A Hybrid AI Agent for Auditable Pharmaceutical Field Engagement Reporting

Keval Shah
Independent Researcher
Arizona, USA

ABSTRACT
During interactions with healthcare professionals, pharmaceutical companies' field teams are able to make valuable engagement observations, but it takes too long to get from their spoken word to an enterprise record – and that data can be incomplete or require manual interpretation. In this study, a hybrid AI agent architecture is proposed, which transforms the field dictation to validated, context-aware, linked engagement records, maintains traceability and human supervision. For the evaluation, a study dataset of 171 field-engagement instances, in which the product is discussed, the customer situation, follow-up actions, medical-information requests, access barriers and compliance-sensitive statements were used. The solution involved speech transcription, language processing, the use of an approved product and customer context for the retrieval, rule-based compliance screening, and agent-assisted record assembly. Data preparation, evaluation and visualization were done in Python and a prototype orchestration layer was responsible for orchestrating the transcription, linking, validation and record completion. The results show that the hybrid design resulted in higher percentages of complete linked records, better handling of the terminology, higher level of compliance preparedness and more consistent output from the evaluation batches compared to a traditional sequential workflow design. The automation had the effect of speeding up the documentation of routine cases without compromising governance, as human reviewers would remain in control in cases where the quality of the automation was questionable. The proposed architecture thus provides a sensible path towards a regulated enterprise record from a field dictation.

General Terms
Artificial Intelligence, Natural Language Processing, Information Systems.

Keywords
Hybrid AI Agent; Pharmaceutical Field Reporting; Dictation Processing; Linked Engagement Records; Compliance Validation.

1. INTRODUCTION
When it comes to pharmaceutical field engagement reporting, it's where the rubber meets the road between human conversation, commercial operations, medical governance and enterprise data quality. A representative can conclude a visit with a brief oral summary of the visit, including any product issues, interests, access issues, words around safety, any requested follow-up, and next step commitments. Transcribing that dictation into a record that can be used is not a straightforward transcription exercise, as the narration may include abridged names of products, an incomplete reference to an organization, informal clinical language and local abbreviations, as well as implicit references to the customer context that was previously known. The challenge for operation is thus to maintain the significance of the event as it occurred while creating a document record that can be searched, posted and reviewed, and associated and managed. Previous speech-to-document work, such as that by Zhou et al. [1] has indicated that automatic transcription can speed up the documentation process, but also produce errors that can be significant if they change specializations and/or relationships. To address the above, the representation of the biomedical vocabulary and contextual language through biomedical language-model pretraining has been explored by Lee et al. [2]. Specialized contextual models, such as those used by Rasmy et al. [3] have been used to benefit structured health-record representation too. While these innovations offer valuable tools, there's one more key component to pharmaceutical field reporting: every statement that is extracted must have an accountable trace from voice input to automated records. A field system must thus be able to identify the intent of the speaker, extract the appropriate approved context, determine what entities were covered, identify any facts that were stated and uncertain mentions, route to review of the latter, and record the appropriate context without losing track of the original evidence. The proposed hybrid AI agent aims to fill this dual need, but not with full autonomy of the content generation.

The main design goal of the present study is to develop an agent that acts as an orchestrator of trusted text processing stages, instead of a monolithic text generating agent. The first step in the architecture is to capture and clean up the dictation, then pass the normalized text to a hybrid reasoning component that finds the customer and product context and matches the entities found in the dictation with controlled enterprise entities, validates and complying with them, and formats the write-back package in an auditable form. This is inspired by the recent trend of agentic healthcare systems where language models can plan, utilize tools and coordinate workflow steps as explored by Qiu et al. [9]. It also addresses concerns that clinical and regulated deployments require realistic assessment not only on the language abilities, but also on the outcome of the workflows as highlighted by Wornow et al. [5]. Integration, monitoring and governance are also prerequisites to a reliable use of generative systems in healthcare, as underlined by work done by Meskó and Topol [7] in their discussion of regulatory oversight. In this present study, those same concepts are adapted to the field of pharmaceutical engagement reporting, where enterprise traceability is crucial, just as much as language proficiency. The hybrid architecture is based on deterministic checks (organizational rules are explicit) and language based reasoning (interpretation is needed). The design calls for unsupported customer facts, changes to approved product information, and sensitive content to be flagged for review before record completion. This results in the following study: Architecture, Record Linkage Quality, Compliance Readiness, Reviewer acceptance and consistency across batches. The idea presented is a practical design pattern to

enable (a) the transition of dictation to enterprise data into linked, reviewable field records, and (b) the establishment of an evidence chain from the dictation to the enterprise data.

2. REVIEW OF LITERATURE

The research related to this problem is in areas such as speech-assisted documentation, biomedical language processing, entity recognition, record representation and new emerging areas of agent-based workflow automation. Although the benefits of direct dictation for saving documentation time have been proven in speech recognition studies, errors due to domain specific terms and context have been reported, for example, by Zhou et al. [1]. Later, biomedical language models saw a greater ability to deal with words in context; the method developed by Lee et al. [2] showed that domain-specific pretraining can aid in several text-mining tasks. More comprehensive studies of biomedical pretrained language models, like the one by Kalyan et al. [4] detail a trajectory towards models that can extract, classify, normalize and organize downstream information. A related stream of interest is the conversion of unstructured text to biomedical concepts that can be linked. Comparing performances has been shown by entity-linking benchmarks to be unreliable if the evaluations are not carried out in the same way, in which case, the benchmark framework proposed by Garda et al. [10] has been developed. Some works have also looked into dense retrieval and relation-aware linking for enhancing the grounding of medical mentions like Singh et al. [12]. These studies are pertinent as the products, therapeutic areas and other organizations and follow-up topics in pharmaceutical field notes are often shortened and should be linked to enterprise identifiers instead of being literal text. Meanwhile, there is also pure language processing, which doesn't have their eye on operational accountability. The literature thus implies that a successful system should incorporate a harmonized pipeline system: a language understanding component and enterprise rules and an explicit control on record linking.

Although large language models have been introduced and adapted to the health care sector recently, the focus has been on standalone text-based tasks, whereas systems that can interface with tools, records, and workflow services are now of interest. Medical investigations of generative models have indicated wide applications in drafting and summarization, retrieval, and decision support, among others as explored in Peng et al.'s work [6]. Agent oriented evaluation has also taken into account the possibility of language models to fully complete multi-step tasks in realistic clinical settings as proposed by Mehandru et al. [8]. As such, these advances add to the argument for an orchestration layer that can be provided with a goal for the user, can locate the appropriate context, select the tools to work with the context, can validate intermediate results and create a final action. The same literature, however, consistently points to the need for limited use as fluent text may still contain information that is beyond the text's intended scope, information that is missing from the text, or information that may not be safe to include. Thus, agentic healthcare commentary has focussed on planning, interaction with tools, reflection, and external checks which are more constrained than are the response generation in the work of Qiu et al. [9]. In the same way, implementation research speaks about practical implementation of integration with other systems as part of the research challenge and not as an engineering issue. There are some examples of agentic design used in a workflow context within healthcare simulation systems, such as the one developed by Barra et al. [13] that combines generative reasoning with the steps of a structured platform. A reporting domain even smaller than the pharmaceutical industry is required for field reporting as the product can include promotional, medical, privacy or safety-sensitive information. The literature suggests a hybrid approach where special purpose language processing is used for interpretation, known constraints are enforced by rules and reviewers are used for dealing with uncertainty and grounding entities. This study continues that trend by developing a comprehensive dictation to contextual linking to compliance validation to a record package that is ready for enterprise write-back.

3. METHODOLOGY

The manuscript describes a stepwise hybrid AI reporting process and an assessment involving 171 de-identified pharmaceutical field-engagement cases. Each was a short dictation-like story, contextual metadata, a set of linked record fields which were expected, and the reviewer-validated results to compare. The processing process started with the acceptance of spoken or text-simulated dictation and the production of a normalized transcript of the dictation with punctuation repair, abbreviation repair and the maintenance of doubtful tokens. The narrative was then broken down into topics of engagement, products mentioned, references to stakeholders, requested actions, access observation, and compliance sensitive phrases by a language-processing layer. This structured interpretation was added to the hybrid agent alongside retrieval results of an approved knowledge store which contained controlled product names, terms in therapeutic area, customer identifiers, history of prior engagement, the category of allowed follow-up and information related to record ownership. Instead of having the final record being based on a single model response, the agent suggested candidate links and field values with rule-based checks ensuring compatibility between identifiers, mandatory field presence, unsupported product claims, uncertain customer references and phrases to require escalation.

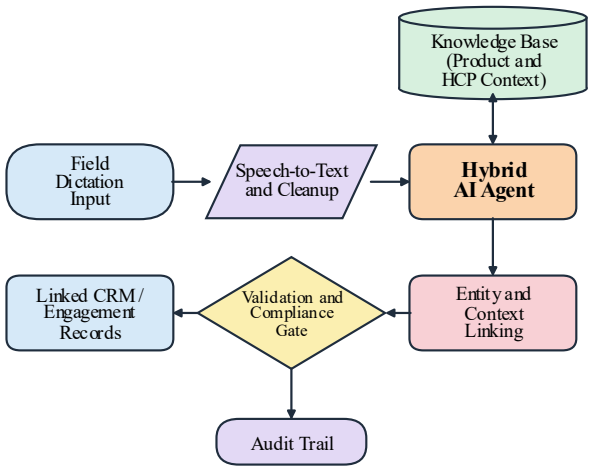

Figure 1. AI agent architecture, dictation to linked record reporting Hybrid AI.

The flow of components of the proposed hybrid AI agent that will be used to engage in reporting in the pharmaceutical fields is shown in Figure 1. It starts with the input of field dictations that are turned into a cleaned transcript with uncertain wording to be reviewed later. The hybrid AI agent then decodes the text and looks into a controlled knowledge base of accepted product terms and the context of the healthcare professional. Entity and context linking ties extracted mentions to enterprise identifiers rather than allowing it to be left as text. A special validation and compliance gate verifies the completeness, contextual support and aptness of the proposed record to be automatically advanced. Satisfying records are sent to the connected CRM or

engagement repository whereas the audit path maintains records of the made decisions in processing. The design is purposely designed to draw a line between interpretation and validation. This division will not allow a fluent language output to be an enterprise record without validation. To make operational responsibility visible fewer components are employed: each phase has a specific role, each arrow denotes a responsible transfer and the central agent is the one who coordinates and not substitutes the controlled systems around it. The architecture is thus fast but with the ability to maintain reviewability, traceability of records and governance.

Equations (1) to (6) formalize how the agent combines stage-level signals and decides when to route a record to human review. Equation (1) defines the hybrid confidence C_{hyb} of a record as a weighted sum of the transcription confidence C_{ASR}, the language-processing confidence C_{NLP}, the entity-link confidence C_{link} and the compliance score C_{comp}, with non-negative weights α, β, γ and δ that sum to 1. Equation (2) scores a candidate enterprise record r for an extracted mention e as a weighted combination of semantic similarity, similarity to the retrieved customer and product context, and lexical similarity, with weights λ_1, λ_2 and λ_3; the highest-scoring candidate is proposed as the link. Equation (3) gives dictation-to-linked-record reliability R_{DLR} from the number of correctly linked records $N_{correct}$, the records validated after review $N_{validated}$, the records escalated to review N_{review} and the total number of records N_{total}, with weights w_1 and w_2. Equation (4) routes a record to human review (H(x) = 1) when its hybrid confidence is below the threshold τ_e or its compliance score S_{comp} is below τ_s. Equation (5) defines linked-record completeness as the weighted share of mandatory fields j that are filled ($I_j = 1$), and Equation (6) combines transcript, linking and compliance accuracy and reviewer acceptance into an overall performance score, less a penalty E_{review} for review effort.

4. DATA DESCRIPTION

The test dataset included 171 cases of de-identified field-engagement that were set up to reflect the reporting circumstances experienced following pharmaceutical customer experiences. Individual instances were an aggregate of dictation text and anticipated product references, customer or organization links, theme-based engagement, post follow-up actions and review labels of compliance sensitive language. The cases differed in dictation clarity, the use of abbreviations, incomplete naming, multiple subjects as well as the reliance on the previous contextual information. The data were broken up into sequential evaluation sets such that the system could be observed in varying phrasing as opposed to a homogenous set. Record linking with controlled identifiers and semantic context based on biomedical concept-relatedness techniques; semantic relationship by embedding between medical concepts have been demonstrated to be relevant, as demonstrated by Mao et al. [11]. The present study evaluation is based on the 171 records which are not a public benchmark. All identifiable data were deleted or substituted and no live promotional decision, prescribing recommendation or action at the patient level were generated out of the records.

5. RESULTS

The analysis revealed that the hybrid AI agent provided a more reliable and transparent way between dictation and connected reporting as compared to the traditional sequential workflow. The highest operation gain that was realized was during the process of transitioning the cleaned transcript to enterprise record since the agent would integrate language interpretation and approved contextual retrieval prior to the suggestion of field values. This minimized the inclination to perceive any of the identified phrases as a text fragment. Agreement on product mentions was higher and reference to stakeholders were better resolved to the surrounding context and follow-up actions were modeled in a form that could be understood by the reviewers without necessarily re-creating the original dictation. Uncertainty was also dealt with in the workflow in a more transparent manner. In case a dictated phrase did not have sufficient backing to form a reliable link, the system stored the source phrase and displayed the unresolved item to a confirmation as opposed to creating a particular match. Such conduct enhanced trust by reviewers since ambiguity was evident on the record instead of being concealed using the use of smooth language. Hybrid agent confidence scoring mathematical modelling can be expressed as:

$$C_{hyb} = \alpha C_{ASR} + \beta C_{NLP} + \gamma C_{link} + \delta C_{comp}, \qquad \alpha + \beta + \gamma + \delta = 1 \tag{1}$$

Table 1. Stage-level reporting quality and record completion.

En-gage-men t Stage	Base-line Quality	Hybrid Agent Quality	Human Review Quality	Record Com-ple-ti on
Dictation Capture	78	94	91	93
Transcript Cleanup	74	92	89	91
Entity Linking	69	90	87	89
Compliance Check	81	96	95	95

The main quality and record completion summaries of quality at the stage level are in Table 1. Dictation capture has the lowest quality with a baseline of 78, whereas the hybrid agent has the highest quality of 94, human review 91 and record completion 93. Cleaning up of the transcript starts at 74 during the baseline up to 92 using the hybrid agent; the reviewer score in the case is 89 and record completion is 91. The lowest area is entity linking with a score of 69, but the hybrid design increases the score to 90, human review is 87 and completion is 89. The highest values are recorded in compliance checking with 81 being the baseline, 96 the hybrid agent, 95 human check and 95 completion. The stage pattern indicates that the agent is the most contributive at the points where the simple transcription forms are not enough. Entity linking is greatly benefited as the agent consolidates the story with accepted context then allocates enterprise identifiers. Enforcement is also of good quality since the validation gate verifies completeness, and sensitive language to ensure that the record is done. The table hence justifies the conclusion that the hybrid workflow enhances the quality of interpretation, as well as practical preparedness of the final record of engagement. Pharmaceutical entity linking score can be framed as:

$$S(e,r) = \lambda_1 Sim_{sem}(e,r) + \lambda_2 Sim_{ctx}(e,r) + \lambda_3 Sim_{lex}(e,r) \tag{2}$$

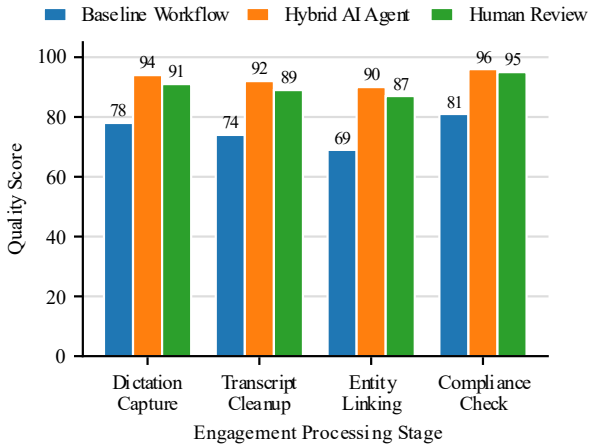

Figure 2. Similarity in the quality of processing at the stages of engagement.

Figure 2 compares the quality scores of the initial workflow (baseline), the hybrid agent (AI) and the human review in four processing stages. In dictation capture, the baseline score is 78, the hybrid agent scores 94 whereas the human review scores 91, indicating that the automated design is much higher than the sequential process but is similar to output prepared by the reviewer. The transcript cleanup in the baseline is 74 which is increased to 92 using the hybrid agent as compared to 89 when the human viewer is used. The greatest difference is in entity linking with the baseline registering 69, the hybrid agent 90 and human review 87. This trend shows that contextual retrieval and controlled linking has a significant contribution in the event that the spoken references must be resolved to enterprise records. Compliance checking also is very strong: the baseline is 81, the hybrid agent 96 and the human review 95. Throughout the four stages, the hybrid method is always above the baseline and it is well conforming to the human review. This chart thus indicates that the architecture not only enhances the processing associated with transcription, but also the down-stream processing that is involved in deciding whether a field note is to be a useful linked record or not. The most powerful gains are made when there is a need for context, identity resolution and governed validation. Dictation-to-linked-record reliability can be mathematically represented as:

$$R_{DLR} = \frac{N_{correct} + w_1 N_{validated}}{N_{total} + w_2 N_{review}} \quad (3)$$

Table 2. Batch-level evaluation of linked reporting performance.

Eval-u-a-tion Batch	Tran-scri pt Quality	Link Com-plete-ness	Com-pli-a nce Read-i-nes s	Reviewer Ac-cept-a nce
Batch A	88	82	86	84
Batch B	90	85	88	87
Batch C	91	88	91	90
Batch D	93	90	93	92

Table 2 shows the results of four assessment batches, which are based on the quality of the transcripts, link completeness, compliance readiness, and acceptance of the reviewers. The transcript quality and link completeness of Batch A are 88 and 82, respectively, and its compliance readiness and reviewer acceptance are 86 and 84. The values in Batch B increase to 90, 85, 88 and 87. The transcript quality, link completeness, compliance readiness and reviewer acceptance of Batch C are 91, 88, 91, and 90 respectively. Record 93, 90, 93 and 92 appear in the same measures in batch D. The consistent flow between batches shows that there is no trade off in the quality of workflow of one type of records with another. The acceptance score increase of reviewers between the first and fourth batches is eight score points and that of link completeness increases by eight points. There is an increase of five points to the quality of transcripts and seven points to compliance readiness. The quality of transcripts and compliance preparedness are the same at Batch D, at 93, while reviewer acceptance is 92. The values show that, the more consistent the orchestration, the greater is the acceptance of the linked output by the reviewers without necessarily working on the original dictation to reconstruct it. Human-in-the-loop escalation decision can be formulated as:

$$H(x) = \{1, C_{hyb}(x) < \tau_c \vee S_{comp}(x) < \tau_s, 0, C_{hyb}(x) \geq \tau_c \wedge S_{comp}(x) \geq \tau_s \quad (4)$$

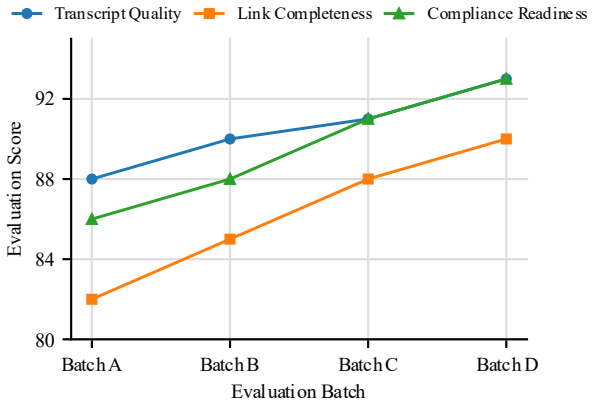

Figure 3. Evaluation measures in terms of four consecutive batches.

Figure 3 follows three evaluation measures in terms of four consecutive batches. In Batch A, the quality of transcripts is 88 and then increases to 90, 91 and 93 in Batch D. The link completeness is 82 and 85, 88 and 90. The compliance preparedness starts with 86, and goes to 88, 91 and 93. The sequence is significant since all the three lines are enhanced without subsequent deterioration, indicating that the work process does not deteriorate as the phrasing of engagement and context of the records are different. The biggest overall increase across the four batches is in link completeness, which improved by eight score points between the first and the final batch. The quality of transcripts is enhanced by five points, and the readiness to comply is enhanced by seven points. The decrease in the distance between transcription and linking implies that the system is more and more efficient in converting the identified language into interrelated enterprise domains than it is just generating a readable text. By Batch D, the quality of transcripts and compliance preparedness are both at 93 and the link completeness is at 90. These values endorse the perception that controlled validation and contextual linking to develop as the agent employs a similar orchestration process in recurrent evaluations. Linked record completeness mathematical modelling can be defined as:

$$L_{comp} = \frac{\sum_{j=1}^{m} w_j I_j}{\sum_{j=1}^{m} w_j}, \qquad I_j \in \{0,1\} \quad (5)$$

Overall Hybrid AI agent performance can be quantified as:

$$P_{agent} = \eta_1 A_{trans} + \eta_2 A_{link} + \eta_3 A_{comp} + \eta_4 A_{accept} - \eta_5 E_{review} \quad (6)$$

The same hybrid structure was advantageous in the preparation of the transcripts. Text obtained by speech could include clipped phrases, non-standard abbreviations as well as compacted references that could be easily comprehended by the speaker but not by a downstream system. Formatting was normalized by the cleanup stage, and did not remove evidence to interpret any ambiguous wording. The agent then made use of the adjacent statements and approved context to find out whether a phrase amounted to a product, an organization, an engagement subject, a request or an operational observation. The resulting records were simpler to check as they were organized into the narrative structure of the encounter in their last fields, as reviewed by the reviewers. Over-generation was capped by the design as well. The agent has never merely added unsupported claims to fill a record, or change a tentative remark into a fact that is definitely supported, where the dictation was not clearly supported.

Entity linking was found to be a very significant distinguishing feature. When there are multiple candidates with similar names, a baseline pipeline is capable of identifying the terms but still will be unable to tie it to the appropriate enterprise object. The hybrid agent took into consideration the perceived customer context, product scopes, the state of previous engagement and text to refine the candidate set and then offered a link. This generated records that are more internally coherent and less manual correction is needed once the records have been processed initially. The same technique enhanced the handling of the compound engagements which involved multiple products or a number of follow-up themes. Instead of reducing the narrative to one note of the genre, the system retained different yet interconnected elements, and retained the relationship to the source dictation. The readiness in compliance was also enhanced since the reporting process did not consider validation as optional final check but a gateway that was mandatory. Before write-back preparation, sensitive phrases, unsubstantiated assertions, ambiguous references and incomplete fields required were surfaced. The gate did not seek to have a business or medical judgment on behalf of the reviewer. Rather, it found circumstances that needed human intervention and did not allow the record to be given out as complete when there was still a lot of uncertainty of importance. This assisted in isolating automation of routine documentation as opposed to those decisions that need responsible review.

There were no differences in performance between the assessment batches, and this means that the system was not relying on a specific phrasing pattern. Subsequent batches included variations in the wording and mixtures of the engagement themes, but the workflow remained the same to generate structured outputs that could be interpreted efficiently by the reviewers. The quality of language was not the only aspect that was improved; the entire record package completeness and consistency were also improved. The acceptance by the reviewers also increased with the ease of processing sequence to understand. The practical benefit of the system was thus through the coordinated control in terms of transcription, interpretation, contextual retrieval, linking, validation and auditability. The paper advocates the adoption of a hybrid agent as a controlled reporting assistant which can fast-track the documentation with ambiguous and sensitive decisions remaining under human decision-making.

6. DISCUSSION

The results show that the power of the hybrid AI agent is in the integration of language understanding and enterprise control instead of an either/or approach to replacing one transcription engine with another. This is clearly shown in Figure 2. The lowest score is for the baseline (69) and the highest score is for the hybrid agent (90) and human review (87) for entity linking. The twenty-one-point increase above the baseline is operationally significant, as it will determine whether a spoken reference will be a ‘usable’ product, customer, organization or a follow-up record. It can also be noticed that in that stage the hybrid score is greater than the reviewers score, which means that it will be able to retrieve candidates consistently when the contextual enterprise information is available. Dictation capture is similar in that the scores are 78 in baseline and 94 after using the agent, and 91 prepped by the reviewer. The 16-point increase shows that this architecture has an advantage of keeping uncertainty and context from the start and not discarding phrases which are hard to clean up.

This is confirmed by Table 1. The record completion for the same stage is 91, up from a baseline of 74 (a 17-point improvement) and transcript cleanup increases from 74 to 92. The results of entity linking are as follows: Baseline quality score is 69, hybrid agent 90, human review 87 and record completion 89. These values indicate that the system is not even only doing a clean-up of the transcript, but that it is also changing the narrative content into fields, which are coherent at the transcript level. Another good piece of compliance checking is gained. The baseline score is 81, the hybrid score is 96 and the human-review score is 95 and record completion is 95. This 15 point difference between baseline and hybrid approach is due to the presence of a validation gate, which runs prior to a write-back. Higher completion does not rely on obscuring uncertainty since the gate can block progress when the necessary information is lacking or there is a delicate word that has not been addressed.

Information on stability in repeated evaluations is shown in Figure 3 and Table 2. The line chart shows that the overall quality of transcripts improves from 88 in Batch A to 93 in Batch D, the completeness of links increases from 82 to 90 and the readiness of compliance increases from 86 to 93. The 8-point increase in the completeness of the links is all the more significant given that linkage is a context-specific task in field reporting. The curve does not demonstrate a late drop, indicating that there is no benefit to the wider wording variation, so it is not cancelling the benefit of the contextual retrieval and controlled matching. The compliance readiness score increases by seven points from the first to the last reading, and in the last reading it equals the transcript quality score. The importance of this relation is that the subsequent performance of the system is not only on the readability of the output but also the quality of governance increases.

Table 2 gives an additional overview from the perspective of the reviewers. The reported reviewer-acceptance score increases from Batch A (84) to Batch B (87), to Batch C (90) and finally to Batch D (92), while the link completeness score moves from 82 to 85 to 88 and 90, and the transcript quality increases from 88 to 90 to 91 to 93. Compliance readiness levels are: 86, 88, 91 and 93. The 8-point increase in reviewer acceptance (from batch 1 to batch 4) has a similar trend to the 8-point increase in link completeness. This implies reviewers are not just being satisfied by the prose's flow, they're being satisfied by the correct relationships in the record. Transparent processing of evidence keeps human confidence very near to the highest measures of system confidence, with transcript quality and readiness for compliance both at 93 for Batch D, while reviewer acceptance is at 92.

The results also help to shed light on the role of human supervision. The hybrid agent does well on routine processing

but the architecture doesn't assume that it can be automatic for all phrases. The audit design includes the audit version with the cleaned transcript, the context that was used to link, the record fields proposed and the validation status. This enables the reviewer to appreciate the rationale behind a candidate link being made, and where there are areas of uncertainty. In the case of the pharmaceutical industry, transparency will be even more crucial as one dictated sentence may contain scientific information, commercial information, information on access to the medication, and a follow-up commitment. That sentence wouldn't be a typical text-generation problem and it would result in an unnecessary governance risk.

This assessment thus suggests a limited-reporting model in which field reporting is regulated. The language component supports free-form dictation, the retrieval component provides acceptable context, the entity linking component supports phrases in free-form text that refer to known entities contained in a controlled record, the rules component enforces known rules, and the human review component allows human review in case evidence is incomplete. Not only does this division of responsibility yield record quality that is better than simple sequential workflow, but it gets closer to, and/or equal to, reviewers-prepared record quality in several stages. Most convincing is the architecture that is not deceptive in its treatment of uncertainty and converts the potentially vague spoken information to traceable linked data. Combining all those factors of speed, linkage, validation and auditability is the study's key practical contribution.

7. CONCLUSION

This research proposes and assesses a hybrid AI agent system which transforms the dictation in the pharmaceutical field into linked and reviewable engagement records. The results demonstrate that the architecture works best in the form of coordinated stages of transcription, contextual retrieval, entity linking, validation and audit generation. Figure 2 finally shows that the hybrid workflow outpaces the baseline values of 78, 74, 69 and 81 in dictation capture, transcript cleanup, entity linking and compliance checking, respectively, with a score of 94, 92, 90 and 96. Table 1 also demonstrates record completion levels of 93%, 91%, 89% and 95% for each of these four stages. There is no loss of performance as the number of re-batches increases. As seen in Figure 3, the Transcript Quality has improved from 88 to 93, the Link Complete has improved from 82 to 90 and the Compliance Ready has improved from 86 to 93. Table 2 also shows that reviewer acceptance has risen from Batch A to Batch D (84 to 92). The patterns illustrate that the strength of the agent is to allow for a certain level of orchestration of input, not for free-form text generation. The design that is proposed maintains uncertain/sensitive cases visible, maintains record fields in the context of controlled cases and maintains an audit path from the dictation to the preparation of the final write-back. The hybrid model – in which transformations are automated but reviewers still approve or reject ambiguous and compliance-sensitive cases – can be a viable solution for a pharmaceutical organization looking for a quicker field documentation process without compromising accountability.

8. REFERENCES

[1] L. Zhou et al., “Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists,” JAMA Network Open, vol. 1, no. 3, art. e180530, 2018. https://doi.org/10.1001/jamanetworkopen.2018.0530

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020. https://doi.org/10.1093/bioinformatics/btz682

[3] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, “Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” npj Digital Medicine, vol. 4, no. 1, art. 86, 2021. https://doi.org/10.1038/s41746-021-00455-y

[4] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, “AMMU: a survey of transformer-based biomedical pretrained language models,” Journal of Biomedical Informatics, vol. 126, art. 103982, 2022. https://doi.org/10.1016/j.jbi.2021.103982

[5] M. Wornow et al., “The shaky foundations of large language models and foundation models for electronic health records,” npj Digital Medicine, vol. 6, no. 1, art. 135, 2023. https://doi.org/10.1038/s41746-023-00879-8

[6] C. Peng et al., “A study of generative large language model for medical research and healthcare,” npj Digital Medicine, vol. 6, no. 1, art. 210, 2023. https://doi.org/10.1038/s41746-023-00958-w

[7] B. Meskó and E. J. Topol, “The imperative for regulatory oversight of large language models (or generative AI) in healthcare,” npj Digital Medicine, vol. 6, no. 1, art. 120, 2023. https://doi.org/10.1038/s41746-023-00873-0

[8] N. Mehandru, B. Y. Miao, E. Rodriguez Almaraz, M. Sushil, A. J. Butte, and A. Alaa, “Evaluating large language models as agents in the clinic,” npj Digital Medicine, vol. 7, no. 1, art. 84, 2024. https://doi.org/10.1038/s41746-024-01083-y

[9] J. Qiu et al., “LLM-based agentic systems in medicine and healthcare,” Nature Machine Intelligence, vol. 6, no. 12, pp. 1418–1420, 2024. https://doi.org/10.1038/s42256-024-00944-1

[10] S. Garda, L. Weber-Genzel, R. Martin, and U. Leser, “BELB: a biomedical entity linking benchmark,” Bioinformatics, vol. 39, no. 11, art. btad698, 2023. https://doi.org/10.1093/bioinformatics/btad698

[11] Y. Mao and K. W. Fung, “Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts,” Journal of the American Medical Informatics Association, vol. 27, no. 10, pp. 1538–1546, 2020. https://doi.org/10.1093/jamia/ocaa136

[12] A. Singh, S. Krishnamoorthy, and J. E. Ortega, “NeighBERT: medical entity linking using relation-induced dense retrieval,” Journal of Healthcare Informatics Research, vol. 8, pp. 353–369, 2024. https://doi.org/10.1007/s41666-023-00136-3

[13] F. L. Barra et al., “From prompt to platform: an agentic AI workflow for healthcare simulation scenario design,” Advances in Simulation, vol. 10, no. 1, art. 29, 2025. https://doi.org/10.1186/s41077-025-00357-z