Multi-Agent AI Evaluation of Donor Suitability in Liver Transplantation

Joseph A. Katz[1,2], Bima J. Hasjim[2], Ghazal Azafar[2], Mamatha Bhat[2]

[1]Faculty of Physiology and Immunology, University of Toronto; [2]Transplant AI initiative, Ajmera Transplant Centre, University Health Network, University of Toronto, ON, Canada

Background

Liver transplantation is the only curative treatment for end-stage liver disease. Despite organ scarcity, many viable livers are discarded, highlighting the need for improved donor evaluation.

AI Agents are specialized Large Language Model (LLM) systems that can learn, reason, and make functional decisions. These agents can be tailored to complete specific tasks and can collaborate by sharing information and simulating a real-world transplant selection committee, referred to as a Multi-Agent model.

Primary Research Question: How effective is a Multi-Agent model at predicting liver transplant decisions using the donor's clinical profile?

Methods

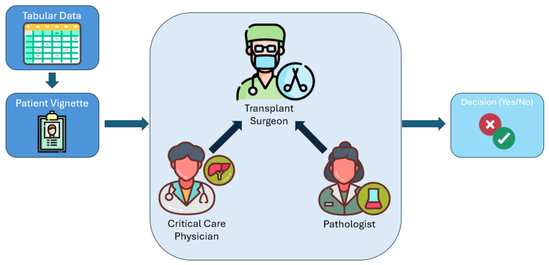


Figure 1. Multi-Agent Committee Workflow

Data Description and Filtering

After merging the SRTR and UNOS Liver Discard dataset, relevant donor clinical features were identified, and the data was cleaned.

Multi-Agent Committee Design

A three-agent AI committee (Transplant Surgeon, Pathologist, Critical Care Physician) was built with a low-temperature vignette-generation agent.

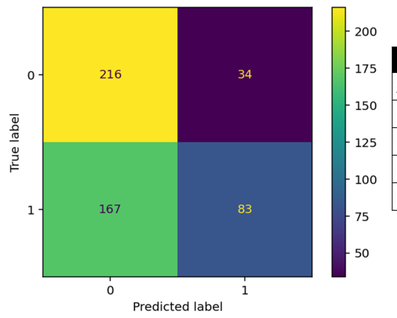
Coding and Prompting

In Python using CrewAI to coordinate agent workflows. Specialized with Prompt engineering and used GPT-4o as the LLM. RAG integration.

Execute Code Using Study Population Data

Results

From a merged dataset of 45,876 patients from the Scientific Registry of Transplant Recipients and the United Network for Organ Sharing, an initial random subsample of 500 patients (250 accepted, 250 rejected) was evaluated using tabular data, without prompting for missing variables. After revising the approach, a second random subsample of 500 patients (250 accepted, 250 rejected) was evaluated using AI-Agent-generated vignette information, with new prompting strategies to handle missing data.



Metric	Value
Accuracy	0.60
Specificity	0.33
Sensitivity	0.864
Precision	0.56
F1 Score	0.68

Figure 2. Multi-Agent Model Performance on 500-Patient Prompted Vignette Non-RAG Subsample

Of the 500 patients of the second subsample, the Multi-Agent model correctly identified 216 accepted (true positives) and 83 rejected (true negatives) donor cases. However, it misclassified 167 rejected cases as acceptable (false positives) and missed 34 viable organs (false negatives),

The model demonstrated high sensitivity with strong identification of viable donor cases, but lower specificity in distinguishing unsuitable ones. The performance metrics reflect this, with an accuracy of 59.8% and an F1 score of 0.682.

Variable	Total (n = 500)	FP (n = 167)	TP (n = 216)	P-value
Age (Years)	44.16 (17.54)	44.17 (17.04)	39.57 (16.37)	<0.001
Aspartate Transaminase	111.17 (206.59)	110.04 (183.21)	82.12 (148.06)	0.002
Biopsy Macrovesicular Fat	15.10 (18.71)	11.34 (11.37)	5.84 (7.53)	<0.001
Biopsy Microvesicular Fat	14.30 (20.71)	11.88 (17.78)	7.64 (15.54)	<0.001
Serum PO2	220.73 (149.79)	208.52 (146.77)	245.83 (154.82)	0.001
FiO2	77.26 (25.95)	73.89 (26.26)	81.13 (24.55)	0.034

Figure 3. Donor Characteristics Stratified by Model Classification (True Positive vs. False Positive)

False positive cases were associated with higher donor age, elevated aspartate transaminase, and increased macrovesicular fat on biopsy. Higher microvesicular fat was also more prevalent in false positives. Additionally, donors in the false positive group had higher PO_2 levels and FiO_2 values.

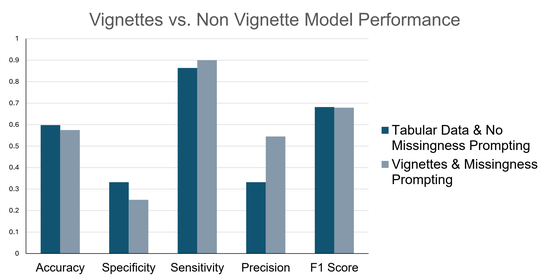
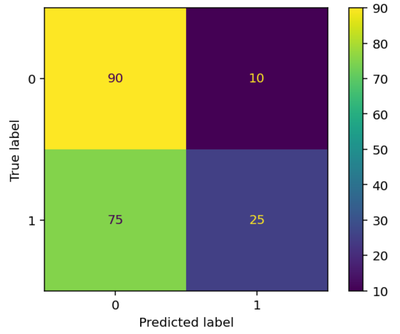


Figure 4. Performance Comparison: Tabular Input Without vs. Vignette-Based Input With Missingness Prompting

Using vignettes with proper missingness prompting resulted in slightly higher sensitivity and precision compared to tabular data without missing data prompting. Accuracy and F1 score remained similar between the two input formats. Specificity was lower when using vignettes.

Next, integrating Retrieval-Augmented Generation (RAG) using North American guidelines from BC Transplant Provincial Health Services and European guidelines from the European Association for the Study of the Liver (EASL), 200 patients (100 accepted v 100 rejected) were run through the model.



Metric	Value
Accuracy	0.58
Specificity	0.25
Sensitivity	0.90
Precision	0.55
F1 Score	0.68

Figure 5. Multi-Agent Model Performance on 200-Patient Prompted Vignette and RAG Subsample

Of the 200 patients of the second subsample, the Multi-Agent model correctly identified 90 accepted (true positives) and 25 rejected (true negatives) donor cases. However, it misclassified 75 rejected cases as acceptable (false positives) and missed 10 viable organs (false negatives),

The model demonstrated higher sensitivity with strong identification of viable donor cases, but lower specificity in distinguishing unsuitable ones. The performance metrics reflect this, with an accuracy of 58% and an F1 score of 0.68.

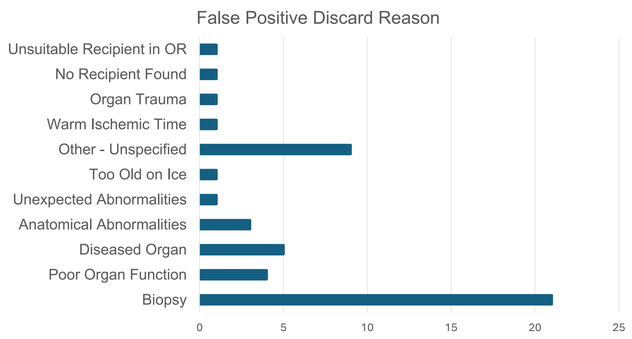


Figure 6. Provided Reason for Discard of RAG-tested False Positive Donors

Of the false positives, the majority come from high microvesicular or macrovesicular fat content as observed under biopsy. Fat contents of 30% and greater had been accepted, with justifications based on RAG-content and possible workarounds clinically. Other reasons for discard are partially covered by the content in the RAG-integrated guidelines, but they could be outside of the scope of the parameters fed into the model for each patient. 62 parameters were selected of 450 based on statistical significance to organ decision status ($\alpha = 0.05$ and< 20% missingness).

Discussion

Agents appeared more permissive in their decision-making, being more approving of livers from older or critically ill donors and those with higher fat content on biopsy.

Despite relatively low specificity, the high sensitivity suggests potential for reducing liver discard by identifying more viable grafts.

There were instances where missing variables led to false positives, highlighting a limitation of the dataset, as certain features may have been excluded if they were not deemed significant for donor approval, leaving agents to make decisions without access to potentially important clinical information.

The model showed positive results in identifying suitable grafts, but further refinement is needed. Enhancing specificity and incorporating more granular clinical reasoning could improve performance.

Moreover, clinical guidelines used did not fully reflect true clinical reasoning and thresholds used for allograft donation decision-making.

Overall, this approach shows promise but requires improvement.

Clinical significance

- Multi-Agent models may be able to identify livers fit for transplant that would otherwise be rejected during standard quality review.

Limitations of results

- Low specificity suggests agents without training material may struggle to appropriately weight key clinical variables, limiting accuracy.

Missingness

- If agents encountered missing variables, they may have lacked critical information needed for accurate decision-making, potentially affecting the reliability of their assessments.

RAG Guideline Sources

- Accuracy of judgements with RAG depend on quality of sources of knowledge integrated into model. Having clinically-relevant descriptive guidelines would increase the accuracy and specificity

Figure 6. Discussion in a nutshell

Next Steps

- Refine North American and European guidelines included in RAG
- Few-shot prompting as a possible inclusion in Agent design.
- Expand Parameters included in decision-making