.
- [4] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquisition*, vol. 5, no. 2, pp. 199–220, 1993, doi: 10.1006/knac.1993.1008.
- [5] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit.*, 2007, pp. 629–633, doi: 10.1109/ICDAR.2007.4376991.
- [6] Z. Shen et al., "LayoutParser: A unified toolkit for deep learning based document image analysis," in *Document Analysis and Recognition – ICDAR*

- 2021, J. Lladós, D. Lopresti, and S. Uchida, Eds. Cham, Switzerland: Springer, 2021, pp. 131–146, doi: 10.1007/978-3-030-86549-8_9.
- [7] Y. Xu et al., “LayoutLM: Pre-training of text and layout for document image understanding,” in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2020, pp. 1192–1200, doi: 10.1145/3394486.3403172.
- [8] Y. Xu et al., “LayoutLMv2: Multi-modal pre-training for visually-rich document understanding,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 2579–2591, doi: 10.18653/v1/2021.acl-long.201.
- [9] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” *Linguisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007, doi: 10.1075/li.30.1.03nad.
- [10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in Proc. 2016 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2016, pp. 260–270, doi: 10.18653/v1/N16-1030.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [12] E. Alsentzer et al., “Publicly available clinical BERT embeddings,” in Proc. 2nd Clin. Natural Lang. Process. Workshop, 2019, pp. 72–78, doi: 10.18653/v1/W19-1909.
- [13] J. Lee et al., “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.
- [14] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androustopoulos, “LEGAL-BERT: The muppets straight out of law school,” in Findings Assoc. Comput. Linguistics: EMNLP 2020, 2020, pp. 2898–2904, doi: 10.18653/v1/2020.findings-emnlp.261.
- [15] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, 2002, doi: 10.1145/505282.505283.
- [16] A. Ratner et al., “Snorkel: Rapid training data creation with weak supervision,” *Proc. VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017, doi: 10.14778/3157794.3157797.
- [17] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [18] S. Amershi et al., “Guidelines for human-AI interaction,” in Proc. 2019 CHI Conf. Human Factors Comput. Syst., 2019, Paper 3, doi: 10.1145/3290605.3300233.
- [19] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005, doi: 10.1109/TKDE.2005.99.
- [20] R. Burke, “Hybrid recommender systems: Survey and experiments,” *User Modeling User-Adapted Interact.*, vol. 12, no. 4, pp. 331–370, 2002, doi: 10.1023/A:1021240730564.
- [21] F. Ricci, L. Rokach, and B. Shapira, Eds., *Recommender Systems Handbook*, 2nd ed. New York, NY, USA: Springer, 2015, doi: 10.1007/978-1-4899-7637-6.
- [22] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” *ACM Comput. Surveys*, vol. 52, no. 1, Art. no. 5, 2019, doi: 10.1145/3285029.
- [23] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [24] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
- [25] A. Barredo Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
- [26] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

- [27] M. Mitchell et al., “Model cards for model reporting,” in Proc. Conf. Fairness, Accountability, Transparency, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.
- [28] T. Gebru et al., “Datasheets for datasets,” Commun. ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.