0.0.1 Transformation algorithm

This subsection compares the transformation algorithms that we have tested and evaluates their suitability for achieving the prototype's goal (Table 1).

The first algorithms we considered were ANN Poplazarova et al. 2020. ANN should be more accurate as it has been validated by a big company, based on the paper "Ethical decision-making in biopharmaceutical research and development applying values using the TRIP/TIPP model". The model is accessible via API and has been previously proven. But, it is only work for certain context.

Secondly, sentiment analysis, it can be done locally. But, it has no proof of accuracy and has not been previously proven. It does not have pretrained models and the technique is quite old as well.

Next, we tried several Large-Language Model (LLM). The script we used for comparison are all available on the Github(cinapr 2024).

The first LLM is GPT4ALL. It utilizes several HuggingFace pre-trained models as the base. The models were chosen based on their performance recommendations on the Hugging Face website. The inclusion criteria included models that were in English, lightweight, and capable of running on the device. Therefore, four models were selected:

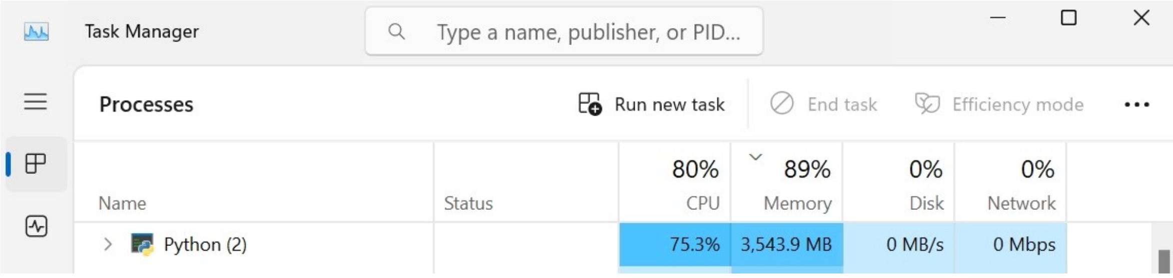
```
mistral-7b-openorca.gguf2.Q4_0.gguf (424.12 seconds)
Nous-Hermes-2-Mistral-7B-DPO.Q4_0.gguf (544.64 seconds)
orca-mini-3b-gguf2-q4_0.gguf (152.23 seconds)
replit-code-v1_5-3b-newbpe-q4_0.gguf (0.96 seconds)
```

Due to the limited availability of device resources, instead of fine-tuning and comparing them, we first asked a simple question to compare how quickly each model could respond. When asked, *"What is London?"*, the `orca-mini-3b-gguf2-q4_0.gguf` model responded in 152.23 seconds, while `Nous-Hermes-2-Mistral-7B-DPO.Q4_0.gguf` and `mistral-7b-openorca.gguf2.Q4_0.gguf` responded in 544.64 seconds and 424.12 seconds, respectively. Unfortunately, the `replit-code-v1_5-3b-newbpe-q4_0.gguf` model was unloadable due to the wrong number of tensors.

Even though this approach was considered the best for data privacy concerns since it runs locally, it was deemed unsuitable due to the excessive time taken to answer basic questions. The fastest response took more than two minutes, which is impractical for our prototype that needs to loop through numerous user stories.

Next, we compared GPT2. GPT2 is run on a local server. It took 1405.1124 seconds to be trained, and the classification compared to the ECCOLA sample dataset returned a 3.61% similarity. On average, the LLM took 0.195 seconds to classify each user story sample, showing better speed performance than GPT4ALL. Unfortunately, its reasoning is still weak, and future fine-tuning is not feasible as the basic fine-tuning already consumes more than 80% of the CPU and RAM of the available machine (Figure 1).

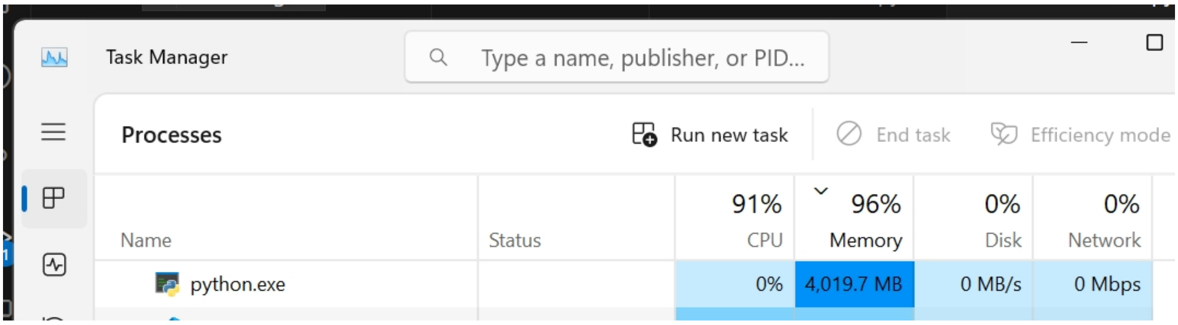
The next algorithm we compared, after fine-tuning, was BART. The base model for BART is not located on the client machine. It took 2059.9524 seconds to be trained, and the clas-



Name	Status	80% CPU	89% Memory	0% Disk	0% Network
Python (2)		75.3%	3,543.9 MB	0 MB/s	0 Mbps

Figure 1: Task Manager Screenshot Showing GPT2 Fine-Tuning Consumption

sification compared to the ECCOLA sample dataset returned a 7.22% similarity. It showed better speed performance than GPT4ALL. Unfortunately, its reasoning remains weak, and future fine-tuning is not possible since the basic fine-tuning consumes more than 90% of the CPU and RAM of the available machine (Figure 2). Moreover, compared to GPT2, BART still falls short as the base model is hosted on an external server.



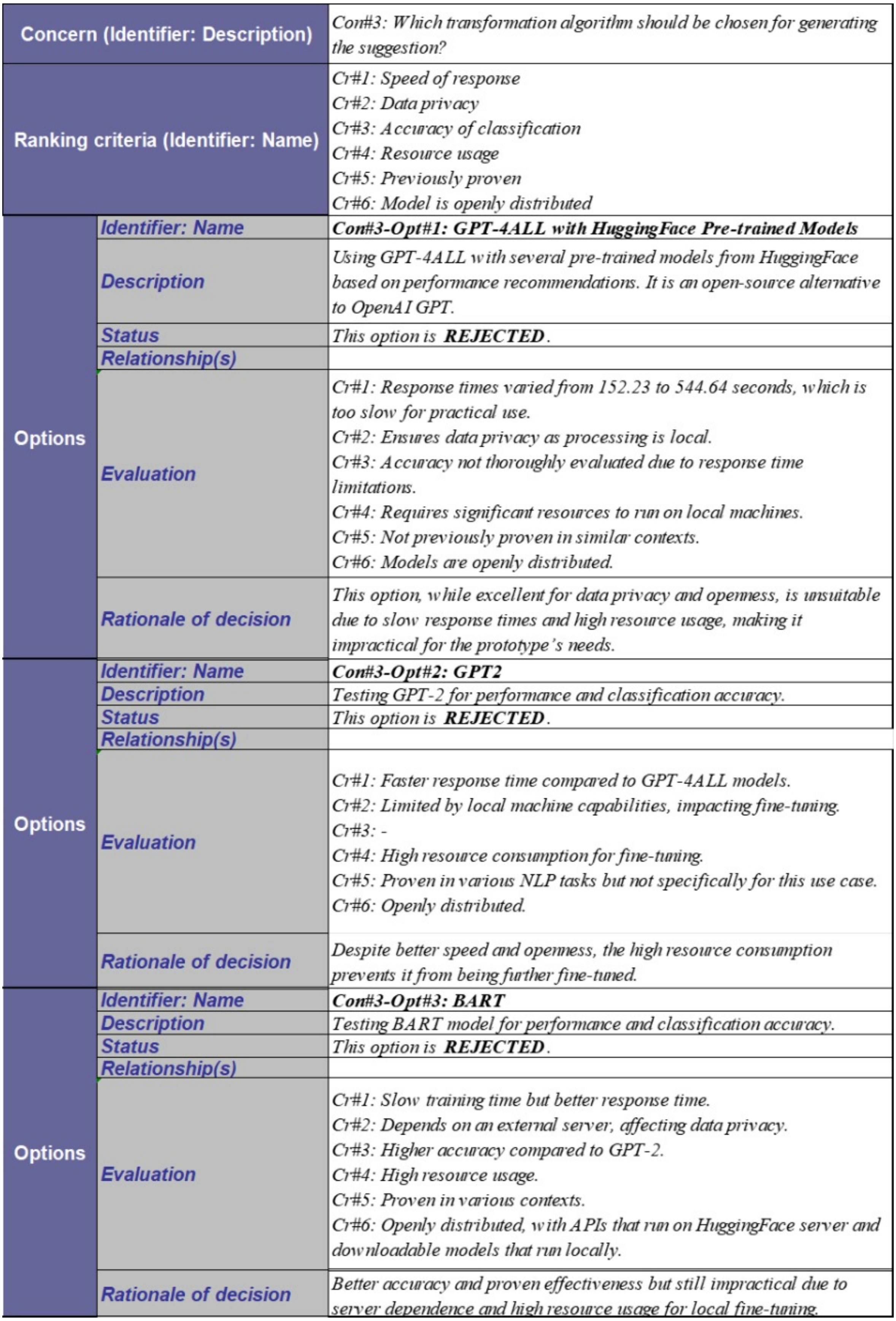
Name	Status	91% CPU	96% Memory	0% Disk	0% Network
python.exe		0%	4,019.7 MB	0 MB/s	0 Mbps

Figure 2: Task manager screenshot that showing BART fine-tuning consumption

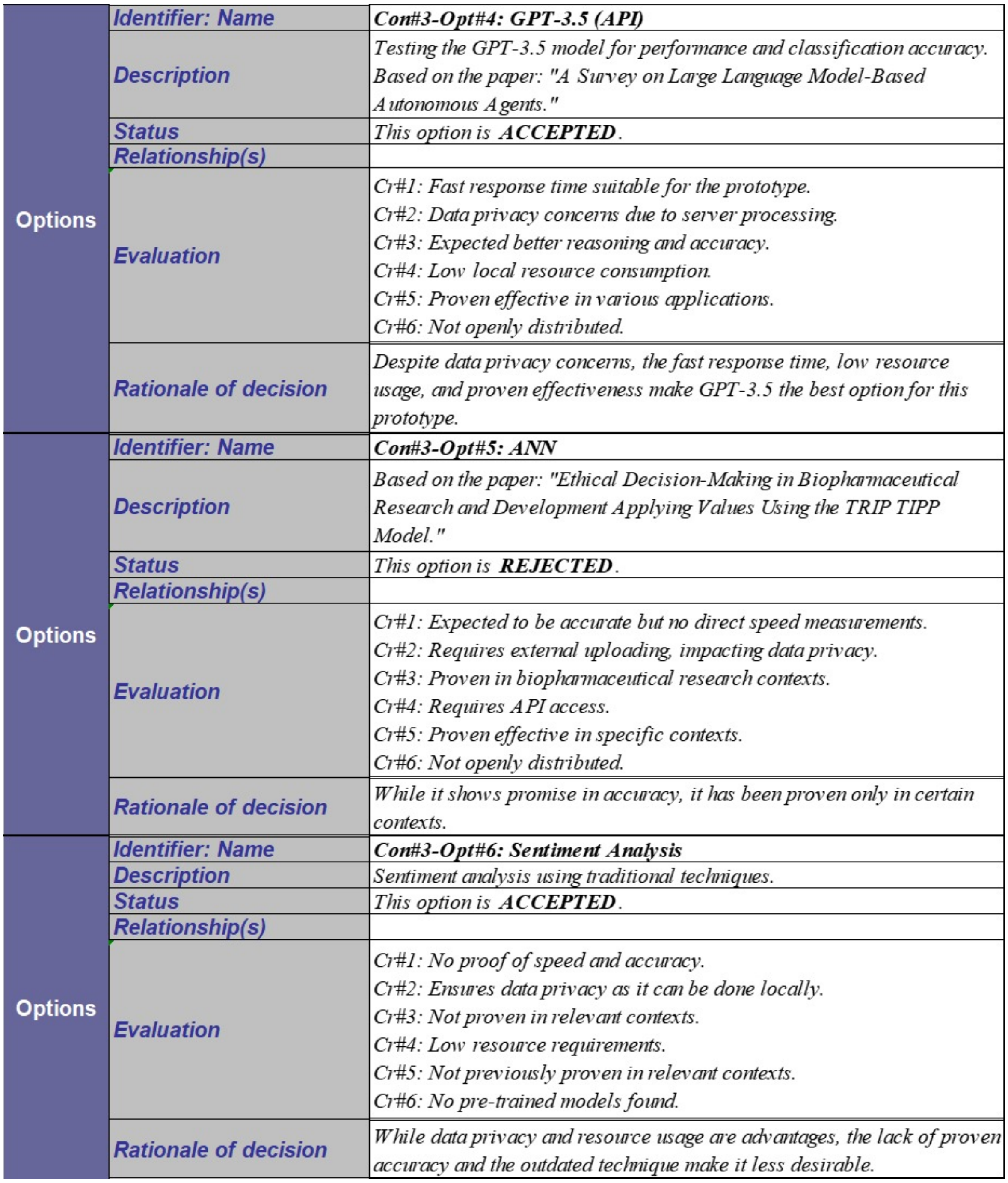
The fourth LLM algorithm we compared was GPT-3.5, which runs on a server owned by OpenAI. It took 963.47 seconds to be trained, and the classification compared to the ECCOLA sample dataset returned 26.32%. Although it lacks in terms of privacy, as it runs on an OpenAI server, we have reviewed the privacy policies provided by OpenAI (OpenAI n.d.(a)OpenAI 2023OpenAI n.d.(b)OpenAI n.d.(c)). Since the LLM process does not run on a local machine, this algorithm offers the best performance in terms of speed and minimal CPU and RAM consumption. Therefore, better reasoning can be expected from this algorithm.

From the perspective of machine availability, speed, and project timeline, we chose GPT-3.5, which runs on an OpenAI server. However, we recommend that future researchers with better machine or server availability consider deeper fine-tuning of GPT2 to improve its reasoning skills. GPT2, which runs locally, would be better for user story data privacy.

Table 1: Transformation Algorithm



Concern (Identifier: Description)		*Con#3: Which transformation algorithm should be chosen for generating the suggestion?*
Ranking criteria (Identifier: Name)		*Cr#1: Speed of response* *Cr#2: Data privacy* *Cr#3: Accuracy of classification* *Cr#4: Resource usage* *Cr#5: Previously proven* *Cr#6: Model is openly distributed*
Options	Identifier: Name	***Con#3-Opt#1: GPT-4ALL with HuggingFace Pre-trained Models***
	Description	*Using GPT-4ALL with several pre-trained models from HuggingFace based on performance recommendations. It is an open-source alternative to OpenAI GPT.*
	Status	*This option is **REJECTED**.*
	Relationship(s)	
	Evaluation	*Cr#1: Response times varied from 152.23 to 544.64 seconds, which is too slow for practical use.* *Cr#2: Ensures data privacy as processing is local.* *Cr#3: Accuracy not thoroughly evaluated due to response time limitations.* *Cr#4: Requires significant resources to run on local machines.* *Cr#5: Not previously proven in similar contexts.* *Cr#6: Models are openly distributed.*
	Rationale of decision	*This option, while excellent for data privacy and openness, is unsuitable due to slow response times and high resource usage, making it impractical for the prototype's needs.*
Options	Identifier: Name	***Con#3-Opt#2: GPT2***
	Description	*Testing GPT-2 for performance and classification accuracy.*
	Status	*This option is **REJECTED**.*
	Relationship(s)	
	Evaluation	*Cr#1: Faster response time compared to GPT-4ALL models.* *Cr#2: Limited by local machine capabilities, impacting fine-tuning.* *Cr#3: -* *Cr#4: High resource consumption for fine-tuning.* *Cr#5: Proven in various NLP tasks but not specifically for this use case.* *Cr#6: Openly distributed.*
	Rationale of decision	*Despite better speed and openness, the high resource consumption prevents it from being further fine-tuned.*
Options	Identifier: Name	***Con#3-Opt#3: BART***
	Description	*Testing BART model for performance and classification accuracy.*
	Status	*This option is **REJECTED**.*
	Relationship(s)	
	Evaluation	*Cr#1: Slow training time but better response time.* *Cr#2: Depends on an external server, affecting data privacy.* *Cr#3: Higher accuracy compared to GPT-2.* *Cr#4: High resource usage.* *Cr#5: Proven in various contexts.* *Cr#6: Openly distributed, with APIs that run on HuggingFace server and downloadable models that run locally.*
	Rationale of decision	*Better accuracy and proven effectiveness but still impractical due to server dependence and high resource usage for local fine-tuning.*



Options	Identifier: Name	**Con#3-Opt#4: GPT-3.5 (API)**
	Description	*Testing the GPT-3.5 model for performance and classification accuracy. Based on the paper: "A Survey on Large Language Model-Based Autonomous Agents."*
	Status	*This option is* ***ACCEPTED****.*
	Relationship(s)	
	Evaluation	*Cr#1: Fast response time suitable for the prototype.* *Cr#2: Data privacy concerns due to server processing.* *Cr#3: Expected better reasoning and accuracy.* *Cr#4: Low local resource consumption.* *Cr#5: Proven effective in various applications.* *Cr#6: Not openly distributed.*
	Rationale of decision	*Despite data privacy concerns, the fast response time, low resource usage, and proven effectiveness make GPT-3.5 the best option for this prototype.*
Options	Identifier: Name	**Con#3-Opt#5: ANN**
	Description	*Based on the paper: "Ethical Decision-Making in Biopharmaceutical Research and Development Applying Values Using the TRIP TIPP Model."*
	Status	*This option is* ***REJECTED****.*
	Relationship(s)	
	Evaluation	*Cr#1: Expected to be accurate but no direct speed measurements.* *Cr#2: Requires external uploading, impacting data privacy.* *Cr#3: Proven in biopharmaceutical research contexts.* *Cr#4: Requires API access.* *Cr#5: Proven effective in specific contexts.* *Cr#6: Not openly distributed.*
	Rationale of decision	*While it shows promise in accuracy, it has been proven only in certain contexts.*
Options	Identifier: Name	**Con#3-Opt#6: Sentiment Analysis**
	Description	*Sentiment analysis using traditional techniques.*
	Status	*This option is* ***ACCEPTED****.*
	Relationship(s)	
	Evaluation	*Cr#1: No proof of speed and accuracy.* *Cr#2: Ensures data privacy as it can be done locally.* *Cr#3: Not proven in relevant contexts.* *Cr#4: Low resource requirements.* *Cr#5: Not previously proven in relevant contexts.* *Cr#6: No pre-trained models found.*
	Rationale of decision	*While data privacy and resource usage are advantages, the lack of proven accuracy and the outdated technique make it less desirable.*