Manage My Spend
System card

Prepared for Manage My Spend

Report Analyst: Nicole Lincoln

Report Supervisors: Erin Cairney, Nick Whitehouse

January 2, 2025

1. Introduction

This System Card describes the use and performance of AI in Manage My Spend.

Manage My Spend leverages a proprietary Retrieval Augmented Generative AI system integrated with OpenAI foundation models to automate invoice dispute processing for in-house legal teams working with outside counsel. By emailing or uploading invoices to the platform, users can automatically reconcile them against their preset spend guidelines. The platform offers an Artificial Intelligence chat feature for querying spend data, with visual insights and an intuitive interface for managing budgets, guidelines, invoices and a support assistant chat for user inquiries.

The purpose of Manage My Spend is to streamline invoice processing and provide insights into legal spend data.

AI Platform Risk Level: Not classified as high-risk.

Feedback Mechanism: Users can report issues or suggest improvements through contact@managemyspend.com.

2. Performance benchmarking

To evaluate the performance of Manage My Spend, we conducted rigorous human benchmarking using a diverse group of seasoned paralegals and attorneys. These experts reviewed 49 "ground truth" invoices according to Manage My Spends default billing guidelines, providing an accurate baseline of human performance in both accuracy and time. Following this, we tested two versions of Manage My Spend: a Preview version and a Full version. The results were analyzed to compare the system's accuracy and efficiency against human benchmarks.

The benchmarking revealed that both Manage My Spend models vastly outperformed human reviewers in terms of speed, completing invoice reviews over 10 times faster. Accuracy varied across versions, with the Preview version falling short of human standards by 15%, largely due to a tendency to over dispute. However, the Full version demonstrated accuracy equivalent to human reviewers, achieving parity with the benchmark.

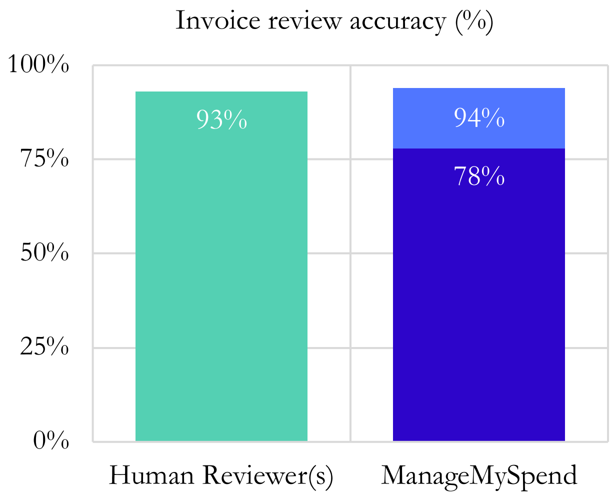

Chart 1: Review accuracy benchmarking

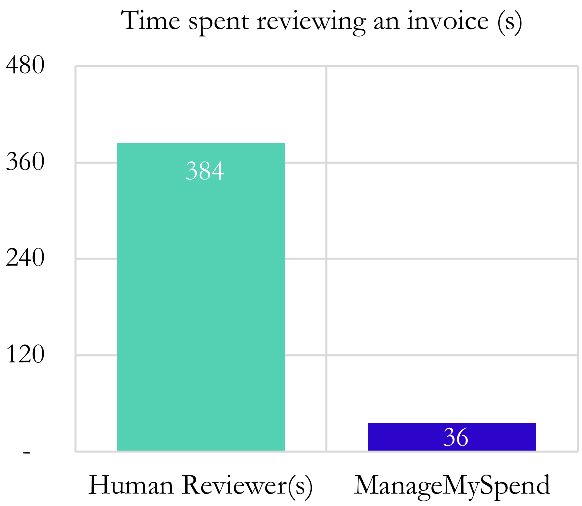

Chart 2: Review time benchmarking

These benchmarking results highlight the transformative impact of Manage My Spend on invoice review processes. The Full version achieves human-level accuracy while eliminating the need for human intervention in invoice reviews, completing the task 10 times faster than traditional manual methods. This translates into substantial operational efficiencies, allowing organizations to redirect resources away from time-intensive, repetitive tasks toward higher-value activities.

The Preview version's tendency to over-dispute invoices reflects the complexity of aligning automated decision-making with nuanced billing scenarios. While the Full version of Manage My Spend successfully addressed this issue, it is important to acknowledge that benchmarking was conducted on a relatively small dataset. This introduces a risk that over-disputing could still occur in broader, real-world applications if the model corrections fail to generalize adequately.

Organizations leveraging Manage My Spend can achieve substantial cost and time savings. By automating invoice reviews, organizations can enforce best-practice billing guidelines, saving an average of 5-8% on legal spend. This not only ensures compliance and accelerates invoice approvals but also eliminates manual bottlenecks. With its human level accuracy, Manage My Spends AI Agent capability enables teams to scale efficiently while maintaining exceptional control and precision.

3. Model overview

The system utilizes generative AI from OpenAI to automate spend management processes. The first model use focuses on invoice approval and dispute resolution, automating the evaluation of invoices by applying uploaded spend guidelines. This model is classified as Generally Available[1] (GA) and carries a limited risk of providing inaccurate answers. The second model use provides a chat-based conversational interface that allows users to query spend data, retrieve insights—presented both textually and visually through charts—and access platform support features. Similarly, this model is also Generally Available[1] (GA) and shares the limited risk of potential inaccuracies in generated responses.

Table: 1: Model Use 1 (Invoice review)

Model Use	Invoice review
Description	The AI automates invoice processing by approving or disputing invoices based on uploaded spend guidelines.
Type	Generative AI (OpenAI GPT-4o)
Status	GA - Generally Available
Risk(s)	Limited - inaccurate answers

Table 2: Model Use 2 (Chat)

Model Use	Chat
Description	A conversational interface to query spend data, retrieve insights (both textual and visual through graphs and bars), and access support for platform features.
Type	Generative AI (OpenAI GPT-4o)
Status	GA - Generally Available
Risk(s)	Limited - inaccurate answers

Detailed Model Use Cards

The following detailed use cards provide a deeper understanding of how Manage My Spend leverages OpenAI's GPT-4o. As MMS leverages OpenAI GPT-4o, industry best practice safety, alignment, and transparency are built in. Please refer to the GPT-4o system card here.

Table 3: Detailed model card – Invoice review

Model Use	**Invoice Approval/Dispute**
Description / intended use	**Input** Invoices received in PDF or .docx format via email. AI extracts relevant details and cross-checks this data against the uploaded spend guidelines and budget limits. Each invoice is classified as either approved or disputed based on adherence to these parameters. **Output** Approved invoices: Forwarded automatically to the accounts department if a pre-determined approval threshold is met. Disputed invoices: Generates an automate dispute email to the sender and outlines the reasoning. If the invoice exceeds any present budget thresholder, it triggers a proactive email notification to the user.
Status	GA – Generally Available
Risk(s)	Inaccurate disputes and dispute reasoning
Type	Generative AI
Model/architecture used	gpt-4o-2024-08-06
Performance metrics	**Dispute accuracy**: Preview 78% / Full 94% (indicating how often the AI correctly identified whether an invoice should be approved or disputed). **Reasoning accuracy**: 51% / Full 83% (reflecting the AI's ability to provide accurate reasoning for disputing an invoice).
Evaluation approach	Binary scoring against determined success criteria for each invoice. For disputed invoices, reasoning was classified into true positive, false positive or false negative against a validation set.
Validation set	49 legal invoices of various length, format and type.

Data source used to fine-tune, test and evaluate.	Synthetic invoices based on historical invoice data specifically created to reflect realistic billing scenarios while maintaining privacy and compliance. Features include: • Varying lengths to reflect diverse billing formats • Anonymized timekeeper details for privacy • Fictional law firms, matter numbers and matter descriptions to avoid referencing real entities. • Fabricated billing narratives to resemble typical invoice descriptions.
Supported input	Legal invoices in .pdf format
Expected output	JSON
Limitations and comments	**Data Quality Dependency:** Performance can degrade with incomplete invoices that are missing expected information. **Language Support:** Primarily validated on English invoices in USD; limited support for multilingual documents and invoices in other currencies. **Calculations**: Challenges in accurately handling deadlines, fiscal month determinations and daily billing calculations, which could lead to errors

Table 4: Detailed model card – Chat

Model Use	**Chat**
Description / intended use	**Input** User queries related to spend data (e.g. What's our total spend this month?", "Show me invoices that were disputed"). Requests for visual data insights (e.g. graphs, charts or other artifacts). Support queries related to creating spend guidelines or other platform features. **Output** Spend Data Queries: Relevant spend data displayed in text form. Visual data insights provided as requested, such as graphs and charts. Support functionality: Contextual responses to user queries, acting as a personal assistant for navigating the platform and its features. Guidance on creating spend guidelines or using other functionalities.
Status	GA – Generally Available
Risk(s)	Limited – Inaccurate answers
Type	Generative AI

Model/architecture used	gpt-4o-2024-08-06
Performance metrics	79% accuracy
Evaluation approach	Similarity and accuracy scoring against detailed success criteria for each established AI chat question and conversation
Validation set	25 AI conversations of varying length across 49 invoices of various length, format and type.
Data source	Anonymized and synthetic invoices specifically created to reflect realistic billing scenarios while maintaining privacy and compliance. Features include: Varying lengths to reflect diverse billing formats Anonymized timekeeper details for privacy Fictional law firms, matter numbers and matter descriptions to avoid referencing real entities. Fabricated billing narratives to resemble typical invoice descriptions.
Supported input	Plain Text
Expected output	JSON
Limitations and comments	**Data Quality Dependency:** Performance can degrade with incomplete invoices that are missing expected information. **Language Support:** Primarily validated on English invoices in USD; limited support for multilingual documents and invoices in other currencies. **Calculations**: Challenges in accurately performing mathematical calculations. **User Input**: Output and ability to answer questions is dependent upon user input containing field information on where data can be found.

4. Risk assessment

The risk assessment process for the system is rigorous and grounded in industry best practices, incorporating red teaming exercises and repeated testing to thoroughly evaluate the likelihood and impact of potential issues. This approach ensures a comprehensive understanding of the system's vulnerabilities and provides a framework for mitigating risks effectively.

The assessment examines four key areas—accuracy, bias/discrimination, alignment, and security—to identify both strengths and areas requiring attention. Each category is assessed based on its potential for disruption and the measures in place to address these risks, ensuring that the system maintains high reliability, ethical standards, and robust protection for sensitive data.

Table 5: Risk Assessment Matrix

	Score	Negligible	Minor	Moderate	Significant	Severe
Accuracy	12			X		
Bias / Discrimination	3	X				
Alignment	8		X			
Security	6			X		

1. **Accuracy:** The system demonstrates a **Moderate** risk related to accuracy, particularly in its ability to extract line-item information and generate appropriate queries. These errors could lead to incorrect categorizations or decisions, such as mistakenly approving or disputing invoices. This risk is inherent in the reliance on automated processes to interpret and act upon complex or ambiguous data. This risk is largely mitigated by user oversight integrated into critical decision points, ensuring that any inaccuracies can be easily identified and rectified.

2. **Bias and discrimination**: The risk of bias or discrimination is assessed as **Negligible** due to the system's focus on invoice content, which inherently reduces its exposure to personal demographic data. However, as with any AI system operating in a specific domain, there is a possibility of reflecting existing biases within the legal field. For example, decisions influenced by patterns in legal spend might inadvertently perpetuate systemic inequities. This bias is significantly mitigated through the systems strict adherence to billing guidelines, which, users have complete control over.

3. **Alignment**: is a key strength of the system, particularly in its specialized design for legal spend use cases. It demonstrates a high level of alignment when responding to domain-specific queries and executing tasks directly tied to its intended purpose. However, we identified a **Minor** risk as challenges may arise when users present queries that are unclear, overly general, or outside the scope of legal spend management. In such cases, the system might produce outputs that fail to meet user expectations. This misalignment risk is

mitigated by clear user documentation, intuitive interface design, and ongoing training to improve the system's contextual understanding and adaptability.

4. **Security**: The system is assessed as having a low risk of security breaches due to robust measures, including encryption, access controls, and adherence to industry standards for data protection. However, the nature of the data handled—often involving attorney-client privileged information—elevates the potential impact of any breach to a **moderate** level. A breach could compromise sensitive legal and financial information, leading to reputational damage and compliance issues. These risks are mitigated through the use of strict data governance policies to safeguard confidential. The system undergoes regular security testing and uses highly secure cloud infrastructure that is independently verified.

5. Conclusion

The design of Manage My Spend is underpinned by a rigorous use of retrieval-augmented generative AI, validated against synthetic, anonymized data that faithfully replicates real-world billing scenarios. The structured lifecycle framework, categorizing the model's maturity from Experiment to General Availability, ensures the solution has progressed through clearly defined stages of development with appropriate risk management at each phase. This methodical approach underscores the platform's reliability and operational maturity.

The risk assessment framework is comprehensive and pragmatic, focusing on accuracy, alignment, bias, and security. While risks related to data dependency and multilingual support are acknowledged, the system's reliance on billing guidelines and user oversight for critical decisions provides effective mitigations. The clear identification of limitations, such as challenges with incomplete data and currency handling, reinforces the transparency of the evaluation process and supports trust in the system's deployment readiness.

Overall, the methodologies applied in the performance evaluation, model validation, and risk assessment demonstrate a high degree of rigor. These approaches provide confidence that Manage My Spend is not only well-suited for its intended use but also capable of maintaining its performance and reliability in practical, real-world applications. While continuous monitoring and iterative improvement remain essential, the system is a robust and responsible solution for automating legal invoice reviews.

6. Definition and guidance

The lifecycle of an AI product involves distinct phases, each representing a specific stage of development and maturity. These phases help define the readiness of the AI solution for deployment and use. Understanding and categorizing these phases allows for a structured approach to risk management, ensuring that development is aligned with acceptable standards of performance, security, and alignment.

In this section, we outline the Model Phase Definitions and Model Risk Scoring Definition framework we use when evaluating AI products.

This structured approach ensures that AI solutions are developed, deployed, and monitored responsibly, minimizing risks and maximizing value.

Model phase definition

When developing AI products, we use the following definitions to categorize the maturity of the solution, from Experiment to End of Life. Each stage of development carries a level of risk that we take account of when assessing overall risk.

Code	Status	Description	Risk
EX	Experiment	The AI solution is not completely understood, standards are not defined, and no performance is being measured.	Critical
AA	Alpha	AI solution is proposed, standards are yet to be defined. Performance of the AI shows applicability to the problem, but significant testing, adjustment and validation remain.	Very High
BA	Beta	Standards are defined, AI performance does not consistently meet these standards.	High
LA	Limited Availability	AI meets or is within 5% of acceptable standards but use at scale in the wild is unknown and requires monitoring and adjustment.	Medium
GA	General Availability	Meets or exceeds acceptable standards and use at scale is monitored and well known.	Low
DD	Depreciated	AI is no longer being improved and is actively being phased out from use in products.	Low
US	Unsupported	AI is no longer being maintained or supported.	High

Model risk scoring definition

After identifying the phase definition, we review the product for Accuracy, Bias/Discrimination, Alignment and Security, scoring risk based on likelihood and impact.

		Impact				
		Negligible 1	Minor 2	Moderate 3	Significant 4	Severe 5
Likelihood	Very likely 5	5	10	15	20	25
	Likely 4	4	8	12	16	20
	Possible 3	3	6	9	12	15
	Unlikely 2	2	4	6	8	10
	Very unlikely 1	1	2	3	4	5