



LLMs and GenAI Tools to Depict Contributions of Human Systems to Spaceflight Tasks Execution

Crew Health and Performance Probabilistic Risk Assessment Project (CHP-PRA)

Mona Matar, Ph.D.¹

Henry Arthur²

Hunter Rehm, Ph.D.³

2025 HRP IWS - *Galveston, TX*

Jan 31, 2025

Goal



- The Performance Risk Model (PRisM) is a tool that predicts astronauts' performance under different scenarios

Goal



- The Performance Risk Model (PRisM) is a tool that predicts astronauts' performance under different scenarios
- PRisM requires knowing what human systems are needed to complete a task successfully

Goal



- The Performance Risk Model (PRisM) is a tool that predicts astronauts' performance under different scenarios
- PRisM requires knowing what human systems are needed to complete a task successfully
- Use GenAI to automate mapping each Mars task¹ to the Human System Performance Categories (HSPCs)

¹ Stuster, et al., 2018. Human Exploration of Mars: Preliminary Lists of Crew Tasks. Washington, DC: NASA.

Goal



- The Performance Risk Model (PRisM) is a tool that predicts astronauts' performance under different scenarios
- PRisM requires knowing what human systems are needed to complete a task successfully
- Use GenAI to automate mapping each Mars task¹ to the Human System Performance Categories (HSPCs)
- Design an experiment to accept the use of GenAI

¹ Stuster, et al., 2018. Human Exploration of Mars: Preliminary Lists of Crew Tasks. Washington, DC: NASA.

Mars Tasks and HSPCs



MARS
TASK

	Cognitive	Cardiopulmonary	Lower Extremity	Upper Extremity	Fine-motor Control	Sensorimotor Coordination
Communicate to Mission Control	1	0	0	0	1	0

Mars Tasks and HSPCs



MARS
TASK

	Cognitive	Cardiopulmonary	Lower Extremity	Upper Extremity	Fine-motor Control	Sensorimotor Coordination
Communicate to Mission Control	1	0	0	0	1	0
...x926 tasks

Mars Tasks and HSPCs



- Linguistically understand the task and each HSPC

MARS
TASK

	Cognitive	Cardiopulmonary	Lower Extremity	Upper Extremity	Fine-motor Control	Sensorimotor Coordination
Communicate to Mission Control	1	0	0	0	1	0
...x926 tasks

Human System Performance Categories



Definitions:

Cognitive: Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.

Examples:

Operating technical equipment, communicating instructions.

Human System Performance Categories



Definitions:	Examples:
Cognitive: Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.	Operating technical equipment, communicating instructions.
Cardiopulmonary: Tasks requiring strenuous physical exertion above normal daily activity levels.	Conducting an EVA, operating heavy equipment manually.

Human System Performance Categories



Definitions:	Examples:
Cognitive: Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.	Operating technical equipment, communicating instructions.
Cardiopulmonary: Tasks requiring strenuous physical exertion above normal daily activity levels.	Conducting an EVA, operating heavy equipment manually.
Lower Extremity: Tasks requiring strenuous engagement of the legs.	Walking in a pressure suit in gravity, exercising.

Human System Performance Categories



Definitions:	Examples:
Cognitive: Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.	Operating technical equipment, communicating instructions.
Cardiopulmonary: Tasks requiring strenuous physical exertion above normal daily activity levels.	Conducting an EVA, operating heavy equipment manually.
Lower Extremity: Tasks requiring strenuous engagement of the legs.	Walking in a pressure suit in gravity, exercising.
Upper Extremity: Tasks requiring strenuous engagement of the arms or shoulders.	Climbing, handling heavy objects.

Human System Performance Categories



Definitions:	Examples:
Cognitive: Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.	Operating technical equipment, communicating instructions.
Cardiopulmonary: Tasks requiring strenuous physical exertion above normal daily activity levels.	Conducting an EVA, operating heavy equipment manually.
Lower Extremity: Tasks requiring strenuous engagement of the legs.	Walking in a pressure suit in gravity, exercising.
Upper Extremity: Tasks requiring strenuous engagement of the arms or shoulders.	Climbing, handling heavy objects.
Fine-motor Control: Tasks requiring precise and coordinated movements of the hand or fingers.	Pressing buttons, writing.

Human System Performance Categories



Definitions:	Examples:
Cognitive: Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.	Operating technical equipment, communicating instructions.
Cardiopulmonary: Tasks requiring strenuous physical exertion above normal daily activity levels.	Conducting an EVA, operating heavy equipment manually.
Lower Extremity: Tasks requiring strenuous engagement of the legs.	Walking in a pressure suit in gravity, exercising.
Upper Extremity: Tasks requiring strenuous engagement of the arms or shoulders.	Climbing, handling heavy objects.
Fine-motor Control: Tasks requiring precise and coordinated movements of the hand or fingers.	Pressing buttons, writing.
Sensorimotor Coordination: Tasks requiring maintenance of balance or stability through integration of vestibular inputs. This does not apply to tasks in microgravity.	Walking in gravity.

Prompt



How can Claude help you today?

Prompt



How can Claude help you today?

<Role>

Your role is to classify astronaut tasks by performance impact to support mission-level risk evaluation.

Prompt



How can Claude help you today?

<Role>

Your role is to classify astronaut tasks by performance impact to support mission-level risk evaluation.

<Rules>

If the task satisfies the category definition, assign a 1. If the task does not satisfy the category definition, assign a 0.

Prompt



How can Claude help you today?

<Role>

Your role is to classify astronaut tasks by performance impact to support mission-level risk evaluation.

<Rules>

If the task satisfies the category definition, assign a 1. If the task does not satisfy the category definition, assign a 0.

<Information>

A (g) means that the task takes place in gravity and a (mg) means microgravity.

...

Prompt



How can Claude help you today?

<Role>

Your role is to classify astronaut tasks by performance impact to support mission-level risk evaluation.

<Rules>

If the task satisfies the category definition, assign a 1. If the task does not satisfy the category definition, assign a 0.

<Information>

A (g) means that the task takes place in gravity and a (mg) means microgravity.

...

<Categories>

1 - Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.
Examples: operating technical equipment, communicating instructions.

...

Prompt



How can Claude help you today?

<Role>

Your role is to classify astronaut tasks by performance impact to support mission-level risk evaluation.

<Rules>

If the task satisfies the category definition, assign a 1. If the task does not satisfy the category definition, assign a 0.

<Information>

A (g) means that the task takes place in gravity and a (mg) means microgravity.

...

<Categories>

1 - Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking.
Examples: operating technical equipment, communicating instructions.

...

<Output>

Output JSON with the Task and Task Number exactly as entered and a 1 or 0 for each category.


```
{"Task": "", "1": , "2": , "3": , "4": , "5": , "6": , }
```

Example Output



Claude 3.5 Sonnet

```
{"Task Number": "1.3", "Task": "Communicate observations/evaluations to other crew and MCC personnel, verbally using communications system", "Cognitive": 1, "Cardiopulmonary": 0, "Lower Extremity": 0, "Upper Extremity": 0, "Fine-motor Control": 0, "Sensorimotor Coordination": 0}, {"Task Number": "2.3", "Task": "Move (translate) to crew seat, wearing pressure suit, to prepare for launch from Earth.", "Cognitive": 0, "Cardiopulmonary": 0, "Lower Extremity": 1, "Upper Extremity": 0, "Fine-motor Control": 0, "Sensorimotor Coordination": 1}, {"Task Number": "3.2", "Task": "Press edge key/button and F1 panel switch, manually with gloved hand, to activate LCG Pump.", "Cognitive": 0, "Cardiopulmonary": 0, "Lower Extremity": 0, "Upper Extremity": 0, "Fine-motor Control": 1, "Sensorimotor Coordination": 0}, {"Task Number": "4.1", "Task": "Secure seat restraints, wearing pressure suit/gloves, to prepare for launch from Earth.", "Cognitive": 0, "Cardiopulmonary": 0, "Lower Extremity": 0, "Upper Extremity": 0, "Fine-motor Control": 1, "Sensorimotor Coordination": 0}, {"Task Number": "5.1", "Task": "Use Earth Ascent Vehicle (EAV) waste management system for liquid waste (i.e., toilet/bodily function).", "Cognitive": 0, "Cardiopulmonary": 0, "Lower Extremity": 0, "Upper Extremity": 0, "Fine-motor Control": 1, "Sensorimotor Coordination": 0}
```

 Copy  Retry  



Claude can make mistakes. Please double-check responses.

Example Output



Task Number	Task Definition	HSPC Definition	Cognitive	Cardiopulmonary	Lower Extremity	Upper Extremity	Fine-motor Control	Sensorimotor Coordination
			Tasks requiring vigilant attention, working memory, or complex scanning and visual tracking. Examples: operating technical equipment, communicating instructions.	Tasks requiring strenuous physical exertion above normal daily activity levels. Examples: conducting an EVA, operating heavy equipment manually.	Tasks requiring strenuous engagement of the legs. Examples: walking in a pressure suit in gravity, exercising.	Tasks requiring strenuous engagement of the arms or shoulders. Examples: climbing, handling heavy objects.	Tasks requiring precise and coordinated movements of the hand or fingers. Examples: pressing buttons, writing.	Tasks requiring maintenance of balance or stability through integration of vestibular inputs. This does not apply to tasks in microgravity. Examples: walking in gravity. SEE NOTE.
1.3	Communicate observations/evaluations to other crew and MCC personnel, verbally		1	0	0	0	0	0
2.3	Move (translate) to crew seat, wearing pressure suit, to prepare for launch from Earth.		0	0	0	0	0	1
3.2	Press edge key/button and F1 panel switch, manually with gloved hand, to activate LCG Pump.		1	0	0	0	1	0
4.1	Secure seat restraints, wearing pressure suit/gloves, to prepare for launch from Earth.		0	0	0	0	1	0
5.1	Use Earth Ascent Vehicle (EAV) waste management system for liquid waste (i.e., toilet/bodily function).		0	0	0	0	1	0
6.1	Access and open emesis bag located in pressure suit/seat pocket, manually with gloved hand, to prepare for		0	0	0	0	1	0
7.1	Prepare meal for crew consumption, manually, using Earth Ascent Vehicle (EAV) food hydration/heating		0	0	0	0	1	0
8.6	Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if		1	0	0	0	1	0
9.1	Observe crew position displays and activate controls, wearing pressure suit, to prepare for launch from Earth.		1	0	0	0	1	0

Experimental Setup



AI Classification

Experimental Setup



Individual Classification

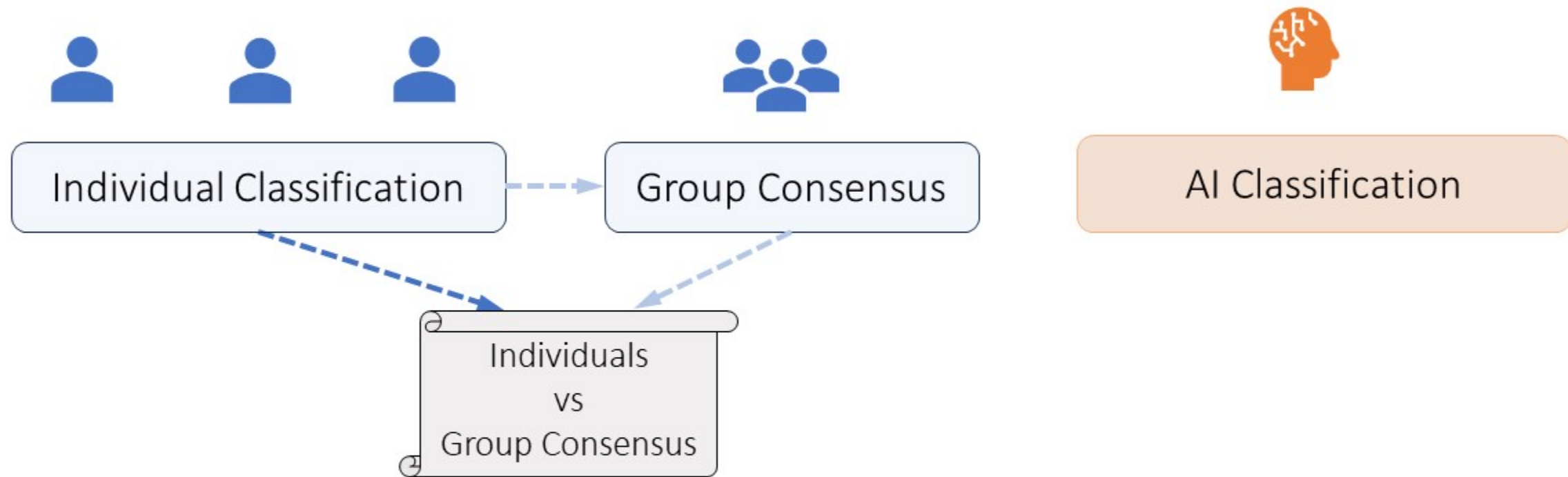


AI Classification

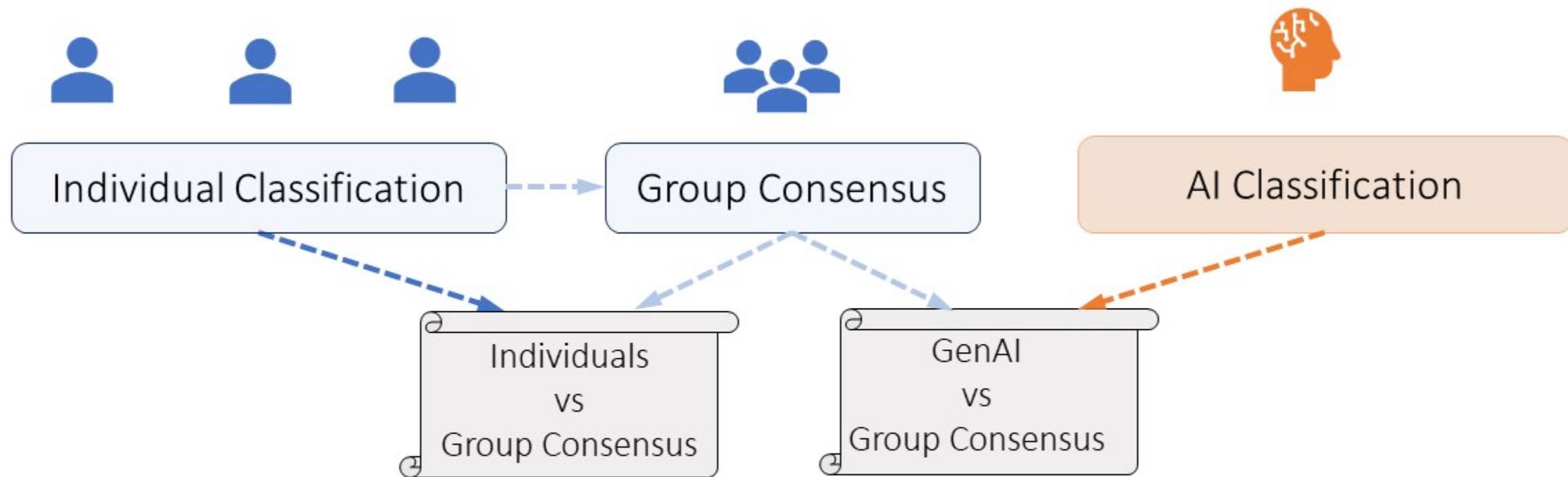
Experimental Setup



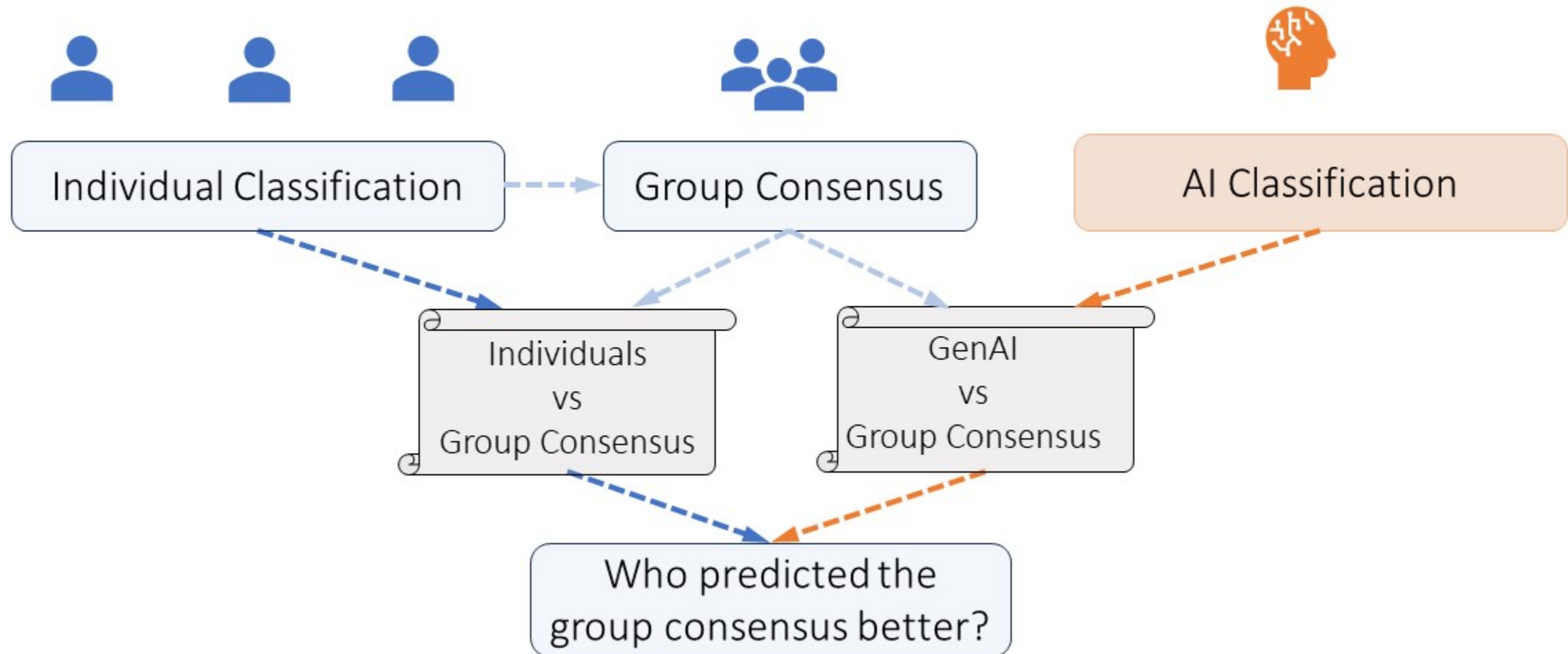
Experimental Setup



Experimental Setup



Experimental Setup



Results



Individuals vs Group Consensus

Category	F1 Score
Cognitive	0.91
Cardiopulmonary	0.74
Lower Extremity	0.71
Upper Extremity	0.76
Fine-motor Control	0.85
Sensorimotor Coordination	0.51

Matthews correlation coefficient 0.72

Results



Individuals vs Group Consensus

Category	F1 Score
Cognitive	0.91
Cardiopulmonary	0.74
Lower Extremity	0.71
Upper Extremity	0.76
Fine-motor Control	0.85
Sensorimotor Coordination	0.51

Matthews correlation coefficient 0.72

GenAI vs Group Consensus

Category	F1 Score	
Cognitive	0.92	(+1)
Cardiopulmonary	0.82	(+8)
Lower Extremity	0.83	(+12)
Upper Extremity	0.79	(+3)
Fine-motor Control	0.86	(+1)
Sensorimotor Coordination	0.78	(+27)

Matthews correlation coefficient 0.80 (+8)

Results



SMEs in the future

Individuals vs Group Consensus

Category	F1 Score
Cognitive	0.91
Cardiopulmonary	0.74
Lower Extremity	0.71
Upper Extremity	0.76
Fine-motor Control	0.85
Sensorimotor Coordination	0.51

Matthews correlation coefficient 0.72

GenAI vs Group Consensus

Category	F1 Score	
Cognitive	0.92	(+1)
Cardiopulmonary	0.82	(+8)
Lower Extremity	0.83	(+12)
Upper Extremity	0.79	(+3)
Fine-motor Control	0.86	(+1)
Sensorimotor Coordination	0.78	(+27)

Matthews correlation coefficient 0.80 (+8)

Time Efficiency



100 tasks took each
person:
~ 2 hours

Time Efficiency



100 tasks took each
person:
~ 2 hours



100 tasks took AI:
~ 4 minutes

+

SMEs review for validation

Limitations



- 100 tasks for validation

Limitations



- 100 tasks for validation
- 3 human reviewers with similar background

Limitations



- 100 tasks for validation
- 3 human reviewers with similar background
- Browser version lacks ability to control for variation

Limitations



- 100 tasks for validation
- 3 human reviewers with similar background
- Browser version lacks ability to control for variation
- Some task descriptions are vague – hard for humans and AI

Conclusion



- Benefits of GenAI use: Reliability, automation, speed, cost

Conclusion and Future Work



- Benefits of GenAI use: Reliability, automation, speed, cost

- Pick human reviewers from different backgrounds

Conclusion and Future Work



- Benefits of GenAI use: Reliability, automation, speed, cost
- Pick human reviewers from different backgrounds
- SMEs review the group consensus mapping

Conclusion and Future Work



- Benefits of GenAI use: Reliability, automation, speed, cost

- Pick human reviewers from different **backgrounds**
- **SMEs** review the group consensus mapping
- GenAI map **all tasks** to HSPCs

Conclusion and Future Work



- Benefits of GenAI use: Reliability, automation, speed, cost

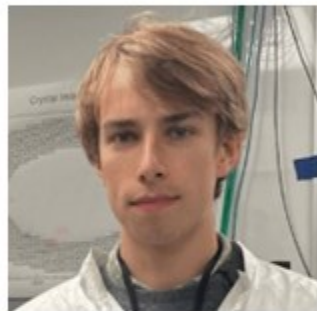
- Pick human reviewers from different **backgrounds**
- **SMEs** review the group consensus mapping
- GenAI map **all tasks** to HSPCs
- Use GenAI with **API** to control for variation

Conclusion and Future Work



- Benefits of GenAI use: Reliability, automation, speed, cost
- Pick human reviewers from different backgrounds
- SMEs review the group consensus mapping
- GenAI map all tasks to HSPCs
- Use GenAI with API to control for variation
- Refine HSPCs and their description with input from SMEs

Team





Thank you



Back-up Slides

How an LLM processes natural language



Input

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.



How an LLM processes natural language



Input

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.



Tokens

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.

How an LLM processes natural language



Input

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.



Tokens

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.

Vectors

[9408, 3580, 21963, 11, 34410, 316, 5888, 33143, 2420, 11, 316, 322, 58856, 28289, 14, 12580, 103279, 538, 8155, 558]

How an LLM processes natural language



Input

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.



Tokens

Enter control inputs, manually to override automated system, to reconfigure CM/SM heaters if necessary.

Vectors

[9408, 3580, 21963, 11, 34410, 316, 5888, 33143, 2420, 11, 316, 322, 58856, 28289, 14, 12580, 103279, 538, 8155, 558]

Prediction

{ "Cognitive": 1, "Cardiopulmonary": 0, "Lower Extremity": 0, "Upper Extremity": 0, "Fine-motor Control": 1, "Sensorimotor Coordination": 0 }



Prompt engineering

- Creating model-specific optimal instructions

Finetuning

- Training a model on domain-specific data

Retrieval Augmented Generation

- Providing a knowledge base (PDFs) that a model can reference

General purpose LLMs

- ChatGPT, Claude, Gemini

Evaluation Metrics



Recall: Did we find all the positives? (**Completeness**)

Precision: When we called something positive, how often were we right? (**Accuracy of Positive Predictions**)

F1 Score

- **Purpose:** Balances precision and recall.
- **Best For:** Evaluating imbalanced datasets.
- **Range:** 0 (worst) to 1 (best).

$$F1 = 2 \cdot \frac{\textit{Precision} \cdot \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

Matthews Correlation Coefficient (MCC)

- **Purpose:** Measures overall prediction quality, accounting for all confusion matrix components.
- **Best For:** Binary and multiclass tasks.
- **Range:** -1 (inverse) to 1 (perfect).

Results



SMEs in the future

Individuals vs Group Consensus

Category	F1 Score	Support
Cognitive	0.91	75
Cardiopulmonary	0.74	10
Lower Extremity	0.71	8
Upper Extremity	0.76	13
Fine-motor Control	0.85	65
Sensorimotor Coordination	0.51	16
Matthews correlation coefficient	0.72	

GenAI vs Group Consensus

Category	F1 Score	
Cognitive	0.92	(+1)
Cardiopulmonary	0.82	(+8)
Lower Extremity	0.83	(+12)
Upper Extremity	0.79	(+3)
Fine-motor Control	0.86	(+1)
Sensorimotor Coordination	0.78	(+27)
Matthews correlation coefficient	0.80	(+8)