

The Open University  
Faculty of Science, Technology, Engineering & Mathematics  
Knowledge Media Institute (KMi)

# **Designing a Personal Awareness Agent to Ameliorate Information Overload**

David Goddard

## SUPPLEMENTAL MATERIAL



 [orcid.org/0000-0002-5565-9692](https://orcid.org/0000-0002-5565-9692)

Supervised by Dr. Paul Mulholland, Dr. Lara Piccolo & Dr. Enrico Daga

# Copyright & Licensing

© 2025 **David Goddard**. All rights reserved.

This supplemental material is licensed under CC BY-SA 4.0.

- CC Creative Commons
- BY Attribution Required
- SA Share Alike

To view a copy of this license, visit: <https://creativecommons.org/licenses/by-sa/4.0/>.

## Document Object Identifiers

The following DOIs have been defined for this work:

- This consolidated document: [doi:10.21954/ou.rd.29293850](https://doi.org/10.21954/ou.rd.29293850)
- Thesis to which these supplements relate: [doi:10.21954/ou.ro.00104313](https://doi.org/10.21954/ou.ro.00104313)
- Overall collection of thesis materials: [doi:10.21954/ou.rd.29293850](https://doi.org/10.21954/ou.rd.29293850)

## Document Information

This document is a consolidated collection of the supplemental materials documents for the PhD thesis that have been individually published as Supplements S0 to S12. The DOIs of the individual standalone documents are listed in the page footers.

# Contents

<b>Title Page</b>	<b>1</b>
<b>S0 Index of Thesis Supplemental Material</b>	<b>3</b>
<b>S1 Design and Development Log</b>	<b>7</b>
<b>S2 Survey Questions</b>	<b>37</b>
<b>S3 Survey Demographics</b>	<b>45</b>
<b>S4 Personas</b>	<b>51</b>
<b>S5 Persona Scenarios</b>	<b>59</b>
S5.1 Susan . . . . .	59
S5.2 Adam . . . . .	61
S5.3 Phoebe . . . . .	62
S5.4 Kenton . . . . .	64
S5.5 Usha . . . . .	65
<b>S6 Synthetic Content Materials</b>	<b>69</b>
S6.1 Simulation Schemas . . . . .	69
S6.2 Simulated Content Prompting . . . . .	73
S6.3 Dramatis Personae . . . . .	75
S6.4 Entities . . . . .	78
S6.5 Content Scheduling . . . . .	80
<b>S7 Synthetic Content Topics</b>	<b>83</b>
S7.1 Susan . . . . .	83
S7.2 Adam . . . . .	87
S7.3 Phoebe . . . . .	91
S7.4 Kenton . . . . .	93
S7.5 Usha . . . . .	96
<b>S8 Synthetic Evaluation Schemas &amp; Prompting</b>	<b>103</b>
S8.1 Evaluation Schemas . . . . .	103
S8.2 Evaluation Schemas - Data Service . . . . .	117
S8.3 Static Prompt Elements . . . . .	124
S8.4 Evaluation JSON Examples . . . . .	133
S8.5 Evaluation Processor . . . . .	140

<b>S9 Synthetic Evaluation Fine Tuning Examples</b>	<b>155</b>
S9.1 Example – Train AGREE . . . . .	155
S9.2 Example – Train DISAGREE . . . . .	157
<b>S10 Study Configuration &amp; Information</b>	<b>161</b>
S10.1 Technical Details Sheet Example . . . . .	162
S10.2 Persona Configurations . . . . .	165
<b>S11 Study Detailed Results</b>	<b>179</b>
S11.1 Statistics . . . . .	180
S11.2 Selection of Evaluation Items . . . . .	200
S11.3 Evaluation Token Usage . . . . .	221
<b>S12 Source Code &amp; Additional Materials</b>	<b>225</b>
S12.1 Source Code . . . . .	225
S12.2 L <sup>A</sup> T <sub>E</sub> X . . . . .	226

# **Supplement S0**

## **Index of Thesis Supplemental Material**

# **Index of Thesis Supplemental Material**

This document contains links to resources relating to the PhD thesis by David Goddard:  
*Designing a Personal Awareness Agent to Ameliorate Information Overload*.

The thesis is available at: [doi:10.21954/ou.ro.00104313](https://doi.org/10.21954/ou.ro.00104313).

A GitHub repository containing resources is available at:

[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) (<https://github.com/revisionist/ou-phd-resources>).

A collection of materials for this research is available at: [doi:10.21954/ou.rd.c.7588025](https://doi.org/10.21954/ou.rd.c.7588025).

## **S1 Design and Development Log**

[doi:10.21954/ou.rd.28045163](https://doi.org/10.21954/ou.rd.28045163) Unabridged version of the Design and Development Log, providing supporting material for the RtD process.

## **S2 Survey Questions**

[doi:10.21954/ou.rd.28045166](https://doi.org/10.21954/ou.rd.28045166) List of all questions from the survey in Chapter 5.

## **S3 Survey Demographics**

[doi:10.21954/ou.rd.28045442](https://doi.org/10.21954/ou.rd.28045442) Information on key demographics of respondents to the Chapter 5 survey.

## **S4 Final Personas**

[doi:10.21954/ou.rd.28045454](https://doi.org/10.21954/ou.rd.28045454) PATHY Personas that were derived using the process covered in Chapter 5.

## **S5 Persona Scenarios**

[doi:10.21954/ou.rd.28045460](https://doi.org/10.21954/ou.rd.28045460) Persona Scenarios that have been written for the personas developed in Chapter 5.

## **S6 Synthetic Content Materials**

[doi:10.21954/ou.rd.28045469](https://doi.org/10.21954/ou.rd.28045469) Information relating to Chapter 7 Synthetic Content, including schemas and example data for content requests and simplified Python code example.

**S7 Synthetic Content Topics**

**doi:10.21954/ou.rd.28045490** Full list of simulated content topics for personas as referenced in Chapter 7.

**S8 Synthetic Evaluation Schemas & Prompting**

**doi:10.21954/ou.rd.28045547** Details of prompting resources used in the Synthetic Evaluation process discussed in Chapter 8.

**S9 Synthetic Evaluation Fine Tuning Examples**

**doi:10.21954/ou.rd.28045562** Examples relating to fine tuning within the Synthetic Evaluation process.

**S10 Study Configuration & Information**

**doi:10.21954/ou.rd.28045577** Study configuration and technical details sheet provided to study participants.

**S11 Study Detailed Results**

**doi:10.21954/ou.rd.28045580** Detailed results data for the Chapter 9 study.

**S12 Source Code & Additional Materials**

**doi:10.21954/ou.rd.28045598** Information on source code for software used in this research, and additional materials.

These supplements were originally published in December 2024 and January 2025. They were subsequently updated in June 2025 (v2) with revised formatting; however, the substantive content remains unchanged from the originals.

# **Supplement S1**

## **Design and Development Log**

## S1 Design and Development Log

---

# Design and Development Log

This document contains the full version of the Design and Development Log, providing supporting material for the Research Through Design process we followed in our work.

<b>2017-08-20</b>	<b>Define Awareness Agent Eclipse project</b>
<i>AwAg Platform</i>	Taking a previous simple test application, we created a new Eclipse Java project to as the basis for an extensible platform that allows for modularity and multiple implementation types, with Platform, Service and Application sub-projects.
<b>2018-09-26</b>	<b>Add Cognos reporting</b>
<i>AwAg Platform</i>	Initial approach to extracting information from project, using a database-based reporting tool.
<b>2019-07-19</b>	<b>Switch from JBot to jSlack</b>
<i>AwAg Platform</i>	Our original Slack test code used a relatively limited library called JBot <sup>a</sup> <sup>b</sup> , designed to get you up and running quickly with Slack bots. This was fine for initially proving some concepts, but the full range of our design depended on being able to fully access the Slack API, so the project was transitioned to the full-featured jSlack API <sup>cde</sup> .
	<sup>a</sup> <a href="https://github.com/rampatra/jbot">https://github.com/rampatra/jbot</a>
	<sup>b</sup> <a href="https://blog.rampatra.com/how-to-make-a-slack-bot-in-java">https://blog.rampatra.com/how-to-make-a-slack-bot-in-java</a> [ <a href="https://perma.cc/ME5K-7AXW">https://perma.cc/ME5K-7AXW</a> ]
	<sup>c</sup> <a href="https://tools.slack.dev/java-slack-sdk/guides/web-api-basics/">https://tools.slack.dev/java-slack-sdk/guides/web-api-basics/</a>
	<sup>d</sup> <a href="https://perma.cc/KL7M-D8HX">https://perma.cc/KL7M-D8HX</a>
	<sup>e</sup> jSlack later transitioned to <a href="https://github.com/slackapi/java-slack-sdk">https://github.com/slackapi/java-slack-sdk</a> , having been taken on with its original developer <a href="https://github.com/seratch">https://github.com/seratch</a> by Slack itself
<b>2019-07-25</b>	<b>Add Cloudant-based module configuration</b>
<i>AwAg Platform</i>	To address the significant configuration needs of the design scope, and to facilitate multiple agents running in cloud-based environments, we transitioned the application to use a commercial CouchDB object store for configuration data.
<b>2019-07-26</b>	<b>Add Slack RTM support</b>
<i>AwAg Acquire</i>	To better support real-time monitoring of content in Slack, we added the now-legacy RTM API <sup>a</sup> to the platform, enabling us to listen for content more effectively.

---

<sup>a</sup><https://api.slack.com/legacy/rtm> [<https://perma.cc/QP7L-5FZB>]

## S1 Design and Development Log

---

<b>2019-09-24</b>	<b>First implementation of a Content Item</b>
<i>AwAg Platform</i>	Initial testing used Slack content more or less natively, which was effective for the scope of the first prototype, but it became clear from our work that data abstraction and standardisation was needed to build a system that would be flexible and extensible enough to accommodate many diverse data sources. This point marked the start of developing the Content Item concept, in the first instance by applying a simple layer of abstraction over native Slack items.
<b>2019-09-24</b>	<b>Add concepts of Base Operation and Service Operations</b>
<i>AwAg Platform</i>	This was the start of a development branch that did not end up going anywhere. We initially designed the agent platform with the concept of several operation types - instructions for the agent to do something. While the support for this is still present, testing with real sources showed that a different type of control was needed - with what eventually became the administration side of the Interact service.
<b>2019-10-15</b>	<b>Add the concept of Augmentation Items</b>
<i>AwAg Augment</i>	We had tried several approaches so far to enhancing content to make it more easily actionable, with some issues arising: while it was certainly possible to summarise items or apply a classification, there was no organised framework within which this could sit, and the approach was not extensible. The previous successful introduction of the Content Item concept gave us an opportunity to address this issue in a structured way, by extending the CI structure to include Augmentations so that we could be both flexible about what enhancements we could add and also have a mechanism for abstracting this and managing/addressing the information.
<b>2019-10-16</b>	<b>Backtrack on Operation changes introduced on September 24</b>
<i>AwAg Platform</i>	These had proved impractical and not entirely useful in testing.
<b>2019-10-16</b>	<b>Introduce a Consignment and Queue based approach</b>
<i>AwAg Platform</i>	Moving away from the previous content/service operations approach, we added a wrapper for Content Items named Content Consignments (essentially just a collection of CIs) and added a queue system for these. This would then go on to form the design paradigm for intra-service content transfer in the Agent via the Exchange Service.
<b>2019-10-17</b>	<b>Add simple text based explanation mechanism for Augmentations</b>
<i>AwAg Augment</i>	First introduction of an explainability mechanism within the augmentation framework.

## S1 Design and Development Log

---

**2019-10-19**

*AwAg  
Platform*

### **Reorganise modules into Acquire, Augment, Control etc.**

While this development change was little more than a code refactoring at this stage, this was a significant design development. It marked the introduction of the concept of high level components within the Awareness Agent system model, with the creation of the Acquire and Augment module types to handle those logical operations. This was a result of initial work to expand the scope of these aspects of functionality - early testing showed us that a consistent modular approach was needed to allow for the future expansion of scope that we were looking for.

**2019-10-19**

*AwAg  
Platform*

### **Introduce Triage service**

Looking at the content flow through the prototype Awareness Agent, we realised that there was a need for a first point of contact assessment of incoming content - a triage process to borrow from medical terminology. We added the Triage service to act as the first port of call for all content coming from Acquire services, adding an ability to discard unwanted content based on simple heuristics before going on to Augmentation.

**2019-10-21**

*AwAg CI*

### **Expand and improve metadata handling**

Recognising a need to have more metadata information about content available further down the line, we improved the way that metadata is stored in individual CIs, and also added increased metadata at the consignment level.

**2019-10-21**

*AwAg  
Platform*

### **Add Discard service**

Before this point we had simply been dropping unwanted content by not passing it on to the next stage; however we recognised that there was utility in tracking discarded content - both from a scientific and application development perspective. By adding a service to which all unwanted items could be sent, we acquired to capability to easily record information about discarded items and also take any other actions appropriate at that point. This approach also had greater consistency with the overall CI queue-service flow model, with the Discard service fitting into a structured lifecycle for every CI.

**2019-10-26**

*AwAg CI*

### **Add support for new Slack-specific metadata fields**

We started formalising the concept of Extended Fields in the Content Item after testing showed us that we needed to be able to surface some source-specific information later in the CI lifecycle (for example at the point of augmentation or interaction). We recognised that this would be a generic need - sources other than Slack would need their own specific metadata - and we developed extended fields to fit this information within a standard CI structure in a manner consistent with the standard fields approach that we had already adopted.

## S1 Design and Development Log

---

**2019-11-01**

AwAg CI

### Transition CI data to using JSON-LD

As a consequence of the development of extended fields, we looked again at the basic data structure of the Content Item, which had hitherto been entirely bespoke JSON. We recognised that JSON-LD<sup>a</sup> and the related concept of linked data notifications (LDN)<sup>b,c</sup> provided us with an academically rigorous and supported mechanism for working with and sharing data from multiple sources, so we transitioned the internal CI structure to one compatible with JSON-LD.

<sup>a</sup><https://www.w3.org/TR/json-ld/>

<sup>b</sup><https://www.w3.org/TR/ldn/>

<sup>c</sup><https://csarven.ca/linked-data-notifications> [<https://perma.cc/7SVP-JPBV>]

**2019-11-09**

AwAg CI

### Distinguish standard & extended fields with prefix

As part of the transition to JSON-LD we looked again at how properties (fields) in the Content Item were named. We had three categories of property: native (existing in the source data item), standard, and extended. We decided to take an approach of using Compact IRIs (CURIEs)<sup>a</sup>, with the namespace awag mapped to Awareness Agent schema <http://parse.net/awag/0.1/> in the @Context. We gave standard fields (where they existed as independent properties) a shorthand namespace of awag:std, and extended fields a namespace of awag:xtd. We made a decision to use unchanged names for native properties from the source system rather than renaming or adding an additional namespace. There had been a number of ways that we might approach this, and we retrospectively discuss this further the Reflections section.

<sup>a</sup><https://www.w3.org/TR/2010/NOTE-curie-20101216/> [<https://perma.cc/443S-8ELP>]

**2020-09-28**

AwAg  
Acquire

### Add LDN module type

We added the basic framework for supporting LDN as an Acquire source via a client for an LDN inbox. We implemented our own local Solid server (nodeSolidServer<sup>a</sup> running on FreeBSD via PM2<sup>b,c</sup>). We ultimately never progressed beyond the proof of concept stage for LDN Acquire support via Solid – although the technology was potentially useful, but did not have the level of actual content that we wanted to run a study.

<sup>a</sup><https://github.com/nodeSolidServer>

<sup>b</sup><https://www.npmjs.com/package/pm2> [<https://perma.cc/8ER6-KGAJ>]

<sup>c</sup><https://gist.github.com/PieterScheffers/457769f2090c6b69cd9d>

## S1 Design and Development Log

---

<b>2020-10-06</b>	<b>Use Juneau to serialise to/from Json</b>
<i>AwAg Platform</i>	We were storing JSON objects in a commerical CouchDB application, using the supplied libraries to store objects (and implicitly serialise them) - we found that this was actually generating incorrect JSON for some of our objects after a round trip, so we moved to Apache Juneau <sup>a</sup> for this serialisation also, giving us more control over the process (we had already used Juneau extensively for JSON serialisation and REST service provision).
	<sup>a</sup> <a href="https://juneau.apache.org/">https://juneau.apache.org/</a> [ <a href="https://perma.cc/6ELF-TA3B">https://perma.cc/6ELF-TA3B</a> ]
<b>2020-10-08</b>	<b>Add tracking of Content Items</b>
<i>AwAg CI</i>	This was a necessary step in developing the mechanism for CI exchanging between agents - we added a persistent tracking mechanism to record CIs that the agent had already originated or seen, so that if offered back to us in a future exchange we could decline them (or take appropriate actions such as looking for new augmentations added by a partner agent)
<b>2020-10-08</b>	<b>Add Augment engine and structure</b>
<i>AwAg Augment</i>	Significant internal change to the organisation of the Augment process, adding a new engine to support it.
<b>2020-11-08</b>	<b>Add basic outbound exchange functionality via MQTT</b>
<i>AwAg Exchange</i>	The initial implementation of the Exchange component uses MQTT; we implemented code to manage the round trip serialisation of Content Items to/from binary data to be sent and received from MQTT.
<b>2020-11-11</b>	<b>Change MQTT serialisation to Kryo</b>
<i>AwAg Exchange</i>	We found in testing that serialising our complex Content Item structure to binary – and being able to successfully instantiate the returned deserialised item – is actually quite difficult. We tried here with the Kryo library <sup>a</sup> instead of the original approach of using native Java bytestream functionality.
	<sup>a</sup> <a href="https://github.com/EsotericSoftware/kryo">https://github.com/EsotericSoftware/kryo</a>
<b>2021-04-20</b>	<b>Change to use Slack Conversations API</b>
<i>AwAg Acquire &amp; Interact</i>	Slack made a change from using “Channels” terminology to referring to “Conversations” instead, so we reflected that change. <sup>ab</sup>
	<sup>a</sup> <a href="https://api.slack.com/docs/conversations-api">https://api.slack.com/docs/conversations-api</a> [ <a href="https://perma.cc/W7QM-5FK7">https://perma.cc/W7QM-5FK7</a> ]
	<sup>b</sup> <a href="https://api.slack.com/methods/conversations.list">https://api.slack.com/methods/conversations.list</a> [ <a href="https://perma.cc/6Z72-XCZN">https://perma.cc/6Z72-XCZN</a> ]

## S1 Design and Development Log

---

**2021-04-21**

*AwAg  
Exchange*

### Rework MQTT serialisation again

We had continued to experience problems with the Content Item lifecycle even with the Kryo library, so changed approach once more. This (successful) iteration uses Apache Juneau for object-JSON serialisation, and Apache SerializationUtils<sup>a</sup> for string-binary conversion. The Juneau approach had been a well tested feature of other elements of the agent so far, so we were able to get consistent object-string conversion using it, and the Apache serialisation tools worked reliably for the string-binary conversion.

<sup>a</sup><https://commons.apache.org/proper/commons-lang/apidocs/org/apache/commons/lang3/Ser>  
[ializationUtils.html \[https://perma.cc/UUR5-57VM\]](https://perma.cc/UUR5-57VM)

**2021-05-14**

*AwAg CI &  
Interact*

### Add 'Prominence'

Adding the concept of 'prominence' a categorical code that can be assigned to a Content Item having possible values LOW, NORMAL, HIGH, HIGHER. We developed this as a way of surfacing augmentations: the prominence is a property at the Content Item level, and is effectively another Standard Field, but the mechanism for setting it is left open; the assumption is that the Augment engine or indeed later services can adjust prominence as they process the CI based on augmentations or other factors. Prominence can then be used by Allocate or Interact services to change how the CI is processed or displayed. The design rationale for this was that it is an easy to check standard way of indicating how prominent a CI should be. We introduced this in response to difficulties we experienced using augmentations for this task – they have the information, but not accessible in a standard way. Having Prominence available allows components to make high level decisions about a CI without knowing the internal augmentation structure.

**2021-05-15**

*AwAg CI*

### Add permalink feature to Slack content item

Testing with user interface prototypes showed that the agent would benefit from being able to easily direct the user to the original item in context in the source system via a URL. In the case of Slack content, there was not such URL available directly, but we could construct one from the message content and current context within the Acquire service. We added this as a new extended Slack field (PERMALINK), but also mapped this as a Standard Field (URL) so that other types of CI can be used in the same way.

## S1 Design and Development Log

---

<b>2021-06-08</b>	<b>Add Slack Service</b>
AwAg Interact	This marked the start of adding two-way interactivity to the Slack Interact component, exposing functionality to receive events from Slack in response to user actions <sup>a</sup> . The first iteration of this was to add and test basic functionality; we would then go on to expand this and integrate with the Awareness Agent framework based on testing results.
	<sup>a</sup> Initially using the now-legacy RTM API – <a href="https://api.slack.com/legacy/rtm">https://api.slack.com/legacy/rtm</a>
<b>2021-06-20</b>	<b>Improve Slack interactivity</b>
AwAg Interact	Initial testing of the Slack service was successful, so we expanded the types of interaction that were possible and genericised the process by adding a SimpleSlackAction object to the code. We used this to add a set of internal queues for Slack actions that were based on abstracted Slack request objects, with a separate request handler component to process these. We did this to take an approach consistent with the design ethos of the rest of the agent, which we had found in testing to be adaptive to change and different types of request.
<b>2022-01-04</b>	<b>Initial creation of awagml</b>
ML Service	We added a new back-end service, referred to as “awagml”, to initially support generating classifications for items. The first instance was a simple classifier to test this functionality - initially testing with both MultinomialNB and SGDClassifier (SVM) classifiers using scikit-learn [Pedregosa et al., 2011]. While the Awareness Agent itself is a Java application, we decided to implement a separate ML service in Python for the following reasons: 1) availability and quality of ML and associated packages in Python, 2) support for rapid prototyping, and 3) consistency with design principle of keeping back-end ML commoditised and separate.
<b>2022-01-11</b>	<b>Add add_dataset_row</b>
ML Service	Initial work to allow us to add training items to the ML model and refit on the fly via an API call.
<b>2022-04-27</b>	<b>Implement Simple Classification Augmentation</b>
AwAg Augment	We added a single value classification for CIs using the new awagml back-end service. This passes CI content to the service and gets a classification back, which is added to the CI as an Augmentation. This first instance had no concept of multiple ML models, with awagml supporting only a single model for classification.

## S1 Design and Development Log

---

<b>2022-05-10</b>	<b>Initial support for multiple data dictionaries</b>
<i>ML Service</i>	The first step to moving to a multiple model design was to add support for multiple data dictionaries at the back end, identifiable by ID. This allowed us to manage different datasets for different models.
<b>2022-05-11</b>	<b>Service initialisation improvements</b>
<i>ML Service</i>	We implemented a hierarchical data structure for model data persisted to disk, with automatic creation and fitting of data sets on startup.
<b>2022-05-16</b>	<b>Add multi-model support</b>
<i>ML Service</i>	Completed initial implementation of multi-model support at the back end.
<b>2022-05-16</b>	<b>Add multi-model support</b>
<i>AwAg UDM</i>	Adding support for passing a model ID to the awagml back end was the next step following the addition of support for multiple models to awagml. This would allow us to move from proving the concept of classification augmentations to supporting full multi-model usage.
<b>2022-05-26</b>	<b>Add support for multi-model augmentations</b>
<i>AwAg UDM</i>	Having added model ID support in the previous iteration, we now changed the processing so that models could be specified in the augment instance configuration rather than a relatively fixed ID. This allowed us to drive the process dynamically using the augmentation configuration.
<b>2022-05-29</b>	<b>Add multi-channel output using classification to channel name map</b>
<i>AwAg UDM &amp; Interact</i>	We introduced a main element of User-Driven ML by implementing full multi-channel output in the Interact layer for augmented Content Items, where the model ID (classification name) and classification value are mapped to a Slack output channel. We made several user interaction design changes to enable this, changing from a single channel interaction paradigm to using multiple channels, with channels mapped to classification names and values. This required us to add support in the Interact configuration for the necessary mapping table in parallel to Augment configuration to apply matching augmentations.

## S1 Design and Development Log

---

<b>2022-05-31</b> <i>ML Service</i>	<b>Add support for multi-model classification training</b> Then process of training awagml had previously been a manual server-side command-line operation, where training text content was organised in directories for fitting and model use. This had been sufficient to prove the concept, but would not support the User-Directed ML process that we had in mind. We added an API to allow training requests to be made to awagml via REST calls, so that the Awareness Agent could directly initiate training events using its own 'live' data.
<b>2022-05-31</b> <i>AwAg UDML</i>	<b>Add support for multi-model classification training</b> Having successfully tested the process of applying multi-model classifications, we then looked to bring the whole model lifecycle within the scope of the Awareness Agent. We expanded our Slack interact handler to handle requests generated by the user interacting with published items, and added the necessary UI artefacts into those items themselves. With this we were able to test the full process of an item being classified, sent to a specific channel, then re-classified into a different channel by the user, with a concomitant back end request to re-train the item in awagml using the newly added training service.
<b>2022-06-01</b> <i>AwAg Augment</i>	<b>Change so that Model ID can be independent of Augmentation Name</b> Our original implementation depended on a match between the augmentation name and the model ID to use for a given augmentation. We found that this was limiting flexibility: while there was generally a relationship between the two this was not a given, and we might also in future want to be able to assign the same back-end model to different augmentations. We changed the design to instead use a configurable mapping between the two.
<b>2022-06-02</b> <i>ML Service</i>	<b>Add model lifecycle methods</b> To support the full model lifecycle under dynamic control of the Awareness Agent we added back end services to create/delete models and to add or remove classifications from models.

## S1 Design and Development Log

---

<b>2022-06-10</b>	<b>Remove RTM and replace with Events API</b>
AwAg Acquire	The Slack RTM API <sup>a</sup> that we had hitherto been using to receive callbacks from Slack had been deprecated in favour of the Events API <sup>b</sup> . While this was a forced change it was also a good opportunity for design re-evaluation: the RTM API involved opening a websocket to listen for events, whereas the Events API is subscription based, where a client subscribes to events and receives these to its configured URI as a HTTP REST call. This change allowed us to coalesce on two adaptable mechanisms for the Acquire service to use: a listener (acts as a service and receives HTTP calls, as is the case with the new Slack Events API) and a client (makes outgoing calls to a service on a scheduled or event-based basis). This meant that we were eventually able to design these mechanisms into a framework that could be applied to the majority of potential Acquire sources. The payload of REST calls to a listener type Acquire service would differ from source to source, but the service mechanism would be the same.
	<sup>a</sup> <a href="https://api.slack.com/legacy/rtm">https://api.slack.com/legacy/rtm</a> [ <a href="https://perma.cc/QP7L-5FZB">https://perma.cc/QP7L-5FZB</a> ]
	<sup>b</sup> <a href="https://api.slack.com/events">https://api.slack.com/events</a> [ <a href="https://perma.cc/MTM2-G27X">https://perma.cc/MTM2-G27X</a> ]
<b>2022-06-14</b>	<b>Add ML model management REST calls</b>
AwAg UDM	Previously the awagml service relied on manual processes at the back end for model management, but we had enhanced this to provide a REST interface for these tasks (model creation, deletion, changes to classifications). We now adapted the agent itself to match this, bringing full control over the model lifecycle within its purview. This was an important step for user-empowerment, with the user now able to control models via the interface. It also had benefits for the agent deployment and integration processes, a parallel consideration.
<b>2022-06-14</b>	<b>Fix issues arising from testing</b>
ML Service	We addressed a number of minor issues that had become apparent during testing, such as fit errors under certain conditions.

## S1 Design and Development Log

---

### **2023-02-21 Add item summarisation based on OpenAI/ChatGPT**

*AwAg  
Augment*

This was the first instance of being able to exploit the revolution in Large Language Models such as ChatGPT. The agent up to this point had used a very simple technique for producing a shortened summarisation of content items for display: it would truncate long items, as well as removing any HTML content. We were able to enhance the augmentation processing, swapping out the original Summarise augmente with a new one that passed the CI data to the OpenAI API and asked for a summary, which it then included as the summary augmentation (in the case of a failure, we used the original technique as a backstop). This greatly improved the utility of the presented CIs but importantly it also vindicated the design approach we had taken: we demonstrated that the summary augmentation could be seamlessly swapped out, taking advantage of a powerful commodity ML service that had not been available to us at project inception.

### **2023-02-28 Separate Slack Acquire and Interact functions**

*AwAg  
Acquire &  
Interact*

The support for using Slack had grown somewhat organically, resulting in an implementation that did not have a clear demarcation between using Slack as an Acquire and as an Interact source. As a consequence of the Acquire formalisation work that we had been able to do as a result of switching to the Events API, we were able to complete a separation of these logical functions.

## S1 Design and Development Log

---

<b>2023-03-03</b>	<b>Support for Slack workspace installs</b>
<i>AwAg Platform</i>	Prior to this, the agent had still required manual installation into a workspace <sup>a</sup> , with the administrator manually creating and assigning tokens within Slack and then adding these to the agent configuration. This was obviously not a sustainable approach, and it limited both agent and user autonomy. We added support for the Slack OAuth 2.0 authentication and installation flow <sup>c</sup> , exposing additional services for Slack to make related calls to the agent, and the internal processes to obtain and store the API tokens issued in this process. Our work to logically separate the Acquire and Interact components of the Slack implementation meant that we were also then able to install the Acquire Listener and the Interact service independently – to different workspaces as required. This was an important feature, because it meant that a Listener could be installed to multiple Slack workspaces with limited read-only authorisation scopes, while only the Interact service needed greater permissions. The design concept for this was that the listener should be installed to workspaces that the agent's owner was a member of (such as company workspaces) without the need for elevated privileges purely to acquire content, while the Interact instance would be installed to a workspace that the owner controlled, dedicated just to the Awareness Agent interaction process. This addressed various potential security and deployability concerns that had we had identified during the course of prototyping.

<sup>a</sup><https://slack.com/help/articles/217626298-getting-started-for-workspace-creators>

<sup>b</sup><https://perma.cc/A7AH-R8PW>

<sup>c</sup><https://api.slack.com/authentication/oauth-v2> [<https://perma.cc/ER4A-H5PS>]

<b>2023-03-10</b>	<b>Add an applicationIdentifier/contextIdentifier</b>
<i>AwAg Platform</i>	While working on the process of setting up a study to test the Awareness Agent we realised that there were a number of technical points where there was a conflict if multiple agents shared the same Java JVM (i.e. deployed to the same application server). In particular the use of singletons for internal queues and other items. We implemented a named application identifier to be used internally to address this.

<b>2023-03-16</b>	<b>Add support for multiple agents</b>
<i>ML Service</i>	To better support the study process, we added support for multiple agents to the back end. The initial test version had not used any form of client identification, which was obviously not suitable for scaling up.

## S1 Design and Development Log

---

**2023-03-27**

*AwAg  
Acquire*

### **Initial work to add RSS as an Acquire source**

We added RSS as a second type of Acquire source, with a Client type acquire service configured to pull one or more RSS feeds on a scheduled basis<sup>a</sup>. This was a validation of the approach that we had designed for Acquire Services and Content Items generally, as it showed that we were able to relatively seamlessly add in a different type of source that not only had different content but also a different mechanism for acquiring content. Iterative testing of this required a few minor changes to be made – mainly in relation to the particular content properties of RSS items – but importantly we found these to be non-breaking changes for the existing Slack Acquire implementation, and we were able to seamlessly use the Slack Interact implementation to output RSS-sourced items.

<sup>a</sup>Using Quartz Scheduler in cron-type mode: <https://www.quartz-scheduler.net/documentation/quartz-3.x/tutorial/crontriggers.html> [<https://perma.cc/EP6D-R2RZ>]

**2023-04-13**

*AwAg  
Platform*

### **Refactor platform.database to platform.data**

This was a minor refactoring operation, but reflected a significant design development. Previously `platform.database` had housed only the CouchDB client, but we had now decided to start introducing a general concept of a data layer to the agent platform, supported by a new back end service, “awagdata”. We would then go on to use `awagdata` to support many tasks related to the smooth operation of the agents (configuration storage, object persistence, logging) and the study implementation (content item capture, recording interactions with ML services for repeatability and so on).

**2023-04-14**

*ML Service*

### **Better handling of empty data & error logging**

Testing of the model lifecycle – in particular with models in early stage of training – had exposed a few errors at the back end with how data was loaded and fitted.

**2023-04-14**

*Data Service*

### **Add data service for recording summaries**

As part of our testing process we had identified that summary quality was inconsistent, with some items being better summarised by OpenAI than others. To gather more information on this, we decided to add the facility to record the input and output of the summarisation process, so that issues could be reproduced and re-tested after the fact. To do this, we added a new service to the back end, `awagdata`, initially supporting only this functionality, with data persisted in a local SQLite database. We would later go on to expand this data platform for many other uses.

## S1 Design and Development Log

---

**2023-04-16**

*AwAg  
Interact*

### **Implement Summarisation Feedback**

Testing with users showed that summarisation quality was not always perfect; in some cases – particularly where the source content was short – the ChatGPT model would summarise badly or even fabricate summary text<sup>a</sup>. We added a feature to the UI to allow the user to give feedback on the quality of the summarisation; this would capture the item in question, its content and the user's rating of the quality of the summary text. The captured information was then stored using the awagdata service for later use.

---

<sup>a</sup>See also Emsley [2023]

**2023-04-16**

*Data Service*

**Add services to record and retrieve summarisation feedback** We expanded the data service to allow feedback from the end user on summary quality to be recorded and retrieved, by adding services `record_summarisation_feedback` and `get_summarisation_feedback`.

**2023-04-21**

*Data Service*

### **Add services to add and supply “dummy text”**

Our first look at synthetic data was the enhancement to awagdata to allow what was initially referred to as “dummy text” to be stored & retrieved using services `add_dummy_text` and `get_dummy_text`. This text was simulated message content (to be generated elsewhere) that could be served sequentially one item at a time, grouped by a notional “topic”. The design idea was that a future simulation service could periodically call this and get a series of pre-canned simulated “messages” for a given topic, channel or discussion. By pre-generating the content and storing in the data service this way, we could ensure good performance and reliability, as well as the ability to pre-vet content.

**2023-04-24**

*AwAg Sim*

### **Add `data.awagdata.simulation` & `modules.genus.simulate`**

Our focus had by this point shifted towards supporting the study process, and we started to add support for synthetic data within the Awareness Agent, adding code to publish simulated (synthetic) content to Slack workspaces. This uses the awagdata service at the back end to serve up synthesised content to a set of channels. While this functionality is not strictly part of the Awareness Agent proper – it exists for the purpose of studying the agent – including it in the agent codebase was a pragmatic decision, as it could take advantage of existing modules and functionality.

**2023-05-05**

*Data Service*

### **Added combined dummy text generation**

We added the ability to generate simulated “dummy” texts within the data service itself so that this process could be automated, by adding the service `generate_and_add_dummy_text`.

## S1 Design and Development Log

---

**2023-05-10**

*AwAg  
Augment*

### **Make summarisation less specific to news articles**

The original GPT prompting used to generate CI summary augmentations was worded in such a way that it was more suited to summarising news articles. We modified this to make this more generic. We should also note that this was not a code change; prompt wording is picked up from text template files that can be more easily modified.

**2023-06-12**

*AwAg Eval*

### **Add ChatGPT/OpenAI evaluation implementation**

This was the second major addition related to the study, supporting synthetic evaluation. Again this is not a core feature of the Awareness Agent design but is included for pragmatic reasons. The new feature passes the contents of augmented Content Items to the OpenAI API, asking it to evaluate the quality of the augmentation classification - in effect OpenAI marking awagml's work. This was integrated into the agent's CI flow, with copies of CI's being siphoned off at the Allocate stage to be evaluated. The awagdata platform was used to store the evaluations.

**2023-06-14**

*AwAg Eval*

### **Attempt to coerce OpenAI to only output the requested JSON**

The OpenAI API proved slightly unreliable; our design for evaluation was to ask OpenAI to output its evaluation as a JSON document conforming to a supplied schema. It would usually do so, but sometimes would respond in plain text or using an incorrect JSON schema. We had several iterations working on OpenAI prompts to make this more reliable.

**2023-06-15**

*Data Service*

### **Add evaluation data recording**

In response to initial testing of the Evaluate functionality in the agent, we added the back-end facility to record the evaluation data via a new service, `record_evaluation_data`. This records the details of what was evaluated, how (OpenAI settings) and what the response was. The intent was to facilitate the recording of this information to support both debugging and the study.

**2023-06-23**

*AwAg Eval*

### **Add support for Perspectives**

We introduced a concept called Evaluation Perspective, which was an attempt to add another dimension to how we view and run evaluations. The concept was that a perspective was a second way of looking at the evaluation, by asking a slightly different question of OpenAI. For example "has this item been classified correctly?" is one perspective, but we could in theory ask slightly different questions. The new perspective data structure supported the automatic flow of this, with different perspectives being defined in the agent configuration and then automatically passed to OpenAI. However, we found success was limited, mainly because we were not able to devise a convincing alternative question to the default.

## S1 Design and Development Log

---

<b>2023-06-25</b>	<b>Add Flow Monitor</b>
<i>Data Service</i>	Our service-queue design for the Awareness Agent meant that the flow of Content Items through the system should be clearly defined with known entry and exit points. We had found in testing that some items were in practice not appearing in the Interact UI when we had expected to see them. To help address these issues, and to provide general visibility on the internal workings of the agent, we added the Flow Monitor, recording each contact of a CI with a service or queue in awagdata. This meant that we could track the progress of individual items or aggregate flows and identify bottlenecks and other issues.
<b>2023-06-26</b>	<b>Add Available Classifications to enable better evaluation</b>
<i>AwAg Eval</i>	Previously, our evaluation requests to OpenAI did not list the available options that awagml had chosen from when it had made its classification; to improve the quality of evaluation we added this information to the evaluation request. This was relatively easy to do technically, as our design for the Single Classification Augmentation Item already included this information.
<b>2023-06-28</b>	<b>Add promptQueryJson to openaiApiCompletionsQueryInfo</b>
<i>AwAg Eval</i>	We had previously started using awagdata to record the evaluation operation and response for each evaluation request. In order to give us better reproducibility we added the raw prompt query JSON sent to OpenAI to the recorded data.
<b>2023-06-29</b>	<b>Add state tracking &amp; ability to disable Simulate &amp; Acquire</b>
<i>AwAg Platform</i>	In order to facilitate efficient testing we added the ability to selectively enable and disable services via the UI; we had found that the ‘always on’ nature of the services resulted (ironically) in information overload with items being generated and processed when we might want to focus on another aspect of testing. This feature allowed as to easily control this.
<b>2023-07-04</b>	<b>Add support for batching of evaluations</b>
<i>Data Service</i>	We added back end data support to facilitate a process of running evaluations in batch that we anticipated adding to the core agent code.

## S1 Design and Development Log

---

**2023-07-09**

*AwAg Eval*

### **Add batch execution of OpenAI evaluation queries**

Our very first implementation had fired off a single evaluation request per Classification per Content Item per Perspective. This was quickly found to be very expensive in terms of token usage, as each request required significant prompting text, and adding more classifications to evaluate grew the demands substantially. We changed to a structure early on that would treat each CI as a single request, processing the evaluations for each classification. However, this was still not entirely efficient, so we added a facility for items to be combined and batched. That is, a single prompt request would include the common instructions and a certain number of combined CI's to evaluate. The idea was to reduce the amount of common text per item processed. However, this was at the cost of larger individual queries and higher complexity and we hit token limits with the GPT-3.5 models available to us at the time<sup>a</sup>. We had mixed results with this, finding that combining too many items caused failures due to a number of factors - such as token per request limits being hit, and also a higher chance of the request receiving a malformed response. We found that 3 items per request was a reasonably safe number.

---

<sup>a</sup>The gpt-3.5-turbo-1106 model supported a maximum context window of 16,385 tokens, which a batched request of over 5 items could easily hit, while the more capable gpt-4 had not initially been available for API use

**2023-07-11**

*Data Service*

### **Add evaluation failure reporting web services**

We added services to capture evaluation failure data to awagdata in order to support parallel work in the agent platform.

**2023-07-11**

*AwAg Eval*

### **Add evaluation failure recording**

We had added a specific mechanism for recording evaluation failures to awagdata, which we now incorporated into the Evaluation engine. Failures may be caused by an invalid response from OpenAI, network issues, running out of OpenAI quota for example. By testing for and recording which evaluations failed we gained the ability to re-run or debug them in a targetted manner.

## S1 Design and Development Log

---

- 2023-07-25      Add support for newer /chat/completions API for Evaluation**  
AwAg Eval      At this time, OpenAI moved its current API from “Completions” to “Chat Completions”<sup>a</sup>, a change made since we had started work on OpenAI evaluations. This was a significant change for us, and actually very helpful. One change introduced as part of it was to formally support JSON as a response format - we had previously requested a JSON response in the prompt text, with imperfect results. With the new API we could specify a defined response schema, which we could then parse. The changes made to the request side were also significant; we could move from constructing a single text prompt to using a more structured array of message JSON. This helped to a degree with our process of passing a mix of JSON and text in the evaluation request, allowing more structure. However, we noted it was still short of a full JSON request format where the API would be sent a schema-based request rather than being told this on each occasion. The change to an array of system/user messages in the OpenAI request also opened a design possibility to us, as it made it easier to make the request process more modular in terms of request construction and variation. We would go on to use this later.

---

<sup>a</sup><https://help.openai.com/en/articles/7042661-moving-from-completions-to-chat-completions-in-the-openai-api> [<https://perma.cc/KBZ7-MGPP>]

- 2023-09-07      Add support for tags and for recording/fetching raw evaluation items**  
Data Service      We made a number of additions to awagdata to support new evaluation-related functionality in the agent platform, by allowing data items to be stored associated with text tags, and adding new services to store and retrieve raw evaluation items (the CI data required to perform an evaluation operation).

## S1 Design and Development Log

---

<b>2023-09-07</b> <i>AwAg Platform</i>	<b>Add support for recording tags (in evaluation etc.)</b> Tag support was a major study-focussed change that emerged from the first test study iterations. The main driver for this was data management and organisation of tests. We needed to support concepts such as “test run 1” or “evaluation run 1” so that we could identify which data (content items, evaluations etc.) were associated with each run. This would enable us to clearly define the start and end of different study phases, and also allow us to run multiple different evaluations against the same content and identify different runs. We did this by adding the concept of tags. Each write of an item to awagdata would be accompanied by one or more tags, plain text strings. Use source of the tags varied: we could set a “current tag” (or multiple current tags) in the Acquire service via the UI; each incoming CI would then be tagged with these. Similarly for evaluations we added support for input tags (evaluate only those CIs that match the supplied tag) and output tags (i.e. store the evaluation output in awagdata with these tags attached). This allowed us a great deal of control and precision in the evaluation and general study administration process.
<b>2023-09-12</b> <i>AwAg Eval</i>	<b>Add on-demand processing for evaluations based on recorded evaluation items</b> Adding tag support had given us the ability to change our approach to evaluation processing. Our previous design for evaluation was to incorporate this step into the CI flow; however we had found this was not optimal. Sometimes we were not interested in evaluating during some types of tests for example (each evaluation execution has both a monetary and performance cost attached to it). We also wanted the ability to run evaluations in batch after changing something, such as the prompting text. To achieve this we added the ability to enable/disable the background Evaluate service as we had previously done for Simulate and Acquire, and also added a Slack command based facility to run evaluations in batch on demand. At the back end we added an awagdata facility to record all Content Items as they passed through identified by the currently active set of tags. We could then run on-demand evaluation batches against a set of CIs defined by the passed tag, and we were also able to identify and distinguish the output of these runs from the added output tags. We actually found that this was our preferred way of running evaluations, as it gave us in our role of study administrator much more control.

## S1 Design and Development Log

---

<b>2023-09-18</b>	<b>Add Evaluate Explore interaction via Slack UI</b>
AwAg Eval	We needed a mechanism to allow the study administrator to examine evaluations interactively and our first iteration of this was based on the Slack Interact UI, adding an Evaluate Explore feature. This could be used to return lists of evaluation results to the user as Slack messages in the Slack UI. Testing showed this was functional but not as easy to use as desired, so we noted that we would need to revisit this area later.
<b>2023-09-22</b>	<b>Add new object store to replace Cloudant/Bluemix</b>
Data Service	We added our self-implemented JSON object store, named SPOSS, to the awagdata layer, allowing us to persist objects locally with good performance, simplicity of access and certain functionality tailored to our specific requirements.
<b>2023-09-23</b>	<b>Remove dependency on IBM Bluemix/Cloudant</b>
AwAg Platform	We had previously been using the free tier of this commercial CouchDB database, but the T&C's changed so that it was no longer free for our level of usage. We stopped using it and transitioned to our own SPOSS object store service.
<b>2023-10-02</b>	<b>Add awagUi implemented in Angular</b>
AwAg UI	We had not been satisfied with the first attempt of evaluation exploration so we make the decision to design a dedicated Angular based UI for the study administrator to interact with evaluation results. We also added the ability for the user (study participant) to submit feedback data via this UI on the quality of the evaluations. This has awagdata as the back end, which required some additional services to be added for access to data for reporting and interaction in this way. We named this new web based interface "awagUi".
<b>2023-10-04</b>	<b>Allow additional tags to be added when processing evaluations</b>
AwAg Eval	The original approach to on-demand evaluation processing added only a single tag to evaluations from each run. Study testing showed us that we could in fact use more tags to give finer grained control over evaluation outputs, so we added this feature.
<b>2023-10-04</b>	<b>Add mode1/mode2 support to awagUi evaluation explorer</b>
AwAg UI	Initial testing of the evaluation feedback process with the pilot study participants showed some issues with the design of the UI; some concepts were not always clear and in some cases there was clutter in the UI that the user did not want to see. We approached this by designing two presentation modes, one with stripped down explanatory information and a differently organised layout. The user had the choice of switching between these.

## S1 Design and Development Log

---

**2023-10-13**

*ML Service*

### **Add models subdirectory for agents**

We implemented segregation at the back end for data from different agents, so that we could more easily clean down and reset in testing.

**2023-10-15**

*ML Service*

### **Add support for model description**

We added the ability to store/return a descriptive text for each model so that this information could be used by the agent evaluation process. We opted to do this in the back end because the description handling fits well into the model lifecycle.

**2023-10-15**

*AwAg Eval*

### **Add support for model description**

Testing of the quality of evaluations showed cases where the OpenAI model was making clear mistakes about the meaning of classifications in the model that it was asked to evaluate. We addressed this by passing a model description text along with the classification request to awagml and incorporating this in the prompting for the request - which produced noticeably better results. We then added support for this throughout the agent system, providing a mechanism for the user to add and maintain descriptions for models.

**2023-11-19**

*Data Service*

### **Enhancements to data simulation**

We made a number of changes to the data simulation back end, in particular item generation via OpenAI, which we updated to use the new /chat/completions API that we had previously adopted for evaluations. This gave us access to better OpenAI models, with improved quality of generated items.

**2023-11-22**

*AwAg Eval*

### **Add recording of classification actions**

We had found during the pilot study phase that it would be helpful to maintain a record of the actions of the user to classify items – that is, where the user used one of the available UI mechanisms to perform a training event related to a content item (changing classification value or confirm correctness of a classification for example). We had not previously been capturing a log of these actions when they were performed, but realised that having this data would provide important experimental insight. A classification performed by the user is a de facto evaluation of the original classification by the human, mirroring the evaluation by the OpenAI-based evaluator. This would provide us with an additional mechanism for getting feedback data on the quality of the AI evaluation by comparing the agreement with the human and AI evaluations.

## S1 Design and Development Log

---

<b>2023-11-26</b> <i>AwAg</i> <i>Platform</i>	<b>Add/use non-volatile JSON Object Store to facilitate classification feedback</b>  We had previously found that we needed to implement a workaround for limitations in the use of Slack for giving evaluation feedback related to the passing of state/context information. There were limits in the Slack API on how much data could be included in drop-down menu items, and there was a lack of ability to store arbitrary context data in items posted to Slack. We had addressed this by storing a data object – “Classification Feedback Item” in memory server-side and then including a reference ID to this in the posted item. Thus when the user made a change via UI action the submitted ID was used to look up the context information. However our original implementation for this used volatile memory to store these items - the flaw with this was that the published Slack items were not volatile, so there was the possibility of the user accessing these after the reference had been lost from volatile memory due to a server restart for example. We addressed this by instead persisting these objects to the SPOSS object store.
<b>2023-11-28</b> <i>AwAg UI</i>	<b>Change AGREE to CONCUR in awagUi</b>  The pilot study user experienced some difficulty with the process of giving evaluations, as sometimes the wording was not clear to them. One case was when they were being asked whether they agreed with the evaluation of a particular classification. This is not the same as agreeing with the original classification (for example if the classification is incorrect then the human and AI would agree with each other that they disagree with the classification). We addressed this to an extent by using the word “Concur” instead.

## S1 Design and Development Log

---

**2023-11-30**

AwAg UI

### Add training item display to awagUi

The pilot study exposed some limitations with the training process for the user-defined models when used in a bulk context (such as bootstrapping a new model or generating statistically valid amounts of data). When training within Slack we found that the user followed a process that was straightforward but inefficient and limited by UI speed. The user would view an item within its channel in Slack, then click on the classification drop-down, then select a different classification (to change classification) or re-select the same classification (to generate positive training). If reclassified, then they would need to wait for the back end service to process the action and move the item from its old to new channel before proceeding with the next one. If classifying a single item then this was sufficiently performant, but in the study the user would be asked to apply one classification per item per model – so if the user had 5 User-Directed ML models set up they would potentially need to repeat this process 5 times. It was too time consuming. We decided that significant action was needed to address this, and added a new training UI to awagUi. This UI was much faster and more responsive than going via Slack, and also let the user select corrected training classifications on every model defined for the CI in a single operation, significantly reducing the number of clicks and individual processing delays. This new UI would also then go on to be a basis for a secondary Interact UI.

**2023-12-08**

Data Service

### OpenAI file handling

We had made a design decision based on our experiences so far to add the ability to train or fine-tune OpenAI models used for evaluation<sup>a</sup>, so that we could compare the performance of trained vs untrained models. To support this we added a mechanism to generate, manage and upload training data files to OpenAI, and then to use these files to train models. To make the process manageable, we chose to handle the whole process within the awagdata layer. This includes generating training objects from recorded classification actions, storing these within our object store, using them to generate OpenAI files and managing the querying, execution and deletion of models on the OpenAI side. We took this approach because OpenAI provided only bare API support for these operations with no higher management ability; we had found that it was too easy to lose track of the various models and files on the OpenAI system, and not always easy to tell the provenance of these files. Our system tracks history, source and other information related to all artefacts that we create in OpenAI for easy and efficient management.

---

<sup>a</sup><https://platform.openai.com/docs/guides/fine-tuning> [<https://perma.cc/CKJ2-LVVK>]

## S1 Design and Development Log

---

**2024-02-18**

*AwAg Eval*

### **Add Evaluation support to awagdata**

Our original design for Evaluation was to have the processing done within the Java Awareness Agent application at time of CI generation; we later added the ability to batch this and run on-demand evaluations. During the pilot study we found that we had moved more and more towards on-demand tag based evaluation processing and rarely used the original mode. Because of this, we decided to replicate the evaluation functionality in the back end, so that it could be run entirely independently of the agent itself. This was consistent with the agent design concept, which does not include evaluation as a core function, and was also compatible with how we ran offline evaluations: all the data used to run these was already stored in the awagdata layer, and the outputs were also recorded there. Changing our evaluation engine to the awagdata platform meant that we could more easily run and manage evaluation jobs via simple web service calls. We also retained the code in the Java Awareness Agent, because the structured nature of the evaluation request object and the standardised prompt structure meant that all we needed to do was ensure that a few resources were kept synchronised between the two, such as schemas and prompt templates.

**2024-02-18**

*Data Service*

### **Add Fixit route**

We had discovered during the pilot study run that a minor technical error had prevented the proper recording of evaluation items during the course of the study – they should have been recorded to the Data Service as they came in, but this was not happening. This would have been catastrophic for the study as it would prevent any evaluation being run on the data, which could not easily be re-captured. However, the necessary data to construct evaluation items after was in fact being recorded elsewhere within awagdata, so we were able to write code to extract this information and retrospectively construct the necessary evaluation items.

**2024-02-19**

*AwAg Eval*

### **Improve OpenAI prompting**

In response to preliminary results in the pilot study we made several changes to the prompt text passed to OpenAI for evaluation requests. We found that we needed to be much more explicit about some of the things that we wanted the evaluator to do, and also needed to repeat some points in multiple ways to reinforce the message.

## S1 Design and Development Log

---

**2024-02-20**

*AwAg Eval*

### **Support Lightweight Mode option for evaluation**

We had found that our OpenAI evaluation requests had been growing increasingly large, having an impact on both query performance and token usage. To address this we added a “lightweight” mode, where certain content could be omitted by setting a flag. For example, the schema describing the content item is not passed in this mode. The intent was to use this in conjunction with tags to run evaluations in each mode for comparison, while also using different OpenAI models at the back end. Part of this approach was also to use pre-trained (fine-tuned) OpenAI models, particularly in conjunction with lightweight mode, and compare results across models. The previous decision to introduce tags to the system made this practical.

**2024-02-23**

*Data Service*

### **Add standard vs split outputs**

We found that many of the parts of prompts that we were sending to OpenAI had commonality for different functions (evaluations, training item generation etc.), so reorganised the back end code to approach this more efficiently and remove redundancies.

**2024-02-23**

*AwAg UI*

### **Add chat component to awagUi**

To facilitate interactive testing, we added a simple chat UI so that the user<sup>a</sup> could enter some text which would be executed against the OpenAI model; this is functionality already provided by the ChatGPT UI<sup>b</sup>, but that UI did not provide access to trained models, only to vanilla ones. By adding our own UI we could run arbitrary requests against our own trained models, also utilising our own system request prompt elements. This allowed us to more efficiently test model and prompt changes in an iterative fashion.

---

<sup>a</sup>In this case the user is the researcher rather than a study participant

<sup>b</sup><https://chatgpt.com/>

## S1 Design and Development Log

---

**2024-02-23**

*Data Service*

### Add subsets support

We had observed during the pilot study that the volume of content was too great for the participant to process all of it – in itself this was not unexpected and not a problem. However, we found that we would get more useful study results if the user acted on the same subset of items in each phase (initial classification and evaluation feedback), but it was important also to get a good distribution of items and not just have the user process in date order. To support this we added support for subsets to awagdata. We did this by adding the ability to add a set of subsets for the data, where we could define random subsets<sup>a</sup> of the recorded Content Items identified by tag and a subset percent label. So for example we might generate a subset of 25% of the items for tagA, which would be identified by the path /subset/tagA/25. The user-facing UIs could then be filtered to show only items in that pseudo-random subset.

---

<sup>a</sup>Using Python `random.sample()` to generate a pseudo-random sample

**2024-03-06**

*AwAg Eval*

### Edit prompt content

Previous changes had improved evaluation quality, but we had found that sometimes OpenAI would not consider all classifications in its evaluation so we changed the prompting to encourage it to do so.

**2024-03-14**

*Data Service*

### Add statistics and improve reporting

To improve the efficiency of result processing during and after then studies, we added a stats service, that would generate a set of statistics based on the evaluations and other data for a given study/tag, outputting as a JSON document containing a stats “package”. Having such a package automatically generated would greatly improve the ability to generate updated statistics for each study instance.

**2024-03-22**

*AwAg Eval*

### Address occasional incorrect property names in OpenAI response

We had found that sometimes OpenAI would return JSON with incorrect property names, such as switching from camelCase to using\_underscores<sup>a</sup>. We implemented a workaround to check for and correct this. However we also noted that we experienced this more often when using Juneau’s `ObjectMap.toString()` in Java to serialise items to JSON (which generated JSON that did not have quotes around property names). Changing the serialisation code to ensure that property names were always enclosed in quotes largely resolved this issue.

---

<sup>a</sup><https://news.ycombinator.com/item?id=34525139> [<https://perma.cc/4F7Y-CUSD>]

## S1 Design and Development Log

---

<b>2024-04-04</b> <i>AwAg UI</i>	<b>Support for text-likert mismatch via partial feedback</b> We had found in some cases that the text description generated by OpenAI for its feedback did not agree with the Likert value that it had assigned (typically in this case they would be opposite to each other). This was a significant enough problem that we added an element to the evaluation feedback UI to allow the user to flag this easily, allowing us to capture data on the number of affected items.
<b>2024-05-02</b> <i>Data Service</i>	<b>Add agree/disagree exclusion</b> To improve the quality of data used to train models, we added a feature to allow the list of user classification actions to be used for training to be filtered to exclude those where the user agreed with the classification or those where the user had disagreed. We had found that the user was agreeing with the classification chosen by the trained models in the majority of cases (often a very large majority) so the training data sets had been very skewed towards entries where there was agreement. The training data would be more effective with a better balance of content [Dube and Verster, 2023], so that the model would be trained with a similar number of disagreements to agreements. Adding these filters meant that we could build up our evaluations in multiple steps, assembling a training set with a higher proportion of disagree actions than a single unfiltered pass would generate (i.e. oversampling 'Disagree').
<b>2024-05-02</b> <i>Data Service</i>	<b>Add dataset merging</b> To support the process of building up training datasets with varied compositions and using these to fine-tune models, we added a feature to allow the administrator to merge datasets – so for example two datasets that had been generated with different ratios of agree/disagree could be combined into one and then used to train an OpenAI model. This allowed us more control over the composition of fine-tuning datasets.
<b>2024-05-08</b> <i>Data Service</i>	<b>Work on evaluation job processing</b> Our evaluation processor was designed to process items that match a given tag and subset, that did not have a corresponding evaluation already recorded. If an error was encountered during evaluation (for example due to an OpenAI error) then no evaluation would be recorded – so when the evaluation was next run, the failed item would be retried. Due to a small percentage of failures in each run, we found that we needed to run a number of passes before all items were evaluated. We added code to make this process more manageable and transparent.
<b>2024-05-23</b> <i>Data Service</i>	<b>Add tabular statistics, broken down by classification</b> We improved the structure and content of the statistics output to better support the study analysis.

## S1 Design and Development Log

---

**2024-05-27**

*Data Service*

### **Excel stats pack export**

We added the ability to automatically output a statistics package to Excel format for easy analysis, using pandas to output dataframes to Excel worksheets<sup>a</sup> and combine to multi-sheet formatted workbooks.

<sup>a</sup>[https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to\\_excel.html](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_excel.html)

<sup>b</sup><https://perma.cc/AN99-MZMP>

## Bibliography

[Dube and Verster, 2023] Dube, Lindani and Tanja Verster (2023). “Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models”. In: *Data Science in Finance and Economics* 3.4, pp. 354–379. issn: 2769-2140. <http://www.aimspress.com/article/doi/10.3934/DSFE.2023021>.

[Emsley, 2023] Emsley, Robin (2023). “ChatGPT: these are not hallucinations – they’re fabrications and falsifications”. In: *Schizophrenia* 9.1, pp. 4–5. issn: 27546993. doi: [10.1038/s41537-023-00379-4](https://doi.org/10.1038/s41537-023-00379-4).

[Pedregosa et al., 2011] Pedregosa, F et al. (2011). “Scikit-learn: Machine Learning in {P}ython”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

## **Supplement S2**

### **Survey Questions**

## S2 Survey Questions

---

# Survey Questions

Table S2.1 contains a list of all questions from the survey in Chapter 5, along with details of the question type.

Table S2.1: Survey Questions

#	Question Text	Details
1	For what reasons do you interact with other people online?	As part of your job Conducting research or studying For leisure To keep in touch with family/friends Staying informed
<b>2. How do you prefer to receive information when it is something you might need to act on quickly? Please rate on a scale of 1-5, with 1 meaning you dislike that method and 5 meaning you actively prefer that method.</b>		
2.1	Email	Likert scale (1-5)
2.2	Text message or SMS	Likert scale (1-5)
2.3	Using an application or website on a smartphone	Likert scale (1-5)
2.4	Receiving a notification on a smartphone	Likert scale (1-5)
2.5	Using an application or website on a desktop computer	Likert scale (1-5)
2.6	Receiving a notification on a desktop computer	Likert scale (1-5)
2.7	Phone call	Likert scale (1-5)
2.8	Voicemail message	Likert scale (1-5)
2.9	In person	Likert scale (1-5)
2.10	Letter	Likert scale (1-5)
<b>3. How do you prefer to receive information when it is something that might be important but does not need immediate action? Please rate on a scale of 1-5, with 1 meaning you dislike that method and 5 meaning you actively prefer that method.</b>		
3.1	Email	Likert scale (1-5)
3.2	Text message or SMS	Likert scale (1-5)
3.3	Using an application or website on a smartphone	Likert scale (1-5)
3.4	Receiving a notification on a smartphone	Likert scale (1-5)
3.5	Using an application or website on a desktop computer	Likert scale (1-5)
3.6	Receiving a notification on a desktop computer	Likert scale (1-5)
3.7	Phone call	Likert scale (1-5)
3.8	Voicemail message	Likert scale (1-5)
3.9	In person	Likert scale (1-5)

Continued on next page

**S2 Survey Questions****Table S2.1 – continued from previous page**

#	Question Text	Details
3.10	Letter	Likert scale (1-5)
<b>4. How do you prefer to receive information when it is interesting but not particularly important or urgent? Please rate on a scale of 1-5, with 1 meaning you dislike that method and 5 meaning you actively prefer that method.</b>		
4.1	Email	Likert scale (1-5)
4.2	Text message or SMS	Likert scale (1-5)
4.3	Using an application or website on a smartphone	Likert scale (1-5)
4.4	Receiving a notification on a smartphone	Likert scale (1-5)
4.5	Using an application or website on a desktop computer	Likert scale (1-5)
4.6	Receiving a notification on a desktop computer	Likert scale (1-5)
4.7	Phone call	Likert scale (1-5)
4.8	Voicemail message	Likert scale (1-5)
4.8	In person	Likert scale (1-5)
4.8	Letter	Likert scale (1-5)
<b>5. How often do you use the following online services or applications?</b> <b>Choose from: Never, Every month or less often, Every week, Daily, Many times a day</b>		
5.1	Twitter	Single choice
5.2	Facebook	Single choice
5.3	Reddit	Single choice
5.4	Instagram	Single choice
5.5	Snapchat	Single choice
5.6	Instant messaging	Single choice
5.7	Slack	Single choice
5.8	Skype	Single choice
5.9	Business collaboration software	Single choice
6	Which, if any, of the following applications for messaging/collaboration do you use to conduct work on?	Microsoft Skype or Lync IBM Sametime Slack IBM Connections Microsoft Sharepoint Google G Suite (Mail, Docs, Hangouts) WhatsApp Evernote Electronic Laboratory Notebook
Continued on next page		

## S2 Survey Questions

**Table S2.1 – continued from previous page**

#	Question Text	Details
7	Which, if any, of the following applications for messaging/collaboration do you use in a personal or social capacity?	Microsoft Skype or Lync IBM Sametime Slack IBM Connections Microsoft Sharepoint Google G Suite (Mail, Docs, Hangouts) WhatsApp Evernote Electronic Laboratory Notebook
8	Please let us know about any applications or services that you use very frequently that are not listed here.	Free text
9	For this question, please think about your attitude towards online services. Please rate on a scale of 1-5, with 1 meaning you dislike that service and 5 meaning you actively use that service.	Likert scale (1-5)
<p><b>9. Please think about your attitude to being interrupted by notifications from your smartphone or computer originating from online services such as Slack, news apps, Facebook, Twitter, Instagram and messaging applications. Please rate each statement on a scale of 1-5, with 1 meaning you strongly disagree and 5 meaning you strongly agree.</b></p>		
9.1	I receive so much information online that I often miss things that are important or time critical	Likert scale (1-5)
9.2	I don't mind being interrupted when it's about something important	Likert scale (1-5)
9.3	I often receive notifications about things that could have waited for later	Likert scale (1-5)
9.4	Getting interrupted by notifications/alerts when I'm trying to get things done is a problem for me	Likert scale (1-5)
9.5	My online services always get it right about what I want to be interrupted with	Likert scale (1-5)
9.6	Online services always get it right when they judge what I'm interested in	Likert scale (1-5)
<p><b>10. We would like to ask you about how well online services such as Slack, Facebook, news apps, Twitter and Instagram understand your preferences. Please rate each statement on a scale of 1-5, with 1 meaning you strongly disagree and 5 meaning you strongly agree.</b></p>		
10.1	Online services always get it right when they judge what I'm interested in	Likert scale (1-5)
10.2	Not all of the things that I follow (hashtags, people) are equally important to me	Likert scale (1-5)
10.3	I want to be able to tell online services what matters to me most	Likert scale (1-5)
10.4	I don't have enough control over what online services choose for me	Likert scale (1-5)
Continued on next page		

## S2 Survey Questions

**Table S2.1 – continued from previous page**

#	Question Text	Details
10.5	I'm happy to have a computer make decisions about what content I should see	Likert scale (1-5)
10.6	The idea of being able to rank or prioritise hashtags and other content appeals to me	Likert scale (1-5)
10.7	I am happy to put in effort to 'train' the online services in order to see better results	Likert scale (1-5)
<b>11. We would like you to tell us about your trust in online services. Please rate each statement on a scale of 1-5, with 1 meaning you strongly disagree and 5 meaning you strongly agree.</b>		
11.1	I trust online services to make the best decisions about what to show me	Likert scale (1-5)
11.2	I'm happy to share information about my interests and activities with online services if it will improve the service	Likert scale (1-5)
11.3	I am uncomfortable about sharing personal information with online services because I don't know what they do with it	Likert scale (1-5)
11.4	I prefer to keep information about my interests & activities under my control	Likert scale (1-5)
11.5	I'm more willing to let a computer program have personal information if I know that I can control it	Likert scale (1-5)
12	If you have time, please use this space to tell us more about your online habits.	Free text
<b>13. Please tell us about some of your general views on smartphone and connected applications. Please rate each statement on a scale of 1-5, with 1 meaning you strongly disagree and 5 meaning you strongly agree.</b>		
13.1	I always take time to customise the applications and devices that I use	Likert scale (1-5)
13.2	I regularly update and ensure I have the latest version of applications I use	Likert scale (1-5)
13.3	I'm always looking to try new applications and services	Likert scale (1-5)
13.4	I consider myself very tech-savvy	Likert scale (1-5)
13.5	Computers are just tools rather than interesting in themselves	Likert scale (1-5)
13.6	I use all of the features on my phone and connected applications	Likert scale (1-5)
14	Do you use connected/online devices or applications for both work and personal purposes?	Yes/No
<b>14.a. Please can you tell us what you think about the relationship between personal and work use of computers and notifications on connected/online devices. Please rate each statement on a scale of 1-5, with 1 meaning you strongly disagree and 5 meaning you strongly agree. Optional for those answering Yes to question 14.</b>		
Continued on next page		

## S2 Survey Questions

**Table S2.1 – continued from previous page**

#	Question Text	Details
14.a.1	I keep my work and personal applications entirely separate	Likert scale (1-5)
14.a.2	I find it easy to switch off from work	Likert scale (1-5)
14.a.3	I'm happy to receive some work-related notifications during personal time	Likert scale (1-5)
14.a.4	I prefer to keep certain applications (such as instant messaging on my phone) for personal things and not for work	Likert scale (1-5)
14.a.5	I'm happy to see personal notifications while I'm at work	Likert scale (1-5)
14.a.6	My employer is happy for me to receive personal notifications while I'm at work	Likert scale (1-5)
14.a.7	I don't mind both work and personal information coming from the same device or application as long as it gets the timing and content right	Likert scale (1-5)
14.b	If you have any more comments on the distinction between personal and work use of devices, please let us know here. <i>Optional for those answering Yes to question 14</i>	Free text
15	How old are you?	Under 18 18 to 24 25 to 34 35 to 44 45 to 54 55 to 64 65 to 74 75 or older I would rather not say
16	What is your gender?	Female Male I would rather not say Other
17	Are you employed (or self-employed, a business owner, or freelancer)?	Yes No I would rather not say
17.a	Do you work full or part time?	Full time Part time
17.b	Which (if any) of the following describe your employment (select any that apply)?	IT professional Administrative/clerical Knowledge worker Academic researcher Business owner/director Medical professional Educational professional Civil servant Armed forces Police or justice professional Sales/retail

Continued on next page

**S2 Survey Questions****Table S2.1 – continued from previous page**

#	Question Text	Details
18	Are you currently a student?	Yes No I would rather not say
18.a	How do you do most of your studying?	On site at an institution Via distance learning
18.b	Which best describes your current educational status?	School/further education student University student (undergraduate) University student (postgraduate)
19	What is the highest level of education that you have completed?	School / further education Undergraduate degree Masters degree Doctorate Not applicable / rather not say
20	Where do you live?	UK Other European country Africa Antarctica Asia Canada Central America Mexico Middle East Oceania South America USA West Indies I would rather not say
21	What are your primary languages (i.e., the languages you use most frequently)?	Multiple answer
21.a	If you selected Other, please specify	Free text
22	Your name	Free text
23	Your email address	Free text
24	Opt-in to follow-up contact	Please check here if you are happy to participate in follow-up research
25	If there is anything else you would like to add, please enter it below	Free text

## **Supplement S3**

### **Survey Demographics**

### S3 Survey Demographics

---

## Survey Demographics

This document contains information on key demographics of respondents to the Chapter 5 survey.

Table S3.1: Survey Questions and References

Question	Description	Reference
Q1	For what reasons do you interact with other people online?	S3.1
Q15	How old are you?	S3.2
Q16	What is your gender?	S3.3
Q17	Are you employed (or self-employed, a business owner/director)?	S3.4
Q17.b	Which (if any) of the following describe your employment (select any that apply)?	S3.5
Q18	Are you currently a student?	S3.6
Q19	What is the highest level of education that you have completed?	S3.7
Q20	Where do you live?	S3.8

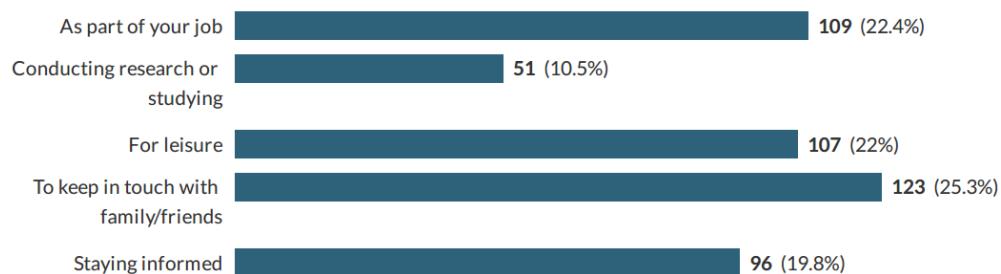


Figure S3.1: Survey Q1: For what reasons do you interact with other people online?

### S3 Survey Demographics

---

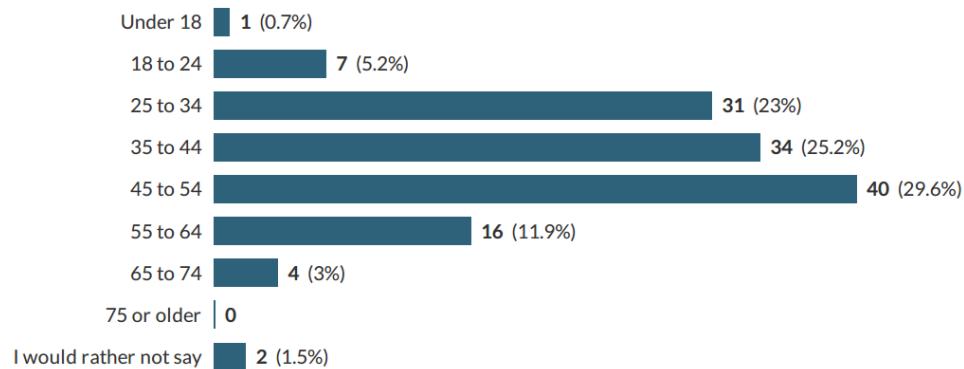


Figure S3.2: Survey Q15: How old are you?

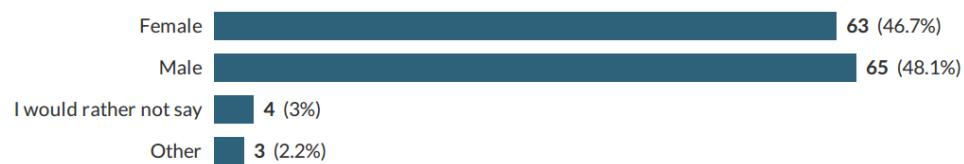


Figure S3.3: Survey Q16: What is your gender?



Figure S3.4: Survey Q17: Are you employed (or self-employed, a business owner/director)?

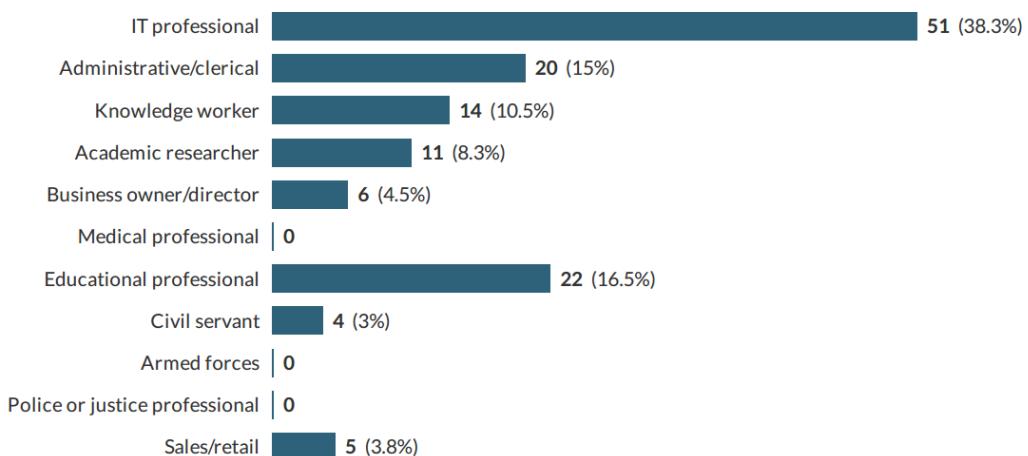


Figure S3.5: Survey Q17.b: Which (if any) of the following describe your employment (select any that apply)?

### S3 Survey Demographics

---

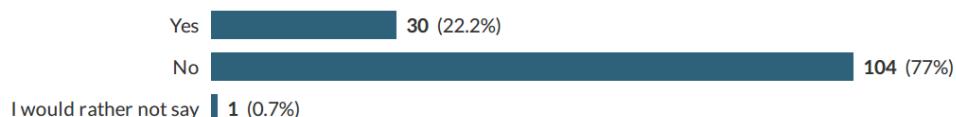


Figure S3.6: Survey Q18: Are you currently a student?

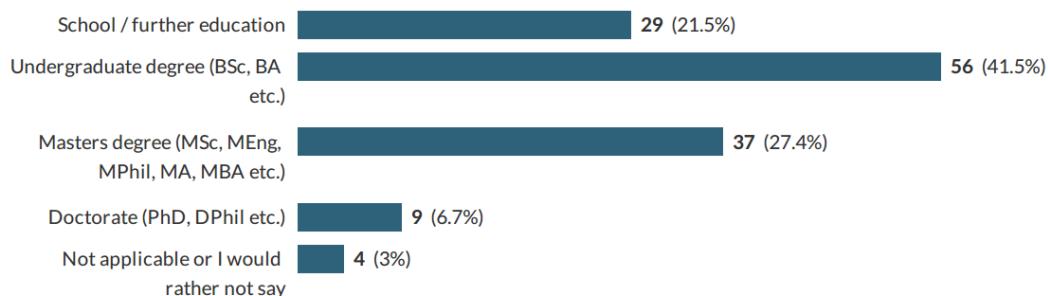


Figure S3.7: Survey Q19: What is the highest level of education that you have completed?

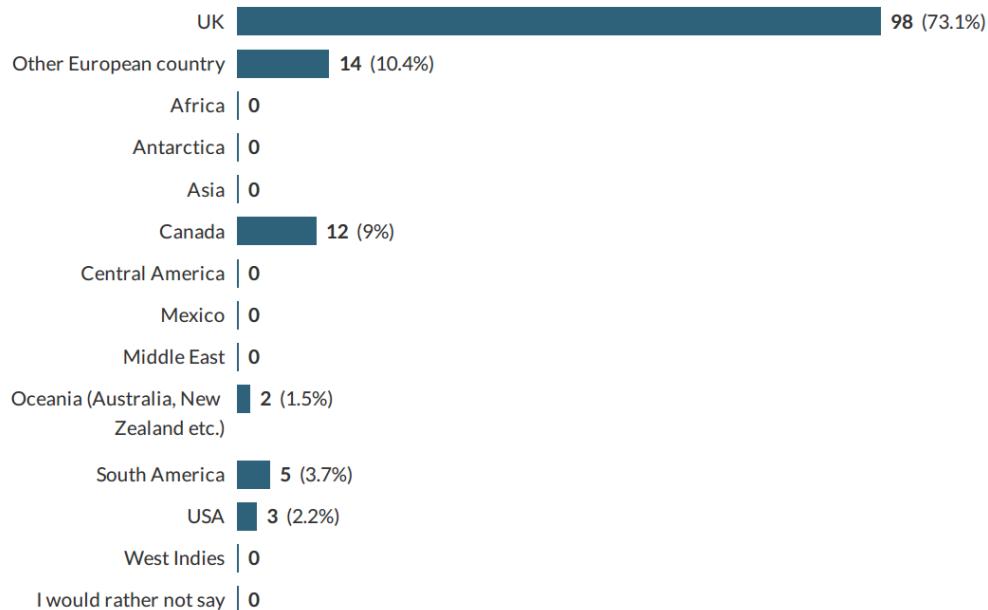


Figure S3.8: Survey Q20: Where do you live?

## **Supplement S4**

### **Personas**

## **Personas**

This document contains the final Personas that were derived using the process described in Chapter 5.

The personas are also available in JSON format at:

[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /personas/personas\_final.json]

- [S4.1 – Susan](#)
- [S4.2 – Adam](#)
- [S4.3 – Phoebe](#)
- [S4.4 – Kenton](#)
- [S4.5 – Usha](#)

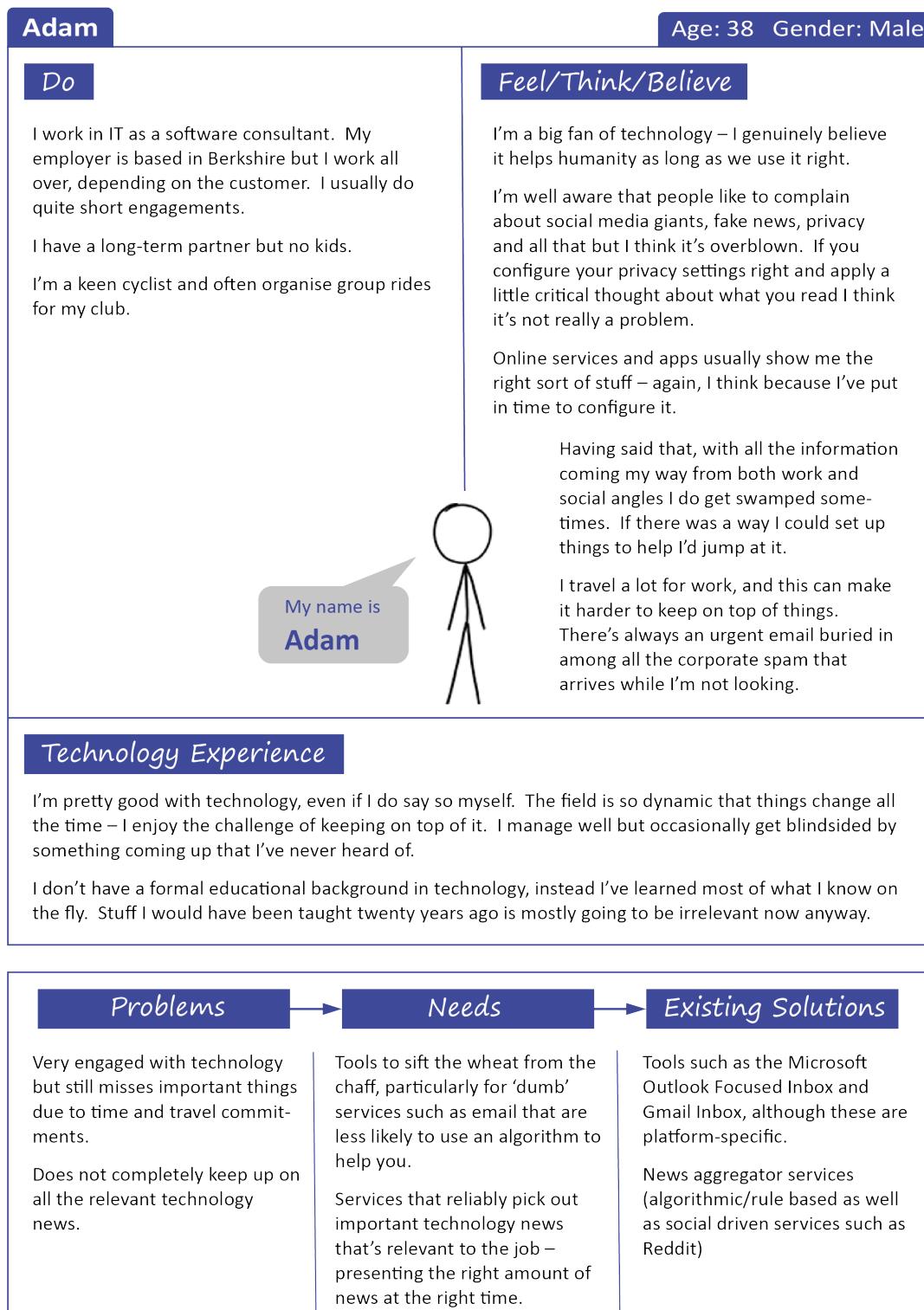
**S4 Personas**

Figure S4.1: Final PATHY Persona “Susan”

<b>Susan</b>								
<p><b>Do</b></p> <p>I live in the Midlands in the UK and work as an administrator at the nearby university.</p> <p>I'm married and have two children – unlike me they both went to university. Since graduating one settled down quite close but the other lives and works in London.</p> <p>I have a few hobbies outside of work – baking, tennis and getting out into the local countryside.</p>	<p><b>Age: 58 Gender: Female</b></p> <p><b>Feel/Think/Believe</b></p> <p>While I enjoy my job, it's not the most important thing in my life; I have no trouble switching off at the end of the day, even when it's been very busy. When I socialise with work colleagues we rarely talk about work (otherwise I probably wouldn't socialise with them).</p> <p>I know my children lead busy lives now, but I miss seeing as much of them as I used to, particularly the youngest who is in London now. They do try and keep in touch but I don't always know what they're up to or how they are doing.</p> <p>I really enjoy my tennis and spend a lot of time helping organise club events.</p> <p>I'm not always entirely on top of things in my personal life because I don't check my email often enough.</p>							
<p><b>Technology Experience</b></p> <p>I use a desktop computer at work for admin, email and maybe a little web browsing. We have one at home too, but it's mostly my husband on that.</p> <p>I've had IT training at work and get along fine with computers – although I prefer to stay in my comfort zone.</p> <p>My son made me get one of those smartphones. I didn't really see the point at first, but it is actually quite useful for staying in touch and organising things. I think I mostly use Facebook and WhatsApp as well as things like the weather app.</p> <p>I admit I do use Facebook quite a lot, but there are a lot of stories about how much they know about you and what they do with that. If it wasn't so handy, I'd use it a lot less.</p>								
<p><b>Problems</b> → <b>Needs</b> → <b>Existing Solutions</b></p> <table border="1"> <thead> <tr> <th>Problems</th> <th>Needs</th> <th>Existing Solutions</th> </tr> </thead> <tbody> <tr> <td>Not as aware of her adult childrens' activities and day to day lives as she would like to be.  Because of relatively low level of engagement with computers at home, Susan sometimes misses items of news or things to act on – particularly when they come in via email or get lost in Facebook feeds.</td> <td>Tools to help her track what is going on with friends, family and hobbies in social media without needing to log in all the time.  A way to ensure that she does not miss important emails.</td> <td>Existing algorithms in social services that select content for users.  Email filtering.  Notification functionality in social media smartphone apps.</td> </tr> </tbody> </table>			Problems	Needs	Existing Solutions	Not as aware of her adult childrens' activities and day to day lives as she would like to be.  Because of relatively low level of engagement with computers at home, Susan sometimes misses items of news or things to act on – particularly when they come in via email or get lost in Facebook feeds.	Tools to help her track what is going on with friends, family and hobbies in social media without needing to log in all the time.  A way to ensure that she does not miss important emails.	Existing algorithms in social services that select content for users.  Email filtering.  Notification functionality in social media smartphone apps.
Problems	Needs	Existing Solutions						
Not as aware of her adult childrens' activities and day to day lives as she would like to be.  Because of relatively low level of engagement with computers at home, Susan sometimes misses items of news or things to act on – particularly when they come in via email or get lost in Facebook feeds.	Tools to help her track what is going on with friends, family and hobbies in social media without needing to log in all the time.  A way to ensure that she does not miss important emails.	Existing algorithms in social services that select content for users.  Email filtering.  Notification functionality in social media smartphone apps.						

**S4 Personas**

Figure S4.2: Final PATHY Persona “Adam”



**S4 Personas**

Figure S4.3: Final PATHY Persona "Phoebe"

<b>Phoebe</b>		<b>Age: 23 Gender: Female</b>						
<p><b>Do</b></p> <p>I'm a recent graduate working for a accountancy firm in London. I'm on the management fast track, meaning that I move around many different parts of the business to learn how they work.</p> <p>I'm currently single, mainly because I have so little time to do things like organise and go on dates.</p> <p>My undergraduate degree was in Chemistry, but I quickly realised that I didn't want to be in a lab for a living.</p>	<p><b>Feel/Think/Believe</b></p>  <p>I have a lot of university friends now scattered all around – I want to stay in touch via social media (and meet up when we can) but I don't always keep on top of what people are up to as many of them post a lot of rubbish on their feeds.</p> <p>I love my job but I feel my work consumes most of my time and I have trouble switching off. There's always a project going on.</p> <p>While I have many social apps that I use a lot, I'm not convinced that that have my best interests at heart and I sometimes wonder if I share too much about myself with them.</p>							
<p><b>Technology Experience</b></p> <p>I guess I'm what people call a digital native. Technology isn't a big thing in its own right for me, but I suppose I depend on it for everything – work and social. While I happily admit that I don't use all the features on my phone (who does, really?), I am always keen to try out new apps and online services if they have something to offer me. If I am really into a particular app, I will take time to ensure that it works just so.</p> <p>I have a notebook computer for work, but I don't actually use it as much as I thought I would when I started – I mainly use it for typing things up and doing admin on the intranet (some of our admin apps don't work on mobile). I use my phone for a lot of my work, particularly for communication and collaborating. I only carry the one phone around with me for convenience, so it runs several work and personal apps.</p> <p>I mainly chat with people using an instant messenger. I pride myself on being able to select a meme for every occasion (although I'm not sure that counts as 'technology experience').</p>								
<p><b>Problems</b> → <b>Needs</b> → <b>Existing Solutions</b></p> <table border="1"> <thead> <tr> <th>Problems</th> <th>Needs</th> <th>Existing Solutions</th> </tr> </thead> <tbody> <tr> <td>Lack of time to stay on top of social media due to pressure of work.  Finds it hard to switch off from work in personal time.</td> <td>Find ways of keeping up with friends on social media that don't consume a lot of time wading through minutiae.  Help making a distinction between work and personal time when using a single device for both.</td> <td>Existing algorithms in social services that select content for users.  Scheduled or manual 'do not disturb' features in apps used for work (i.e. Slack's Do Not Disturb hours).</td> </tr> </tbody> </table>			Problems	Needs	Existing Solutions	Lack of time to stay on top of social media due to pressure of work.  Finds it hard to switch off from work in personal time.	Find ways of keeping up with friends on social media that don't consume a lot of time wading through minutiae.  Help making a distinction between work and personal time when using a single device for both.	Existing algorithms in social services that select content for users.  Scheduled or manual 'do not disturb' features in apps used for work (i.e. Slack's Do Not Disturb hours).
Problems	Needs	Existing Solutions						
Lack of time to stay on top of social media due to pressure of work.  Finds it hard to switch off from work in personal time.	Find ways of keeping up with friends on social media that don't consume a lot of time wading through minutiae.  Help making a distinction between work and personal time when using a single device for both.	Existing algorithms in social services that select content for users.  Scheduled or manual 'do not disturb' features in apps used for work (i.e. Slack's Do Not Disturb hours).						

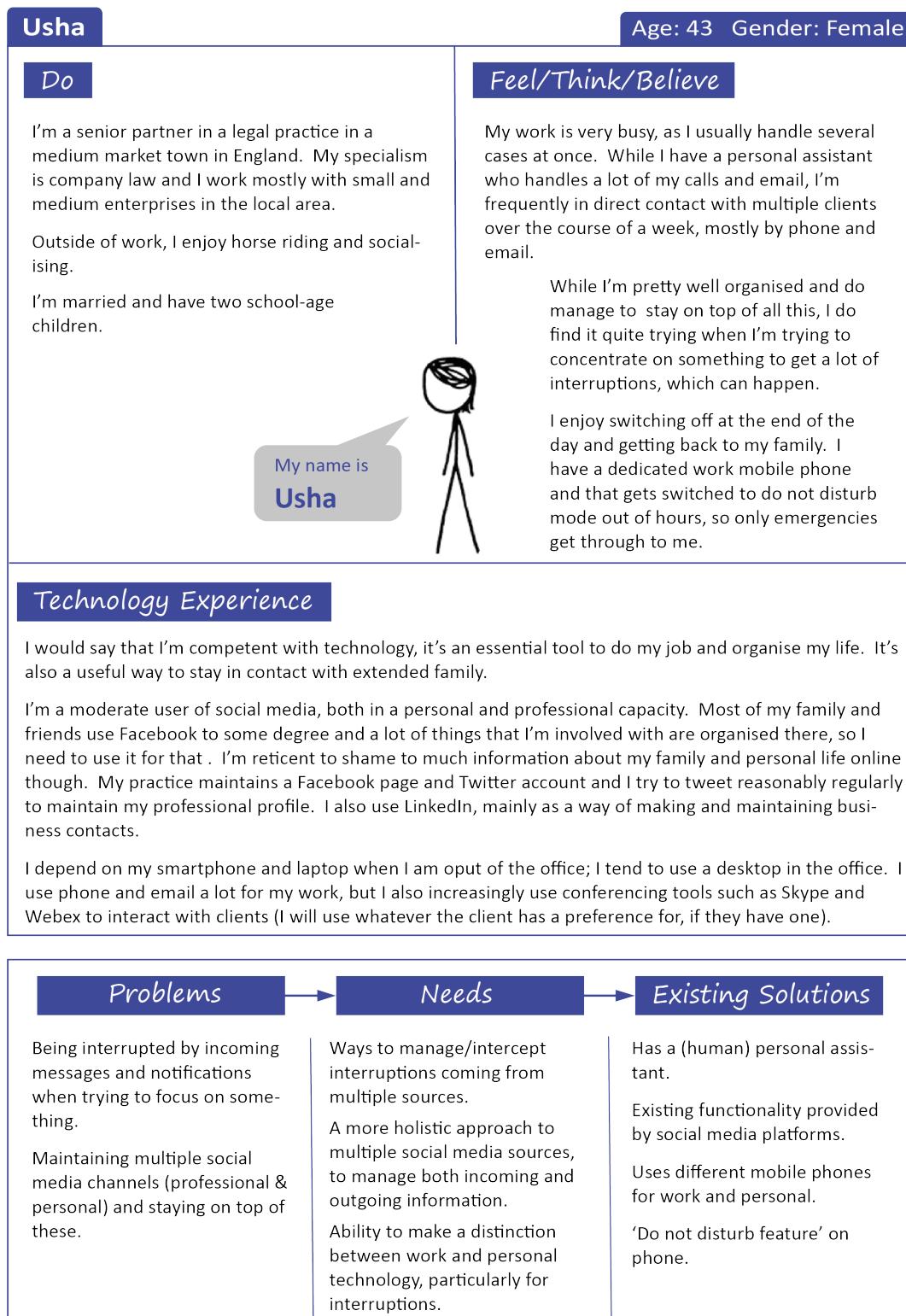
**S4 Personas**

Figure S4.4: Final PATHY Persona "Kenton"

Kenton		Age: 49 Gender: Male						
<b>Do</b>	<p>I work as a client manager for a global consumer products company. You will have heard of us. My role is essentially to drive sales, but looking after existing customers and bringing new ones in. As many of my accounts are themselves global corporates, I end up doing a fair bit of international travel.</p> <p>I never went to university – it didn't interest me at the time – so I started my career at 18 just after A-levels.</p> <p>I'm divorced, with one son who lives with his mum but I see often.</p> <p>Outside work, I watch football and play golf. I used to do it more the other way around when I was younger.</p>							
	<b>Feel/Think/Believe</b>	<p>I have a lot to stay on top of at work. I sometimes miss things, but someone usually reminds me if it's urgent. I prefer not to miss things from important customers though, even if they are not urgent I like to give them a prompt reply. It can be difficult though, when you get things flying at you from all directions at all times.</p> <p>I'm on Facebook, Twitter and all that, like most people are. Not sure how I'd organise my social life without Facebook (how did I manage before?)</p> <p>I hear all the usual scare stories about social media giants, but they're a bit overblown by the media. I don't over-share, but I'm happy to put information out there – I think you get more out of it that way. Of course in an ideal world I'd be happy to have more control.</p>						
<b>Technology Experience</b> <p>I don't think I'm held back by not having a degree. A lot of graduates joining my company haven't got a clue about the real business world anyway. Experience counts for a lot and I've always done well for getting stuck in with new things.</p> <p>I really believe that you get out what you're prepared to put in with technology. I'll put in the time when I can to get things set up right, tell it my preferences or whatever helps.</p> <p>In mainly use office tools like Word and Excel and communications tools such as Webex day to day. We also have a new CRM system that I have to use, as well as the old CRM system that we haven't completely migrated off yet, and the other CRM system that we also use, for reasons I don't fully understand.</p> <p>I travel a lot, so I'm pretty good at managing with a notebook and mobile.</p>								
 <p><b>Problems</b> → <b>Needs</b> → <b>Existing Solutions</b></p> <table border="1"> <thead> <tr> <th>Problems</th> <th>Needs</th> <th>Existing Solutions</th> </tr> </thead> <tbody> <tr> <td>Has to use multiple different systems that are not integrated with each other. Receives many interruptions from different sources that need to be prioritised differently.</td> <td>Methods to handle and prioritise incoming messages.</td> <td>Email rules. Social media functionality. Features of bespoke applications that control notifications.</td> </tr> </tbody> </table>			Problems	Needs	Existing Solutions	Has to use multiple different systems that are not integrated with each other. Receives many interruptions from different sources that need to be prioritised differently.	Methods to handle and prioritise incoming messages.	Email rules. Social media functionality. Features of bespoke applications that control notifications.
Problems	Needs	Existing Solutions						
Has to use multiple different systems that are not integrated with each other. Receives many interruptions from different sources that need to be prioritised differently.	Methods to handle and prioritise incoming messages.	Email rules. Social media functionality. Features of bespoke applications that control notifications.						

**S4 Personas**

Figure S4.5: Final PATHY Persona "Usha"



# **Supplement S5**

## **Persona Scenarios**

## Persona Scenarios

This document contains a number of Persona Scenarios that have been developed for the personas listed in Supplement S4 [[doi:10.21954/ou.rd.28045454](https://doi.org/10.21954/ou.rd.28045454)]. These can be used to provide general design scope and direction for an application intended to serve these personas.

## Contents

- S5.1 – Susan
- S5.2 – Adam
- S5.3 – Phoebe
- S5.4 – Kenton
- S5.5 – Usha

### S5.1 Susan

Table S5.1: Scenario 1 for Persona Susan

---

**Susan - Staying Connected with Family**

---

**Goal:** Susan wants to stay updated on her children's activities without getting overwhelmed by unnecessary notifications.

---

**Context:** Susan is preparing for a tennis match on a Saturday morning and wants to check her notifications quickly.

---

**Day/Time:** Saturday morning

---

**Narrative:**

1. Susan opens her smartphone and checks her notifications.
  2. She wants to see updates from her children on their various social media apps but is overwhelmed by unrelated posts.
  3. Susan opens her Awareness Agent app on the *Family - Children* channel.
  4. The app presents a summarised feed of her children's posts and a tennis match reminder.
  5. Satisfied, Susan heads out for her match without worrying that she missed anything important.
-

**S5 Persona Scenarios**

Table S5.2: Scenario 2 for Persona Susan

<b><u>Susan - Maintaining Focus at Work</u></b>
<b>Goal:</b> Susan wants to manage her work tasks effectively without being distracted by non-work-related notifications.
<b>Context:</b> Susan is at work after lunch on Wednesday, trying to focus on completing an important report.
<b>Day/Time:</b> Wednesday afternoon
<b>Narrative:</b>
1. Susan is working on her report when she receives several notifications on her phone.
2. She realises that many of the notifications are from social media and personal apps.
3. To avoid distraction, Susan opens her Awareness Agent app and filters on the <i>Work</i> channel
4. The app filters out non-work-related notifications, showing only work-related emails and messages.
5. Susan completes her report without interruption and feels more productive.

Table S5.3: Scenario 3 for Persona Susan

<b><u>Susan - Balancing Personal and Professional Life</u></b>
<b>Goal:</b> Susan wants to effectively manage the balance between her personal life and professional responsibilities.
<b>Context:</b> Susan is spending time with her family on a Sunday evening but still needs to keep an eye on work-related notifications in case something urgent arises.
<b>Day/Time:</b> Sunday evening
<b>Narrative:</b>
1. Susan is enjoying a film night with her family but she is getting notifications on her phone.
2. She doesn't want to disrupt the evening with work-related interruptions unless it's something urgent.
3. Susan opens her Awareness Agent app and sets it to the <i>Work - Urgent</i> channel, where only urgent work notifications are allowed through.
4. The app filters out non-urgent emails and messages, notifying her only if there is a critical work issue.
5. Susan continues to enjoy her evening, confident that she won't miss any important work matters.

## S5.2 Adam

Table S5.4: Scenario 1 for Persona Adam

---

### **Adam - Managing Multiple Projects**

---

**Goal:** Adam needs to keep track of multiple ongoing projects without missing important updates or deadlines.

**Context:** Adam is working from home and needs to juggle several tasks across different projects.

---

**Day/Time:** Tuesday morning

---

**Narrative:**

1. Adam starts his day by reviewing his to-do list and checking project updates.
  2. His inbox and messaging is filled with emails from various colleagues and clients, making it hard to prioritise.
  3. Adam opens his Awareness Agent app and accesses his custom channel for priority project-related updates.
  4. The app surfaces his relevant project communications, with other items kept out of the way.
  5. Adam efficiently organises his day, ensuring that nothing important is overlooked.
- 

Table S5.5: Scenario 2 for Persona Adam

---

### **Adam - Balancing Work and Personal Life**

---

**Goal:** Adam wants to maintain a clear boundary between his work commitments and personal time.

**Context:** Adam is taking time to organise his cycling club's annual race championships, but he needs to stay available for urgent work matters.

---

**Day/Time:** Tuesday evening

---

**Narrative:**

1. Adam is in his home office working on logistics for the cycle club championships, and he needs to focus on this task.
  2. He knows that there is also a possibility that he will need to visit a customer tomorrow, and needs to keep an eye out for communications about that.
  3. Adam chooses only the *Personal - Cycle Club* and *Work - Urgent* channels in his Awareness Agent app.
  4. The app filters out non-urgent work emails and messages, alerting him only if there's an emergency.
  5. Adam completes his task, knowing he won't miss anything important at work.
-

**S5 Persona Scenarios**

Table S5.6: Scenario 3 for Persona Adam

<b><u>Adam - Enhancing Productivity During Travel</u></b>
<b>Goal:</b> Adam wants to remain productive while travelling on a business trip, managing both work and personal tasks efficiently.
<b>Context:</b> Adam is on an early train headed to a customer site on a Monday and needs to catch up on work.
<b>Day/Time:</b> Monday morning
<b>Narrative:</b>
1. Adam is travelling by train and wants to use the time to complete some tasks.
2. His phone is receiving a mix of personal and work notifications, making it hard to focus.
3. Adam switches his Awareness Agent app to the <i>Morning Catch-Up</i> channel, which prioritises important work and personal-related notifications & tasks.
4. The app provides a focused list of items for Adam to review during the journey, minimising distractions.
5. Adam completes several work and personal tasks efficiently, making the most of his travel time.

**S5.3 Phoebe**

Table S5.7: Scenario 1 for Persona Phoebe

<b><u>Phoebe - Staying Connected without the Clutter</u></b>
<b>Context:</b> Phoebe wants to keep in touch with her university friends without being overwhelmed by trivial updates and irrelevant notifications.
<b>Day/Time:</b> A typical evening after work, around 8 PM.
<b>Narrative:</b>
1. Phoebe wants to use her social media apps to check in with her university friends.
2. She sees a barrage of notifications about unimportant updates from friends and acquaintances.
3. She feels frustrated and wants to concentrate only on meaningful updates, such as messages about upcoming reunions or significant life events.
4. Phoebe uses her Awareness Agent app on the <i>Friends &amp; Family - Important</i> channel; this shows her only significant posts and notifications about her friends and family coming from multiple social apps.
5. Phoebe feels satisfied, as she can now stay connected without being distracted by trivial updates.

**S5 Persona Scenarios**

Table S5.8: Scenario 2 for Persona Phoebe

<b><u>Phoebe - Managing Work-Life Balance</u></b>
<b>Goal:</b> Phoebe needs to maintain a clear separation between her work and personal life, ensuring that neither one interferes unduly with the other.
<b>Context:</b> Phoebe is working from home in the afternoon, and she wants to keep her work and personal tasks separate.
<b>Day/Time:</b> Thursday afternoon
<b>Narrative:</b>
<ol style="list-style-type: none"> <li>1. Phoebe is finishing up some work tasks at home when she receives several personal notifications on her phone.</li> <li>2. She finds it challenging to stay focused with the constant stream of messages from friends and family.</li> <li>3. Phoebe switches her Awareness Agent app to the <i>Work</i> channel, which prioritises work-related notifications and temporarily mutes personal alerts.</li> <li>4. This allows her to complete her work without distractions.</li> <li>5. After finishing her tasks, Phoebe switches to the <i>Personal</i> channel catching up on her messages and enjoying her evening without work interruptions.</li> </ol>

Table S5.9: Scenario 3 for Persona Phoebe

<b><u>Phoebe - Training the Awareness Agent</u></b>
<b>Goal:</b> Phoebe wants to personalize her Awareness Agent app settings to better suit her unique preferences and routines.
<b>Context:</b> Phoebe is tweaking the settings of her Awareness Agent app on a Monday evening to ensure it matches her lifestyle.
<b>Day/Time:</b> Monday evening
<b>Narrative:</b>
<ol style="list-style-type: none"> <li>1. Phoebe notices that some non-urgent work notifications are still coming through in her <i>Work - Urgent</i> channel, which she finds distracting.</li> <li>2. She decides to spend a short time training in her Awareness Agent app to better tell the difference between urgent and other work notifications.</li> <li>3. After training the app, Phoebe finds that she sees far fewer non-urgent content in this channel and can focus her time better.</li> </ol>

## S5.4 Kenton

Table S5.10: Scenario 1 for Persona Kenton

### **Kenton - Handling Overlapping Deadlines**

**Goal:** Kenton needs to manage multiple overlapping deadlines for different projects without missing any key tasks.

**Context:** Kenton is at his office on Wednesday, trying to prioritize tasks for two projects that have deadlines at the end of the week.

**Day/Time:** Wednesday afternoon

**Narrative:**

1. Kenton is aware that both of his projects have critical deadlines approaching and needs to focus on these.
2. His email and messaging tools are filled with notifications and updates, making it hard to focus.
3. Kenton chooses the *Work - Current Projects* channel on his Awareness Agent app, filtering notifications to only show updates related to current projects.
4. Kenton is able to concentrate on the critical tasks, ensuring he meets both deadlines without missing important details.

Table S5.11: Scenario 2 for Persona Kenton

### **Kenton - Managing Interruptions**

**Goal:** Kenton wants to minimise interruptions during his focus time to enhance productivity while working on detailed tasks.

**Context:** Kenton is working on a detailed report at work and wants to avoid interruptions from non-essential notifications.

**Day/Time:** Tuesday morning

**Narrative:**

1. Kenton begins working on a complex report that requires his full attention.
2. He starts receiving notifications from various apps and services, which disrupt his focus.
3. Kenton uses his Awareness Agent app to activate the *Urgent* channel, which blocks all non-essential content.
4. The app silences all but the most critical alerts, allowing Kenton to concentrate fully on his report safe in the knowledge that he will not miss anything truly urgent.
5. After completing the report, Kenton changed back to the normal channels and addresses any items that require his attention.

**S5 Persona Scenarios**

Table S5.12: Scenario 3 for Persona Kenton

**Kenton - Organising Golf Meetups without the Noise**

**Context:** Kenton is part of a social group chats for both golf and football, but he wants to separate the logistics of arranging golf meetups from the general banter and football discussions.

**Day/Time:** Saturday morning

**Narrative:**

1. Kenton receives a mix of messages in his group chat, including discussions about the latest football match and plans for a golf game next weekend.
2. He finds it difficult to keep track of the logistics for the golf meetup amidst the constant football banter.
3. Kenton uses his Awareness Agent app to activate the *Personal - Golf Arrangements* channel, which contains only content about meeting up to actually play golf.
4. Kenton is able to arrange a golf match for Sunday morning and can get back to the chat.

**S5.5 Usha**

Table S5.13: Scenario 1 for Persona Usha

**Usha - Demarcating Work and Personal Time**

**Goal:** Usha wants to ensure that her work does not intrude on her personal time, maintaining a clear boundary between the two.

**Context:** Usha is at home on a Friday evening, preparing to relax after a busy week at work.

**Day/Time:** Friday evening

**Narrative:**

1. Usha finishes her work for the week and is ready to switch off for the weekend.
2. She begins receiving work-related notifications on her phone, which disrupt her attempts to relax.
3. Usha sets Awareness Agent app to the *Personal or Urgent* channel which filters out all non-urgent work-related content
4. The app hides all unimportant work content, allowing Usha to enjoy her evening without work intruding.
5. She feels at ease, knowing that her work will not intrude on personal time.

**S5 Persona Scenarios**

Table S5.14: Scenario 2 for Persona Usha

**Usha - Dealing with Low Priority Unwanted Work Content**

**Goal:** Usha wants to minimise the number of irrelevant content she views during her work day, so that she can stay focused.

**Context:** Usha is at her office on a Monday morning, preparing to start her work day.

**Day/Time:** Monday morning

**Narrative:**

1. Usha starts the day and notices an influx of notifications and content from apps she doesn't need during work hours.
2. These notifications distract her from her tasks and slow her down.
3. Usha configures her Awareness Agent app to show only higher priority work content during office hours.
4. The app filters out irrelevant content, allowing Usha to focus on her work without unnecessary interruptions.
5. By the end of the day, Usha feels more productive and less distracted.

Table S5.15: Scenario 3 for Persona Usha

**Usha - Customizing Content Preferences**

**Goal:** Usha wants to fine-tune her content preferences to ensure she only receives the most important alerts.

**Context:** Usha is spending her Sunday evening preparing for the upcoming week.

**Day/Time:** Sunday evening

**Narrative:**

1. Usha is preparing for the week ahead and decides to review how the Awareness Agent app handles her content.
2. She's aware that she is receiving too much low-priority work content in the *Work - Important* channel.
3. Usha uses her Awareness Agent app to train the *Work - Important* channel, fine-tuning it to better focus on important content.
4. Usha feels more in control, knowing that she will only be alerted to the most important messages during the week.

## **Supplement S6**

### **Synthetic Content Materials**

# Synthetic Content Materials

This document contains information relating to Synthetic Content, as discussed in Chapter 7.4 of the thesis. This includes schemas and example data for synthetic content requests and a simplified sample of the Python code used to generate these.

The Topic materials for synthetic content are documented separately in Supplement S7 [doi:10.21954/ou.rd.28045490].

## Contents

- S6.1 – Schemas
- S6.2 – Prompting
- S6.3 – Dramatis Personae
- S6.4 – Entities
- S6.5 – Scheduling

### S6.1 Simulation Schemas

This section contains listings of the schemas for synthetic content. These are also available at: doi:10.21954/ou.rd.28044944 [path: /sim/schemas].

Listing S6.1 shows the schema passed to the OpenAI chat completions API tools function<sup>1</sup> by the generate\_simulated\_text() code shown in Listing S6.4.

```
{  
  "$id": "https://parse.net/awag/0.1/simulation-messages-result.schema.json",  
  "$schema": "https://json-schema.org/draft/2020-12/schema",  
  "name": "get_simulated_messages",  
  "description": "Return a set of simulated text messages based on the provided information",  
  "parameters": {  
    "type": "object",  
    "properties": {  
      "data": {  
        "type": "array",  
        "items": {  
          "type": "object",  
          "title": "Simulation results",  
          "description": "Array containing all of the simulated text messages for the provided data",  
          "properties": {  
            "category": {  
              "title": "Category",  
              "type": "string",  
              "enum": ["Work", "Personal", "Social", "Entertainment"]  
            }  
          },  
          "required": ["category"]  
        }  
      }  
    }  
  }  
}
```

<sup>1</sup>[https://cookbook.openai.com/examples/how\\_to\\_call\\_functions\\_with\\_chat\\_models](https://cookbook.openai.com/examples/how_to_call_functions_with_chat_models) [<https://perma.cc/FY7T-WUNA>]

## S6 Synthetic Content Materials

---

```

        "description":"The category ID that was supplied for this set of messages",
        "type":"string"
    },
    "text":{
        "title":"Item Text",
        "description":"The text content of this message",
        "type":"string"
    },
    "userid":{
        "title":"User ID",
        "description":"The ID of the simulated user for this message, this should resemble
            the username",
        "type":"string"
    },
    "username":{
        "title":"User Name",
        "description":"The name of the simulated user for this message (include both first
            name and surname)",
        "type":"string"
    }
},
"required":[
    "category",
    "text",
    "userid",
    "username"
]
}
}
}
}
```

Listing S6.1: Simulation Messages Result Schema

Listing S6.2 shows the schema that describes our Simulation Dramatis Personae document. This is our schema (as opposed to a schema specified by a third party service for example), and describes the Dramatis Personae JSON document that is passed with the simulation request. While no formal mechanism existed in the OpenAI API for passing such input schemas with requests, we opted to pass these as System Messages, with the content itself passed as User Messages, as seen in Listing S6.4. Examples of the *Dramatis Personae* JSON are included in Section S6.3.

```
{
"$id":"https://parse.net/awag/0.1/simulation-dramatis-personae.schema.json",
"$schema":"https://json-schema.org/draft/2020-12/schema",
"type":"array",
"title":"Dramatis Personae List",
"items":[
{
    "type":"object",
    "title":"Group",
    "description":"Collection of identities that are linked in some way",
    "properties":{
        "name":{
            "type":"string",
            "description":"The name of this group of identities"
        }
    }
}
]
```

## S6 Synthetic Content Materials

```
},
"description": {
    "type": "string",
    "description": "Description of the contents of this group; use this value to determine how
        and when to represent/use them"
},
"members": {
    "type": "array",
    "description": "The list of identities in this group",
    "items": [
        {
            "type": "object",
            "title": "Identity",
            "description": "An individual identity",
            "properties": {
                "surname": {
                    "type": "string",
                    "description": "The surname of family name of this identity"
                },
                "firstname": {
                    "type": "string",
                    "description": "The first name of this identity"
                },
                "userid": {
                    "type": "string",
                    "description": "A computer or application username to use for this identity"
                },
                "role": {
                    "type": "string",
                    "description": "The role of this identity within the group, and/or the
                        relationship that the identity has to the principle character"
                }
            },
            "required": [
                "surname",
                "firstname",
                "userid"
            ]
        }
    ]
},
"required": [
    "name",
    "description",
    "members"
]
}
}
```

Listing S6.2: Simulation Dramatis Personae Schema

## S6 Synthetic Content Materials

---

Listing S6.3 shows the schema that describes our Simulation Entities document. This is also our own schema. Examples of the Entities JSON are included in Section S6.4.

```
{
  "$id": "https://parse.net/awag/0.1/simulation-entities.schema.json",
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "title": "Entities Schema",
  "description": "Schema for representing details of entities that can be referenced in simulated messages. Entity types can be companies, clubs or organisations and can include their associated people, and other specific information.",
  "type": "array",
  "items": {
    "type": "object",
    "properties": {
      "name": {
        "type": "string",
        "description": "The name of the entity"
      },
      "also_known_as": {
        "type": "array",
        "description": "Alternate names or abbreviations for the entity",
        "items": {
          "type": "string"
        }
      },
      "type": {
        "type": "string",
        "description": "The type of entity, or the relationship it has to the subject (e.g., Client, Competitor)"
      },
      "notes": {
        "type": "string",
        "description": "Information to use to understand the nature of the entity and how to refer to it in messages"
      },
      "people": {
        "type": "array",
        "description": "List of people associated with the entity that can be referred to in simulated content",
        "items": {
          "type": "object",
          "properties": {
            "name": {
              "type": "string",
              "description": "Name of the person."
            },
            "role": {
              "type": "string",
              "description": "Role of the person within the entity"
            }
          },
          "required": ["name", "role"]
        }
      }
    },
    "required": ["name", "type", "notes", "people"]
  }
}
```

Listing S6.3: Simulation Entities Schema

## S6 Synthetic Content Materials

---

### S6.2 Simulated Content Prompting

Listing S6.4 shows an abridged Python code snippet of the awagdata function for generating simulated content from OpenAI using the /chat/completions API. Listing S6.5 shows the prompt template used by this code.

```
def generate_simulated_text(self, category, topic, item_count, prompt_template, dramatis_personae=None,
                           entities=None, openai_params=None, openai_model=None, is_use_history=False):

    agent_id = self.get_agent_id()
    openai_client = OpenAIWrapper(self.openai_auth_token)

    prompt = prompt_template.format(category=category, topic=topic, item_count=item_count)

    model = openai_model
    if not model:
        model = self.openai_engine

    messages = openai_client.append_system_message(f"This is the schema to use for identities:\n{json.dumps(self.simulation_dramatis_personae_schema)}")

    openai_client.append_system_message(f"This is the schema to use for entities:\n{json.dumps(self.simulation_entities_schema)}", messages)
    openai_client.append_system_message(prompt, messages)
    openai_client.append_user_message(f"This is the topic that I want you to generate content for: {topic}", messages)
    if dramatis_personae:
        openai_client.append_user_message(f"This is identities data (dramatis_personae):\n{json.dumps(self.dramatis_personae)}", messages)
    if entities:
        openai_client.append_user_message(f"This is entities data:\n{json.dumps(entities)}", messages)
    if is_use_history:
        historical_messages = self._get_historical_messages(category, 150)
    openai_client.append_user_message(f"This is a list of already existing messages for this topic. Do not repeat any exact existing message in your content, but you can have your new messages make references to existing ones where appropriate:\n{json.dumps(historical_messages)}", messages)

    tools = [openai_client.get_tools_json_from_function_schema(self.get_simulated_messages_function)]
    tool_choice = openai_client.get_tool_choice_json("get_simulated_messages")

    logger.debug(f"Generated messages: {messages}")

    openai_response, info_json = openai_client.run_chat_completions(
        messages=messages,
        model=model,
        tools=tools,
        tool_choice=tool_choice,
        openai_params=openai_params
    )

    simulated_messages = self._extract_simulated_messages(openai_response, info_json)

    if len(simulated_messages) == 0:
        return self.generate_response("ERROR", f"No data generated", code=500, data=response_data)
    else:
        return self.generate_response("OK", "Items generated successfully", code=200, data=simulated_messages, info=info_json)
```

Listing S6.4: Simulation Messages Generation Code Snippet

## S6 Synthetic Content Materials

```
SIMULATION_MESSAGES_PROMPT_TEMPLATE = """I would like you to generate simulated messages from simulated users that you might see in a messaging application such as Teams, Slack or WhatsApp.

The messages should look like they are real, and relate to a topic that I will provide. You can make up fake names for people, products, projects and companies, as well as using those that are provided.

You may be supplied with a list of entities to use in content generation; these could be companies, organisations, clubs etc. that people might refer to in some messages in the context of the provided topic. The entities list is provided as a JSON object conforming to the schema: https://parse.net/awag/0.1/simulation-entities.schema.json

The 'text' property of your response data should contain the text that you generate. The 'category' property should be a string that identifies the topic that I will give you. The 'userid' and 'username' properties should describe a fictional user that the message is from: 'userid' is a short ID string containing only alphanumeric characters; 'username' is a proper name of the simulated user (first name and last name). The value of 'userid' should be based on the value of 'username' (for example a user with username 'Fred Wilson' might have a userid of 'fredwilson47' or 'fred1')

You should assign many simulated users to messages. You may be supplied with a list of simulated identities to use, in a JSON object conforming to schema: https://parse.net/awag/0.1/simulation-dramatis-personae.schema.json
Some of the identities have a role assigned that indicates a specific relationship to the subject, or a particular role in their organisation. For those without a role assigned, you can still use the identity but assume a generic role for them.

These simulated users can either be the sender of any given message, or referred to in other messages, as appropriate.

If you need to generate your own identities, follow these rules:
- The userid and username combination should be consistent for each time the user is re-used.
- Simulated user names should be realistic actual British or American names
- Simulated identities should have a representative distribution of gender and ethnic background.

To give the appearance of an ongoing discussion, the simulated users should send multiple messages; you should try to use at least 4 distinct simulated user identities for every 10 messages that you generate.

In this case the category string should be: {category}

Please generate {item_count} items."""
```

Listing S6.5: Simulation Messages Prompt Template

## S6 Synthetic Content Materials

---

### S6.3 Dramatis Personae

Figures S6.1 to S6.3 show some abridged examples of *dramatis personae* JSON documents. Full documents can be found at:

[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /sim/persona-data/dramatis-personae].

```
[
  {
    "name": "bwcc-chat-members",
    "description": "Members of Adam's cycle team Borchester Wheelers",
    "members": [
      {
        "surname": "Thompson",
        "firstname": "George",
        "userid": "george.thompson",
        "role": "Club president"
      },
      {
        "surname": "Clarke",
        "firstname": "Adam",
        "userid": "adam.clarke",
        "role": "Big figure with a reputation for being a very strong rider"
      },
      {
        "surname": "Hughes",
        "firstname": "Benjamin",
        "userid": "benjamin.hughes",
        "role": "Road racing coordinator"
      },
      {
        "surname": "Wilson",
        "firstname": "Ethan",
        "userid": "ethan.wilson",
        "role": "Virtual (Zwift) racing coordinator"
      },
      {
        "surname": "Smith",
        "firstname": "Daniel",
        "userid": "daniel.smith",
        "role": "Zwift racing team-mate of Adam"
      }
    ]
  }
]
```

Figure S6.1: Abridged Dramatis Personae - Adam/Cycling

## S6 Synthetic Content Materials

---

```
[
  {
    "name": "work-team",
    "description": "Members of Adam's team at work in Borchester Software, The team is called the Client Technology Group (CTG)",
    "members": [
      {
        "surname": "Walker",
        "firstname": "Charlotte",
        "userid": "walkerc",
        "role": "Adam's team manager"
      },
      {
        "surname": "Khan",
        "firstname": "Jayden",
        "userid": "khanj",
        "role": "New starter in the team, recent graduate"
      },
      {
        "surname": "Singh",
        "firstname": "Oscar",
        "userid": "singho",
        "role": "Close colleague to Adam, often work together"
      },
      {
        "surname": "Jones",
        "firstname": "Amelia",
        "userid": "jonesa",
        "role": ""
      }
    ]
  },
  {
    "name": "work-executives",
    "description": "Members of the executive team of Borchester Software",
    "members": [
      {
        "surname": "Johnson",
        "firstname": "Liam",
        "userid": "johnsonl",
        "role": "CEO"
      },
      {
        "surname": "White",
        "firstname": "Henry",
        "userid": "whiteh",
        "role": "CFO"
      }
    ]
  },
  {
    "name": "work-other",
    "description": "Other employees of Borchester Software",
    "members": [
      {
        "surname": "Chen",
        "firstname": "Youssef",
        "userid": "cheny",
        "role": "Adam's manager's manager"
      },
      {
        "surname": "Khan",
        "firstname": "Mei",
        "userid": "khamn",
        "role": ""
      }
    ]
  }
]
```

Figure S6.2: Abridged Dramatis Personae - Adam/Work

## S6 Synthetic Content Materials

---

```
[
  {
    "name": "family",
    "description": "Members of Susan's immediate family",
    "members": [
      {
        "surname": "Carter",
        "firstname": "Neil",
        "userid": "neil",
        "role": "Susan's husband; works as a pig farmer at a local farm"
      },
      {
        "surname": "Grundy",
        "firstname": "Emma",
        "userid": "emma",
        "role": "Susan and Neil's daughter, married to Ed Grundy"
      },
      {
        "surname": "Grundy",
        "firstname": "Ed",
        "userid": "ed",
        "role": "Susan and Neil's son in law, married to Emma"
      }
    ]
  },
  {
    "name": "friends",
    "description": "Friends and acquaintances of Susan and Neil",
    "members": [
      {
        "surname": "Aldridge",
        "firstname": "Brian",
        "userid": "brian",
        "role": "Alice Aldridge's father; runs Brookfield Farm"
      },
      {
        "surname": "Aldridge",
        "firstname": "Jennifer",
        "userid": "jennifer",
        "role": "Alice Aldridge's mother; runs Brookfield Farm"
      },
      {
        "surname": "McCreary",
        "firstname": "Jack",
        "userid": "jazzer",
        "role": "Also know as Jazzer; works with Neil at Berrow Farm"
      }
    ]
  }
]
```

Figure S6.3: Abridged Dramatis Personae - Susan/Personal

## S6 Synthetic Content Materials

---

### S6.4 Entities

Figures S6.4 to S6.6 show some abridged examples of entities JSON documents. Full documents can be found at: [doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /sim/persona-data/entities].

```
[
  {
    "name": "Borchester Wheelers Cycling Club",
    "also_known_as": [
      "BWCC",
      "Borchester Wheelers",
      "the Wheelers"
    ],
    "type": "Club",
    "notes": "The cycling club that our subject, Adam Macy is a member of",
    "people": []
  },
  {
    "name": "Darrington Dynamos",
    "also_known_as": [
      "Dynamo"
    ],
    "type": "Club",
    "notes": "Darrington Dynamos are BWCC's big local rivals",
    "people": []
  },
  {
    "name": "Zwift Racing League",
    "also_known_as": [
      "ZRL"
    ],
    "type": "Organisation",
    "notes": "ZRL organises the Zwift racing league that BWCC participates in",
    "people": []
  }
]
```

Figure S6.4: Abridged Entities - Adam/Cycling

```
[
  {
    "name": "Berrow Farm",
    "also_known_as": [],
    "type": "Company (Farm)",
    "notes": "This is the farm near Ambridge where Neil Carter works as a pig manager",
    "people": [
      {
        "name": "Justin Elliott",
        "role": "Chair of Borchester Land, who own Berrow Farm"
      },
      {
        "surname": "Jazzer McCreary",
        "role": "Works with Neil at Berrow Farm"
      }
    ]
  },
  {
    "name": "The Bull",
    "also_known_as": [
      "the pub"
    ],
    "type": "Company",
    "notes": "This is the popular village pub in the village of Ambridge, where a lot of Kenton's family and friends live, work and socialise",
    "people": [
      {
        "name": "Fallon Rogers",
        "role": "Landlady"
      },
      {
        "name": "Wayne Tucson",
        "role": "Chef"
      }
    ]
  }
]
```

Figure S6.5: Abridged Entities - Susan/Personal

## S6 Synthetic Content Materials

---

```
[
  {
    "name": "Borchester Software Ltd.",
    "also_known_as": [
      "BIS",
      "Borchester Software"
    ],
    "type": "Employer",
    "notes": "Borchester Software is the employer of the subject, Adam Macy",
    "people": []
  },
  {
    "name": "Colossal Widgets Ltd.",
    "also_known_as": [
      "ColWidgets"
    ],
    "type": "Client",
    "notes": "This is one of Borchester Software's biggest customers",
    "people": [
      {
        "name": "Freddie Widget",
        "role": "CEO"
      },
      {
        "name": "Andy Smithy",
        "role": "Main customer contact"
      }
    ]
  },
  {
    "name": "Big Information Systems Inc",
    "also_known_as": [
      "BIS",
      "BigSys"
    ],
    "type": "Competitor",
    "notes": "This is one of Borchester Software's biggest business competitors",
    "people": [
      {
        "name": "Thomas T Big",
        "role": "CEO"
      },
      {
        "name": "Evan Fournier",
        "role": "Top salesperson"
      }
    ]
  },
  {
    "name": "Tiny Widgets Company",
    "also_known_as": [
      "TWC"
    ],
    "type": "Client",
    "notes": "This is a small client that Borchester Software has recently acquired",
    "people": [
      {
        "name": "Justin Sprocket",
        "role": "CEO"
      },
      {
        "name": "Susan Williams",
        "role": "Main customer contact"
      },
      {
        "name": "Mary Parker",
        "role": "Technical contact"
      }
    ]
  }
]
```

Figure S6.6: Abridged Entities - Adam/Work

## S6 Synthetic Content Materials

### S6.5 Content Scheduling

Listing S6.6 shows the Java Awareness Agent Java code `generateEventTimes()` method that is used to apply a random offset for each event of a few minutes through the hour for scheduled publication of content.

```
private static List<LocalTime> generateEventTimes(LocalTime hour, int volume, int maxOffsetMinutes) {  
    final List<LocalTime> eventTimes = new ArrayList<>();  
  
    if (volume < 1) {  
        return eventTimes;  
    }  
  
    final int minutesPerEvent = 60 / volume;  
    int minutesRemaining = 60;  
  
    final Random rand = new Random();  
  
    for (int i = 0; i < volume; i++) {  
        final int minutesUntilNextEvent = Math.min(minutesPerEvent, minutesRemaining);  
        final LocalTime baseEventTime = hour.plusMinutes(60 - minutesRemaining);  
        final int offsetMinutes = rand.nextInt(maxOffsetMinutes * 2 + 1) - maxOffsetMinutes;  
        final LocalTime eventTime = baseEventTime.plusMinutes(offsetMinutes);  
        eventTimes.add(eventTime);  
        minutesRemaining -= minutesUntilNextEvent;  
    }  
  
    return eventTimes;  
}
```

Listing S6.6: Awareness Agent Sim Event Time Generation (Java)

# **Supplement S7**

## **Synthetic Content Topics**

# Synthetic Content Topics

This document contains a full list of simulated content topics for personas that are referenced in Chapter 7. These are also available in JSON format at:

[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /sim/persona-data/topics]

## Contents

- [S7.1 – Susan](#)
- [S7.2 – Adam](#)
- [S7.3 – Phoebe](#)
- [S7.4 – Kenton](#)
- [S7.5 – Usha](#)

## S7.1 Susan

### S7.1.1 Susan - Work

#### **susan-work-university-announce**

Corporate messenger application chat within the employer of our fictional protagonist, Susan Carter. The messages are exclusively internal university announcements coming from the senior leadership team and PR office of the university that Susan works for, Borchester University. Generate messages that come from all of the senior leaders in the provided list as well as from an anonymous PR account or from Alumni Relations. Messages should occasionally refer to the university's academic achievements, rankings, campuses, faculty changes etc. Message content should be unique and not duplicated. Try and use a good range of content, making up names of internal and external entities such as campuses, buildings, research centres, funding bodies, business partnerships, alumni organisations and government entities. The tone of the messages should be formal business British English with British spellings.

#### **susan-work-university-general**

Workplace messenger application chat within the employer of our fictional protagonist, Susan Carter. The chat is general workplace discussion within the university that Susan works for, Borchester University. The chat is at the overall university staff level and not individual teams, so should mostly not refer to in-team topics. Topics should include discussion about university announcements and courses, upcoming events,

## S7 Synthetic Content Topics

---

academic climate, general news (where appropriate for university staff discussion) and other items people might choose to share with an organisation of this size. This group chat is intended for discussions aimed at the whole university administration - it is not the correct place for team discussions, technical support requests etc. Messages should occasionally refer to the university's international operations, locations, facilities etc. Some messages might be complaints, compliments or other discussion on university policies, changes, facilities, catering etc. as well as some external events; avoid generating duplicates or overly similar repeated messages. The tone of the messages should be informal/professional British English with British spellings.

### **susan-work-team-general**

Corporate messenger application chat for the work team of our fictional protagonist, Susan Carter. The chat is general workplace discussion between members of Susan's team in the Undergraduate Admissions Office at Borchester University. The team is half working in the office and half remote working (some people alternate between the two). About half of the team is part-time. Topics should include discussions about work (admissions queries, processes, asking for help or cover etc.), logistics, talking about out of work meetups and events and other topics that you might expect a work team to discuss. Some of the messages should imply a sense of urgency, such as a request for timely assistance, while most should be more relaxed. While a list of some colleagues in other parts of the university is provided, you should also generate fictional names to reference in messages; avoid generating duplicates or overly similar repeated messages. The tone of the messages should be informal British English with British spellings.

### **susan-work-team-manager**

Corporate messenger application chat between our fictional protagonist, Susan Carter and her manager Eleanor Baker, who heads up the Undergraduate Admissions Office at Borchester University where Susan works. It only includes messages between Susan Carter and Eleanor Baker. The team is half working in the office and half remote working (some people alternate between the two). About half of the team is part-time. This chat is just between Susan and her manager, and topics should include admissions-related discussions, logistics, and usual topics that a manager and their direct report might need to discuss. Some of the messages should be of a type that would normally need a prompt or urgent response. In this scenario, Susan is a solid performer at work, but has a high workload and not all of her colleagues pull their weight equally; avoid generating duplicates or overly similar repeated messages. The tone of the messages should be informal/professional British English with British spellings.

### **S7.1.2 Susan - Personal**

#### **susan-village**

General chat among residents of the fictional village of Ambridge in the county of Dorsetshire, UK, as seen by our protagonist, Susan Carter. The chat is a mostly light-hearted exchange between residents of Ambridge, covering topics including comings & goings around Ambridge, parking problems, postal service & parcel delivery (problems), recommendations for good tradespeople, social events, local crime and plenty of village gossip. This chat is among a general village audience rather than one-to-one messages about individual arrangements. A few of the messages should be of a nature that merits a prompt response, such as a request for assistance or information. Nearby villages include Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Faversham. Some of these locations should occasionally be mentioned in messages. In addition to the supplied list of people, you should make up identities as needed - the chat forum would contain maybe 50-30 members (although not all of these post messages often). The tone of the messages should be informal British English.

#### **susan-family**

General chat between members of a fictional British family, as seen by our protagonist, Susan Carter. The chat is a lighthearted exchange between family members on a variety of topics, with some messages relating to logistics and meeting up but also plenty of village gossip. A few of the messages should be of a nature that merits a prompt response, such as a request for assistance or information. Susan and Neil live in the village of Ambridge in Dorsetshire, while family are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Faversham. The tone of the messages should be informal British English.

#### **susan-neil**

Chat between our fictional protagonist, Susan Carter and her husband Neil. The chat is a exchange a variety of topics, with some messages relating to logistics and meeting up but also discussing how their day is going and talking about family and work events. Try and include reasonable variety. Susan is an administrator in the Undergraduate Admissions Office at Bournemouth University, where she commutes to work 3 days a week and works from home 2 days. Sometimes both Susan and Neil like to comment or complain about their day. A small amount of messages should be of a type that might need a prompt or urgent response. Susan and Neil live in the village of Ambridge, and Neil works with the pigs at nearby Berrow Farm, while family are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Faversham. The tone of the messages should be informal British English.

### **S7.1.3 Susan - Tennis**

#### **susan-tennis-club**

Chat forum within an English tennis club, Felpersham Tennis Club, that our fictional protagonist Susan Carter is member and co-organiser for. Members are from the areas around the cathedral city of Felpersham in Borsetshire where the club is located, including Abridge village where Susan lives, the main county town of Borchester, Penny Hassett and Loxley Barrett. Topics generally relate to tennis club business (fixtures, club elections and meetings, events at the clubhouse etc.), discussion about professional and amateur tennis competitions and other things you might expect members of a tennis club to discuss. Chat will often veer to other topics including current affairs, but there will often be someone along to remind people to stick vaguely to the topic. There is a ‘no politics’ rule that is sometimes broken but quickly reprimanded. There are some general jokes and light banter. The local rival club Darrington Tennis will often get a mention. Generate a range of messages to reflect the above and other topics; avoid generating duplicates or overly similar repeated messages. The tone of the messages should be informal British English with British spellings.

#### **susan-tennis-chat**

Chat between a group of friends who share an interest in tennis. Our protagonist, Susan, is a member of a tennis club which has its own discussion forum, but this is a separate group chat among friends. People mostly come from the fictional village of Ambridge in the county of Borsetshire, UK, or nearby. The group often meet at the tennis club where many are members (Felpersham Tennis Club) or village pub (The Bull) to socialise as well as actually playing tennis. Tennis played is a mixture of social and competitive fixtures. Topics often relate to tennis, but the chat will frequently veer to other topics including village gossip, current affairs and family news and you should generate a range of messages to reflect this. Tennis-related discussion will often include arranging informal one to one or doubles matches, discussion about injuries or technique and speculating on the weather and court availability. People may also mention locations in the main county town of Borchester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham where the club is based. There are some general jokes and light banter; avoid generating duplicates or overly similar repeated messages. The tone of the messages should be informal British English with British spellings.

## **S7.2 Adam**

### **S7.2.1 Adam - Work**

#### **adam-work-company-announce**

Corporate messenger application messages for the workplace of our fictional protagonist, Adam Macy. The messages are exclusively internal corporate announcements coming from the executive and senior leadership team of company that Adam works for, Borchester Software. Generate messages that come from all of the executives in the provided list. Borchester Software is an international company with approximately 500 employees who are based mainly in the UK, USA, Canada, Japan and Germany. Announcements can include (but are not limited to) industry topics, company-wide meetings, reminders about business conduct rules, customer contracts and acquisitions, new internal IT systems, and senior personnel changes. Please ensure that each batch of generated content contains messages that cover the whole range of topics. While a list of some clients is provided, you should also generate fictional client names to reference in those messages that relate to clients. You should also generate fictional internal team names to mention where necessary. There must be variation in message topic and content; avoid re-using the same phrases. The tone of the messages should be relatively formal business English with British spellings.

#### **adam-work-company-general**

Corporate messenger application chat for the workplace of our fictional protagonist, Adam Macy. The chat is general workplace discussion within company that Adam works for, Borchester Software (sometimes referred to by its initials), which is an international company with approximately 500 employees who are based mainly in the UK, USA, Canada, Japan and Germany. Topics should include discussion about company announcements and products, upcoming events, business climate, general news (where appropriate for corporate discussion) and other items people might choose to share with a company of this size. This group chat is intended for discussions aimed at the whole company - it is not the correct place for team discussions, technical support requests, local office matters etc. Message content should be unique and not duplicated. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal/business English with British spellings.

#### **adam-work-team-general**

Corporate messenger application chat for the work team of our fictional protagonist, Adam Macy. The chat is general workplace discussion between members of the same IT consulting team that Adam is in. The company that Adam works for is called Borchester Software (sometimes referred to by its initials) and his team of 10 people is called the Client Technology Group (CTG). The team is spread all over. Topics should

## S7 Synthetic Content Topics

---

include client discussions, logistics, talking about out of work meetups and events, some office gossip and other topics that you might expect a distributed work team to discuss. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Some of the messages should convey some sort of sense of urgency or need for a prompt response. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

### **adam-work-team-manager**

Corporate messenger application chat between our fictional protagonist, Adam Macy and his manager Charlotte Walker. Adam works for Borchester Software (sometimes referred to by its initials) and is in a team of 10 people called the Client Technology Group (CTG), headed by Charlotte. The team is spread all over. This chat is just between Adam and his manager, and topics should include client discussions, logistics, and usual topics that a manager and their direct report might need to discuss. In this scenario, Adam is a strong performer at work, but has a high workload with a lot of customer travel commitments. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Some of the messages should convey some sort of sense of urgency or need for a prompt response. The tone of the messages should be informal British English with British spellings.

### **adam-work-team-client**

Corporate messenger application chat for the work team of our fictional protagonist, Adam Macy. The chat is workplace discussion between members of the same IT consulting team that Adam is in. The company that Adam works for is called Borchester Software (sometimes referred to by its initials) and he is in a team of 10 people called the Client Technology Group (CTG). The team is spread all over. Other chats exist for more general topics, but this chat is specifically about issues, questions and logistics relating to client work. For example, the team might talk about new client engagements, seek help with client technical problems or arrange meetings related to the clients. Some of the messages should convey some sort of sense of urgency or need for a prompt response. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal British English with British spellings.

### **S7.2.2 Adam - Personal**

#### **adam-family-group-chat**

General chat between members of a fictional British family, as seen by our protagonist, Adam Macy. The chat is a lighthearted exchange between family members on a variety of topics, with some messages relating to logistics and meeting up. Other topics might include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. Adam and Ian live in the county town

## S7 Synthetic Content Topics

---

of Borchester in Borsetshire, while family are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English with British spellings.

### **adam-friends-chat**

General chat between a group of friends of our protagonist, Adam Macy and his partner Ian Craig. The chat is a lighthearted exchange between friends on a variety of topics that a group of young to middle aged adult friends might discuss. This would include but is not limited to jokes, meeting arrangements, discussion about what people have been doing, world affairs and local gossip. Adam and Ian live in the county town of Borchester in Borsetshire, while friends are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. Chat participants should be limited to only Adam, Ian and the supplied list of friends (not any family or work colleagues), although messages may refer to any individual. The tone of the messages should be informal British English with some light banter using British spellings.

### **adam-ian-chat**

Chat between our fictional protagonist, Adam Macy and his civil partner Ian Craig. The chat is a exchange a variety of topics, with some messages relating to logistics and meeting up but also discussing how their day is going and talking about family and work events/gossip. Try and include reasonable variety. Adam is an IT consultant, who often works long hours and has to travel, which sometimes causes tension. Adam and Ian live in the county town of Borchester in Borsetshire, while family are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English with British spellings.

### **S7.2.3 Adam - Cycling**

#### **adam-cycling-club-general**

General chat between members of a fictional British cycling club the 'Borchester Wheelers' (also called BWCC). Club members mostly come from the fictional town of Borchester in Borsetshire, UK, or nearby. Topics generally relate to all aspects of cycling but in particular club rides, cycle sport, cycle commuting, and maintenance. For your reference, the club does its big cycle rides on a Sunday morning at 9am (people meet at the car park of a cafe called 'Spoke & Brew' as the club does not have a physical club-house). Popular destinations for club rides include the villages/towns of Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and

## S7 Synthetic Content Topics

---

Lakey Green, and sometimes as far afield as the cathedral city of Felpersham. Club rides could be of the sporting, social or coffee shop variety, and there is a monthly new joiners ride. Some members commute to work by bike and occasionally meet to do their commutes together. Doing laps of Borchester Park is a popular weekend and after work activity, although people often complain about the car traffic there. Many members also cycle on the virtual platform Zwift, although there are different chat groups dedicated to this. People will often discuss any professional cycle races that are happening at the time, as well as occasionally other sports such as rugby or football. People will sometimes chide members to stay on topic if they stray too far into other sports. Sometimes people will send messages reminding members to look at some topic posted on the club's web forum. There are some general jokes and light banter. The tone of the messages should be informal British English with British spellings.

### **adam-cycling-club-zwift**

Chat between members of a fictional British cycling club the 'Borchester Wheelers' (also called BWCC). Club members mostly come from the fictional town of Borchester in Borsetshire, UK, or nearby. This chat is specifically to discuss the virtual cycling platform Zwift, which many members use, and is not the place to discuss outdoor riding. Main topics for Zwift include the Zwift Racing League (ZRL) where the club fields several teams, equipment, technical issues related to Zwift, and nutrition/hydration strategies (beetroot juice is popular). Zwift racing is organised into categories A-C based on rider power/weight and this is often controversial, with people having been known to engage in 'sandbagging' so that they can compete in a lower category with more success. Club members often joke with each other in a lighthearted way about sandbagging. Other running jokes involve the excessive use of spreadsheets and the power or otherwise of various riders. Many messages relate to organising people into teams to compete. For your reference, the club does its big 'real life' cycle rides on Sundays - popular destinations for club rides include the villages/towns of Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green, and sometimes as far afield as the cathedral city of Felpersham. Sometimes non-Zwift messages appear in the chat, which is mostly tolerated. There are some general jokes and light banter. The tone of the messages should be informal British English with British spellings.

## S7.3 Phoebe

### S7.3.1 Phoebe - Work

#### **phoebe-work-team**

Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is general workplace discussion among members of Phoebe's immediate work team, consisting of her manager and fellow fast-track graduates. Topics should include discussions related to work in general but not specific client details. This can include client-related discussions such as requests for help, lunch arrangements, work social events and other topics that you might expect to be discussed in a team within a large London accountancy firm. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Phoebe lives in London, where the company has its main offices. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **phoebe-work-fasttrack**

Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is workplace discussion relating to the fast track programme that Phoebe is on. Members of the chat include fast track participants, mentors, and administrators. Topics should relate to the programme and include requests for advice, discussion of experiences, talk about job rotations, and social gatherings. Include in the messages some announcements from fast track leadership or administrators; these announcements should range from low importance to high importance/urgent. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Phoebe lives in London, where the company has its main offices. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **phoebe-work-manager**

Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is workplace discussion between Phoebe and her manager Megan Foster. Include only messages by Phoebe and Megan in this chat. Topics should be the usual ones you would expect an employee and manager to discuss. Messages should range from low importance to high importance/urgent. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Phoebe lives in London, where the company has its main offices; she sometimes needs to travel for work. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **phoebe-work-announce**

Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is company-wide internal announcements, with messages only coming from senior leadership (or their assistants) or an impersonal announcements account. Topics include client-related announcements (i.e. new clients, project successes), senior appointments, compliance-related communications, recognising individual successes etc. Messages should range from low importance to high importance/urgent. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be formal British business English with British spellings.

#### **S7.3.2 Phoebe - Personal**

##### **phoebe-friends-university**

General chat between a group of friends who all studied together at Borsetshire University and graduated two years ago. This close group of friends stays in touch - mainly via this chat - and sometimes organises meetings. They discuss how each other's careers and personal lives are doing. Topics are wide-ranging, including general jokes and light chat in addition to talking about what they are doing or arranging to meet. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

##### **phoebe-friends-work**

General chat between a small group of friends who currently work together at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is outside of normal work channels and can include arrangements to meet up and office gossip as well as general jokes and light chat. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

##### **phoebe-family-chat**

General chat between members of a fictional British family, as seen by our protagonist, Phoebe Aldridge. The chat is a lighthearted exchange between immediate family members Phoebe, Brian, Jennifer, Debbie, Adam, Kate and Alice, and also wider family members. Do not include messages from non-family. Topics include anything a family might discuss, but can also include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. While Phoebe lives in London now, most of the family live in the village of Ambridge, while wider family and friends are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham where Usha works as a partner in Felpersham Law. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

## S7.4 Kenton

### S7.4.1 Kenton - Work

#### **kenton-work-company-announce**

Corporate messenger application messages within the employer of our fictional protagonist, Kenton Archer. The messages are exclusively internal corporate announcements coming from the executive and senior leadership team of company that Kenton works for, ICG. Generate messages that come from all of the executives in the provided list. Messages should occasionally refer to the company's international operations, R&D, production facilities etc. Refer to the countries that these are located in. Message content should be unique and not duplicated. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be formal business international English.

#### **kenton-work-company-general**

Corporate messenger application chat within the employer of our fictional protagonist, Kenton Archer. The chat is general workplace discussion within the company that Kenton works for, ICG. The chat is at the overall company level and not individual teams, so should mostly not refer to in-country topics. Topics should include discussion about company announcements and products, upcoming events, business climate, general news (where appropriate for corporate discussion) and other items people might choose to share with a company of this size. This group chat is intended for discussions aimed at the whole company - it is not the correct place for team discussions, technical support requests etc. Messages should occasionally refer to the company's international operations, R&D, production facilities etc. Refer to the countries that these are located in. Message content should be unique and not duplicated. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal/business international English.

#### **kenton-work-team-client**

Corporate messenger application chat for the work team of our fictional protagonist, Kenton Archer. The chat is client-related workplace discussion between members of the same business development (sales) team that Kenton is in within ICG. The team is spread all over. Topics should include client discussions, logistics, problem solving/assistance and other topics that a distributed team might discuss relating to their client base. Some of the messages should be of a type that might need a prompt or urgent response. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be informal British English.

#### **kenton-work-team-general**

Corporate messenger application chat for the work team of our fictional protagonist, Kenton Archer. The chat is general workplace discussion between members of the same business development (sales) team that Kenton is in within ICG. The team is spread all over. Topics should include client discussions, logistics, talking about out of work meetups and events and other topics that you might expect a distributed work team to discuss. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be informal British English.

#### **kenton-work-team-manager**

Corporate messenger application chat between our fictional protagonist, Kenton Archer and his manager Evelyn Baker, who heads up the business development (sales) team that Kenton is in within ICG. The team is spread all over. This chat is just between Kenton and his manager, and topics should include client discussions, logistics, and usual topics that a manager and their direct report might need to discuss. Some of the messages should be of a type that would normally need a prompt or urgent response. In this scenario, Kenton is a strong performer at work, but has a high workload with a lot of customer travel commitments. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal British English.

### **S7.4.2 Kenton - Personal**

#### **kenton-family**

Chat between members of the family of our protagonist, Kenton. The chat is a light-hearted exchange between family members on a variety of topics, with some messages relating to logistics and meeting up. A small amount of messages should be of a type that might need a prompt or urgent response. Kenton and Kathy live in the village of Ambridge, and work in the county town of Borchester in Dorsetshire, while family are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.

#### **kenton-friends**

Chat between a group of friends. People mostly come from the fictional village of Ambridge in the county of Dorsetshire, UK, or nearby. Some of the group often meet at the village pub (The Bull) to socialise. Chat will cover a wide range of topics, including village gossip, current affairs and family news, as well as making arrangements and sharing personal news. People may also mention locations in the main county town of Borchester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral

## S7 Synthetic Content Topics

---

city of Felpersham. There are some general jokes and light banter. The tone of the messages should be informal British English.

### **kenton-kathy**

Chat between our fictional protagonist, Kenton and his wife Kathy Archer. The chat is a exchange a variety of topics, with some messages relating to logistics and meeting up but also discussing how their day is going and talking about family and work events. Try and include reasonable variety. Kenton is a client manager for a global consumer products company, who often works long hours and has to travel, which sometimes causes tension. A small amount of messages should be of a type that might need a prompt or urgent response. Kenton and Kathy live in the village of Ambridge, and work in the county town of Borchester in Borsetshire, while family are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.

### **S7.4.3 Kenton - Football**

#### **kenton-football**

Chat between a group of friends who share an interest in football. People mostly come from the fictional village of Ambridge in the county of Borsetshire, UK, or nearby. The group often meet at the village pub (The Bull) to socialise and sometimes watch matches, but also watch the local team and some travel to league matches. Topics generally relate football, but the chat will often veer to other topics including village gossip, current affairs and family news. As well as the football clubs listed in the entities document, people may also mention locations in the main county town of Borchester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham. There are some general jokes and light banter. The tone of the messages should be informal British English.

### **S7.4.4 Kenton - Golf**

#### **kenton-golf**

Chat between a group of friends who play golf together. People mostly come from the fictional village of Ambridge in the county of Borsetshire, UK, or nearby. The group socialise together around golf, but also the village pub, The Bull. Topics generally relate to arranging meetups to play golf, but the chat will often veer to other topics including village gossip, current affairs and family news. As well as the golf clubs listed in the entities document, people may also mention locations in the main county town of Borchester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham. There are some general jokes and light banter. The tone of the messages should be informal British English.

## **S7.5 Usha**

### **S7.5.1 Usha - Work**

#### **usha-work-office**

Company messenger application chat at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. This chat is general workplace discussion among staff. Topics should include discussions related to work in general but not specific client details. This can include cover, logistical arrangements, daily office opening and closing, lunch arrangements, work social events and other topics that you might expect to be discussed in a small company. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **usha-work-partners**

Company messenger application chat at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. This chat is discussion between the three partners including Usha. Topics in this chat should include discussions related to the individuals work as partners of the practice. This can include the strategy and performance of the partnership, staff issues, problems or opportunities with clients, concern about competitors and other leadership topics. Discussion of client issues should be at a higher level rather than details, focussing only on things that might affect the client relationship - the partners have different areas of specialisation and would not generally discuss day to day client matters among themselves. As it is a private chat between the three partners they will be candid about issues and people. Some messages may concern social topics or trivia. A subset of the messages should be of the type that might expect a prompt response. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **usha-work-pa**

Company messenger application chat at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. This chat is discussion between Usha and her personal assistant Evan Thomas. Only include messages from these two people. Topics in this chat should include discussions related to Usha's client work, organising her diary and other similar items that would come up between a partner in a solicitor's practice and her personal assistant. Some messages may concern social topics or trivia. A subset of the messages should be of the type that might expect a prompt

response. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **usha-work-clients**

Business communications (email/messages) at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. These messages are communications directly between Usha and various clients that she is working with. Topics in this chat should include discussions related to the clients' cases and other similar items that would come up between a partner in a solicitor's practice and her clients. A few messages may concern social topics or trivia where Usha has a long relationship with the client. A subset of the messages should be of the type that might expect a prompt response. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. Ensure that you produce incoming messages from clients. The tone of the messages should be British English with British spellings.

### **S7.5.2 Usha - Personal**

#### **usha-family-close**

General chat between members of a fictional British family, as seen by our protagonist, Usha Franks. The chat is a lighthearted exchange between immediate family members Usha, Alan, Amy and Ravi - do not include messages from any other people. Amy and Ravi should refer to Usha and Alan as 'Mum' and 'Dad'. Topics include anything a family might discuss, but can also include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. Usha and Alan live in the village of Ambridge, while wider family and friends are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham where Usha works as a partner in Felpersham Law. The tone of the messages should be informal British English with British spellings.

#### **usha-family-wide**

General chat between members of a fictional British family, as seen by our protagonist, Usha Franks. The chat is a lighthearted exchange between immediate family members Usha, Alan, Amy and Ravi and wider family members including Deepak, Shiv and Satya - do not include messages from non-family. Topics include anything a family might discuss, but can also include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. Usha and Alan live in the village of Ambridge, while wider family and friends are dotted around local

## S7 Synthetic Content Topics

---

villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham where Usha works as a partner in Felpersham Law. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

### **usha-village**

General chat among residents of the fictional village of Ambridge in the county of Dorsetshire, UK, as seen by our protagonist, Usha Franks. The chat is a mostly light-hearted exchange between residents of Ambridge, covering topics including comings & goings around Ambridge, parking problems, postal service & parcel delivery (problems), recommendations for good tradespeople, social events, local crime and plenty of village gossip. This chat is among a general village audience rather than one-to-one messages about individual arrangements. A few of the messages should be of a nature that merits a prompt response, such as a request for assistance or information. Nearby villages include Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. Some of these locations should occasionally be mentioned in messages. In addition to the supplied list of people, you should make up identities as needed - the chat forum would contain maybe 50-30 members (although not all of these post messages often). Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spelling.

### **usha-school-comms**

Messages originating from the school Felpersham Grammar in Felpersham, Dorsetshire, UK, where our protagonist, Usha Franks sends her children. All messages are announcements and notices from school leaders or the school office, directed at parents with children at the school. In addition to the supplied list of people, you should make up identities as needed. Message content should be unique and not duplicated. The tone of the messages should be relatively formal British English with British spelling.

### **usha-school-parents-chat**

Messenger chat between parents with children attending Felpersham Grammar School in Felpersham, Dorsetshire, UK, where our protagonist, Usha Franks sends her children. This is a chat group independent of the school where parents can discuss school and other topics. This could include compliments or more often complaints about the school as well as sharing helpful information and coordinating things. Sometimes the parents group meets up in Felpersham cafes. There are some occasional jokes and off topic discussion. Some parents have a less positive outlook than others and tend to complain, while others are more positive - generate a mix of positive and negative comments. Only parents are in this chat - neither Usha's nor anyone else's children are on it. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spelling.

### **S7.5.3 Usha - Riding**

#### **usha-riding-chat**

General chat between a group of friends with a shared interest in horse riding, including our protagonist Usha Franks. Most of those on the chat are members of Felpersham Riding Club and some stable their horses there. Club members mostly come from the fictional villages of Ambridge and Darrington in the county of Borsetshire, or nearby, although some come from the cathedral city of Felpersham where Usha practices as a solicitor. This is a social chat between friends and acquaintances that share a common interest. Topics generally relate to all aspects of recreational and competitive horse riding, with many messages also being about meeting up for rides. Some but not all ride competitively (i.e. dressage), while most are social riders. There are some non-horse related messages such as general jokes and light chat. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

#### **usha-riding-stables**

General chat for people associated with Felpersham Riding Club and its stables, including our protagonist Usha Franks. The chat is used for communication between members of Felpersham Riding Club and also for announcements and information from club officials and stable managers. Usha stables her horse there. Club members mostly come from the fictional villages of Ambridge and Darrington in the county of Borsetshire, or nearby, although some come from the cathedral city of Felpersham where Usha practices as a solicitor. Topics generally relate to all aspects of recreational and competitive horse riding, with many messages also being about club matters and announcements about competitions, events and happenings at the stables. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.

## **Supplement S8**

### **Synthetic Evaluation Schemas & Prompting**

# Synthetic Evaluation Schemas & Prompting

This document contains details of prompting resources used in the Synthetic Evaluation process discussed in Chapter 8.

## Contents

- S8.1 – Schemas (Evaluation)
- S8.2 – Schemas (Data Service)
- S8.3 – Prompt Elements
- S8.4 – Examples
- S8.5 – Processor

## S8.1 Evaluation Schemas

These schemas are used for evaluation actions - i.e. requests to the evaluation LLM and the responses back. These schemas are also available at: [doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /eval/schemas].

### S8.1.1 Persona Schema

Listing S8.1 shows the Persona Schema, which is used as part of all evaluation requests. This schema is derived from our work on Persona development, as discussed in Section 5.2.

```
{  
  "$id": "https://parse.net/awag/0.1/persona.schema.json",  
  "$schema": "https://json-schema.org/draft/2020-12/schema",  
  "type": "object",  
  "title": "Awareness Agent Evaluation Persona",  
  "description": "Object representing a request to evaluate one or more items in the role of the  
    specified persona from the supplied perspectives",  
  "properties": {  
    "name": {  
      "title": "Persona Name",  
      "description": "The name given to this persona",  
      "type": "string"  
    },  
    "id": {  
      "title": "Persona ID",  
      "description": "The internal ID for this persona (ignore)",  
      "type": "string"  
    }  
  }  
}
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

},
"definition": {
    "title": "Evaluation Persona Definition",
    "description": "The definition of this persona",
    "type": "object",
    "properties": {
        "type": {
            "title": "Persona Type",
            "description": "Internal identifier of the persona type (ignore)",
            "type": "string"
        },
        "age": {
            "title": "Age",
            "description": "The age of the persona",
            "type": "integer"
        },
        "gender": {
            "title": "Gender",
            "description": "What gender does this persona identify as",
            "type": "string"
        },
        "does": {
            "title": "Does",
            "description": "What does this persona do in their work and personal life",
            "type": "string"
        },
        "feelThinkBelieve": {
            "title": "Feel/Think/Believe",
            "description": "What does this persona feel, think and believe; what is their outlook on
                life, priorities etc",
            "type": "string"
        },
        "technologyExperience": {
            "title": "Technology Experience",
            "description": "What experience does this persona have with technology",
            "type": "string"
        },
        "problems": {
            "title": "Problems",
            "description": "What problems does this persona encounter",
            "type": "string"
        },
        "needs": {
            "title": "Needs",
            "description": "What needs does this persona have",
            "type": "string"
        },
        "existingSolutions": {
            "title": "Existing Solutions",
            "description": "What solutions does this persona currently employ to address their
                problems and meet their needs",
            "type": "string"
        }
    }
},
"required": [
    "id",
    "name",
    "definition"
]
}

```

Listing S8.1: Evaluation Persona Schema

## S8 Synthetic Evaluation Schemas & Prompting

---

### S8.1.2 Perspective Schema

Listing S8.2 shows the Perspective Schema, which is used as part of Mode 1 evaluation requests.

```
{
  "$id": "https://parse.net/awag/0.1/perspective.schema.json",
  "$schema": "http://json-schema.org/2020-12/schema#",
  "title": "Awareness Agent Evaluation Perspective",
  "description": "Object representing an evaluation perspective.",
  "type": "object",
  "properties": {
    "perspectiveId": {
      "title": "Perspective ID",
      "description": "A string identifying the perspective",
      "type": "string"
    },
    "perspectiveName": {
      "title": "Perspective Name",
      "description": "The name of the perspective",
      "type": "string"
    },
    "perspectiveText": {
      "title": "Perspective Text",
      "description": "Textual description of the perspective to apply",
      "type": "string"
    }
  },
  "required": [
    "perspectiveId",
    "perspectiveName",
    "perspectiveText"
  ]
}
```

Listing S8.2: Evaluation Perspective Schema

### S8.1.3 Evaluation Request Schema - Mode 1

Listing S8.3 shows the Classification Evaluation Request Schema; this schema is only used with **Mode 1** evaluation requests as discussed in Section 8.2.3.

```
{
  "$id": "https://parse.net/awag/0.1/classification-evaluation-request.schema.json",
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "type": "object",
  "title": "Evaluation Request",
  "description": "Object representing a request to evaluate one or more items in the role of the specified persona from the supplied perspectives",
  "properties": {
    "persona": {
      "title": "Evaluation Persona",
      "description": "The persona to adopt when performing the evaluation",
      "type": "object",
      "properties": {
        "name": {
          "title": "Persona Name",

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        "description": "The name given to this persona",
        "type": "string"
    },
    "id": {
        "title": "Persona ID",
        "description": "The internal ID for this persona (ignore)",
        "type": "string"
    },
    "definition": {
        "title": "Evaluation Persona Definition",
        "description": "The definition of this persona",
        "type": "object",
        "properties": {
            "type": {
                "title": "Persona Type",
                "description": "Internal identifier of the persona type (ignore)",
                "type": "string"
            },
            "age": {
                "title": "Age",
                "description": "The age of the persona",
                "type": "integer"
            },
            "gender": {
                "title": "Gender",
                "description": "What gender does this persona identify as",
                "type": "string"
            },
            "does": {
                "title": "Does",
                "description": "What does this persona do in their work and personal life",
                "type": "string"
            },
            "feelThinkBelieve": {
                "title": "Feel/Think/Believe",
                "description": "What does this persona feel, think and believe; what is their outlook on life, priorities etc",
                "type": "string"
            },
            "technologyExperience": {
                "title": "Technology Experience",
                "description": "What experience does this persona have with technology",
                "type": "string"
            },
            "problems": {
                "title": "Problems",
                "description": "What problems does this persona encounter",
                "type": "string"
            },
            "needs": {
                "title": "Needs",
                "description": "What needs does this persona have",
                "type": "string"
            },
            "existingSolutions": {
                "title": "Existing Solutions",
                "description": "What solutions does this persona currently employ to address their problems and meet their needs",
                "type": "string"
            }
        }
    },
    "required": [
        "name", "id"
    ]
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        ],
    },
    "perspectives": {
        "title": "Perspectives",
        "description": "An array of perspectives to use when evaluating each item. Each perspective defines a way that the item should be evaluated. For each perspective, the persona should consider the evaluation from the angle defined by the perspective",
        "type": "array",
        "items": [
            {
                "title": "Perspective",
                "description": "An individual perspective form which to consider the evaluation",
                "type": "object",
                "properties": {
                    "id": {
                        "title": "Perspective ID",
                        "description": "A string identifying the perspective",
                        "type": "string"
                    },
                    "name": {
                        "title": "Perspective Name",
                        "description": "The name of the perspective",
                        "type": "string"
                    },
                    "text": {
                        "title": "Perspective Text",
                        "description": "Textual description of the perspective to apply",
                        "type": "string"
                    }
                },
                "required": [
                    "id", "name", "text"
                ]
            }
        ]
    },
    "items": {
        "title": "Evaluation Items",
        "description": "An array of items to evaluate. Each of these represents a set of classifications made by the software system that is being evaluated that must be assessed",
        "type": "array",
        "items": [
            {
                "title": "Evaluation Item",
                "description": "An item to evaluate.",
                "type": "object",
                "properties": {
                    "itemId": {
                        "title": "Item ID",
                        "description": "The ID of the item being evaluated",
                        "type": "string"
                    },
                    "sent": {
                        "title": "Sent Date Time",
                        "description": "The date and time that the item was sent or delivered. This should be used in any time-based perspective evaluations",
                        "type": "string",
                        "format": "date-time"
                    },
                    "title": {
                        "title": "Item Title",
                        "description": "The title or short description of the item being evaluated. Use this as part of the evaluation of whether the software system correctly classified the item",
                        "type": [

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        "string",
        "null"
    ],
},
"summary":{
    "title":"Item Summary",
    "description":"Summarised text of the item being evaluated. Use this as part of the
                  evaluation of whether the software system correctly classified the item",
    "type":"string"
},
"source":{
    "title":"Item Source",
    "description":"Description of where the item came from, which will vary according
                  to the type of thing that the item is",
    "type":"object",
    "properties":{
        "type":{
            "title":"Item Type",
            "description":"Short description of the type of the item being evaluated,
                          such as a Slack message or RSS item.",
            "type":[
                "string",
                "null"
            ]
        },
        "originator":{
            "title":"Item Originator",
            "description":"Information on who/what the item came from, which will vary by
                          item type. This could be the sender's email address, an RSS publisher,
                          Slack user etc.",
            "type":[
                "string",
                "null"
            ]
        },
        "channel":{
            "title":"Item Channel",
            "description":"Information on how/where the item was distributed, which will
                          vary by item type. For example, this could be a Slack channel, a mailing
                          list name, or an RSS feed etc.",
            "type":[
                "string",
                "null"
            ]
        }
    },
    "required": [
        "type",
        "originator",
        "channel"
    ]
},
"classifications": {
    "title": "Classifications",
    "description": "An array of classifications that the software system has applied to
                   the item. These are the classifications that should be evaluated for each
                   perspective using the supplied persona",
    "type": "array",
    "items": [
        {
            "type": "object",
            "properties": {
                "name": {
                    "title": "Classification Name",
                    "description": "The name of the classification. Do not use this in your
                                   prompt as it is used internally by the system"
                }
            }
        }
    ]
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        "evaluation as it is only an identifier",
        "type":"string"
    },
    "desc":{
        "title":"Classification Description",
        "description":"The description of the classification. Use this text to
            understand what type of classification this is and the criteria for
            selecting different values",
        "type":"string"
    },
    "selected":{
        "title":"Classification Selection",
        "description":"This is the value that you must evaluate. It is the
            classification value that the source system has chosen for the item
            from the list of available classifications. Note that a value of 'none',
            'not' or null is a special case that does not need to be in
            the list of available classifications; these would mean that the
            system decided to classify the item as not being relevant for this
            classification, which should also be taken into consideration in the
            evaluation",
        "type":"string"
    },
    "available":{
        "title":"Available Classifications",
        "description":"The classification values that were available to the
            system to apply to this classification for this item. Use this to
            understand the context of the chosen classification by examining the
            possible alternatives",
        "type":"array",
        "items":[
            {
                "type":"string"
            }
        ]
    },
    "required":[
        "name",
        "selected",
        "available"
    ]
}
],
"required":[
    "itemId",
    "sent",
    "title",
    "summary",
    "source",
    "classifications"
]
}
],
"required":[
    "persona",
    "perspectives",
    "items"
]
}
}

```

Listing S8.3: Classification Evaluation Request Schema (Mode 1)

## S8 Synthetic Evaluation Schemas & Prompting

---

### S8.1.4 Evaluation Request Schema - Modes 2 & 3

Listing S8.4 shows the Alternative Classification Evaluation Request Schema; **Mode 2** and **Mode 3** evaluation requests conform to this schema, although it is never used directly in the evaluation process.

```
{
  "$id": "https://parse.net/awag/0.1/classification-evaluation-request.alt.schema.json",
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "type": "object",
  "title": "Evaluation Request",
  "description": "Object representing a request to evaluate one or more items in the role of the specified persona from the supplied perspectives",
  "properties": {
    "persona": {
      "title": "Evaluation Persona",
      "description": "The persona to adopt when performing the evaluation",
      "type": "object",
      "properties": {
        "name": {
          "title": "Persona Name",
          "description": "The name given to this persona",
          "type": "string"
        },
        "id": {
          "title": "Persona ID",
          "description": "The internal ID for this persona (ignore)",
          "type": "string"
        },
        "definition": {
          "title": "Evaluation Persona Definition",
          "description": "The definition of this persona",
          "type": "object",
          "properties": {
            "type": {
              "title": "Persona Type",
              "description": "Internal identifier of the persona type (ignore)",
              "type": "string"
            },
            "age": {
              "title": "Age",
              "description": "The age of the persona",
              "type": "integer"
            },
            "gender": {
              "title": "Gender",
              "description": "What gender does this persona identify as",
              "type": "string"
            },
            "does": {
              "title": "Does",
              "description": "What does this persona do in their work and personal life",
              "type": "string"
            },
            "feelThinkBelieve": {
              "title": "Feel/Think/Believe",
              "description": "What does this persona feel, think and believe; what is their outlook on life, priorities etc",
              "type": "string"
            },
            "technologyExperience": {
              "title": "Technology Experience",
              "type": "string"
            }
          }
        }
      }
    }
  }
}
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        "description":"What experience does this persona have with technology",
        "type":"string"
    },
    "problems":{
        "title":"Problems",
        "description":"What problems does this persona encounter",
        "type":"string"
    },
    "needs":{
        "title":"Needs",
        "description":"What needs does this persona have",
        "type":"string"
    },
    "existingSolutions":{
        "title":"Existing Solutions",
        "description":"What solutions does this persona currently employ to address their
                     problems and meet their needs",
        "type":"string"
    }
}
},
"required":[
    "id",
    "name",
    "definition"
]
},
"perspective":{
    "title":"Perspective",
    "description":"The perspective defines a way that the item should be evaluated - the persona
                  should consider the evaluation from the angle defined by the perspective",
    "type":"string"
},
"items":{
    "title":"Items",
    "description":"An array of items to be evaluated",
    "type":"array",
    "items":{
        "type":"object",
        "properties":{

            "id":{
                "title":"Item Identification",
                "description":"Unique identifiers for the item and classification",
                "type":"object",
                "properties":{
                    "item_id":{
                        "title":"Item ID",
                        "description":"The ID of the item being evaluated",
                        "type":"string"
                    },
                    "perspective_id":{
                        "title":"Perspective ID",
                        "description":"A string identifying the perspective",
                        "type":"string"
                    },
                    "classification_id":{
                        "title":"Classification ID",
                        "description":"Identifier for the specific classification being evaluated",
                        "type":"string"
                    }
                }
            },
            "required":[
                "item_id",
                "perspective_id",

```

## S8 Synthetic Evaluation Schemas & Prompting

```

        "classification_id"
    ]
},
"content": {
    "title": "Content",
    "description": "The content of the item being evaluated",
    "type": "string"
},
"classification": {
    "title": "Classification",
    "description": "Details of the classification that was applied to the item",
    "type": "object",
    "properties": {
        "description": {
            "title": "Description",
            "description": "Explanation of the classification's criteria",
            "type": "string"
        },
        "classified_as": {
            "title": "Classified As",
            "description": "This is the value that must be evaluated. It is the classification value that the source system has chosen for the item from the list of available classifications. Note that a value of 'none', 'not' or null is a special case that does not need to be in the list of available classifications; these would mean that the system decided to classify the item as not being relevant for this classification, which should also be taken into consideration in the evaluation",
            "type": "string"
        },
        "from_available_classifications": {
            "title": "Available Classifications",
            "description": "The classification values that were available to the system to apply to this classification for this item",
            "type": "array",
            "items": {
                "type": "string"
            }
        }
    },
    "required": [
        "description",
        "classified_as",
        "from_available_classifications"
    ]
},
"required": [
    "id",
    "content",
    "classification"
]
}
},
"required": [
    "persona",
    "perspective",
    "items"
]
}
}
```

Listing S8.4: Classification Evaluation Request Schema (Mode 2/3)

## S8 Synthetic Evaluation Schemas & Prompting

---

### S8.1.5 Evaluation Result Schema - Mode 1

Listing S8.5 shows the Classification Evaluation Result Schema for **Mode 1** evaluations. This schema defines the `get_evaluations` function passed to the OpenAI chat completions call using the Function Calling technique<sup>1</sup>.

Note that `classificationName` in the schema is actually what we refer to as the Classification ID.

```
{
  "$id": "https://parse.net/awag/0.1/classification-evaluation-result.model.schema.json",
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "name": "get_evaluations",
  "description": "Return a set of evaluations of the provided input data",
  "parameters": {
    "type": "object",
    "properties": {
      "resultItems": {
        "type": "array",
        "items": {
          "type": "object",
          "title": "Evaluation results",
          "description": "Array containing all of the evaluation results for the provided data",
          "properties": {
            "itemId": {
              "title": "Item ID",
              "description": "The ID of the item being evaluated",
              "type": "string"
            },
            "evaluations": {
              "type": "array",
              "items": {
                "type": "object",
                "properties": {
                  "classificationName": {
                    "title": "Evaluated Classification Name",
                    "description": "The name of the classification that this evaluation is for",
                    "type": "string"
                  },
                  "perspectives": {
                    "title": "Evaluated Perspectives",
                    "description": "Array containing all of the evaluation results for this item, with one entry per perspective evaluated",
                    "type": "array",
                    "items": {
                      "type": "object",
                      "title": "Evaluated Perspective",
                      "description": "The evaluation result for the evaluated classification from a specific perspective",
                      "properties": {
                        "perspectiveId": {
                          "title": "Evaluated Perspective ID",
                          "description": "The ID of the perspective that this evaluation is for",
                          "type": "string"
                        },
                        "evaluatedSelection": {
                          "title": "Evaluated Classification Selection",
                          "type": "string"
                        }
                      }
                    }
                  }
                }
              }
            }
          }
        }
      }
    }
  }
}
```

<sup>1</sup>[https://cookbook.openai.com/examples/how\\_to\\_call\\_functions\\_with\\_chat\\_models](https://cookbook.openai.com/examples/how_to_call_functions_with_chat_models) [<https://perma.cc/FY7T-WUNA>]

## S8 Synthetic Evaluation Schemas & Prompting

```
        "description":"The classification selection chosen by the system,  
        that the evaluation has been applied to",  
        "type":"string"  
    },  
    "evaluationLikert":{  
        "title":"Evaluation Likert Scale Value",  
        "description":"A likert scale ranging from 1 (not at all in  
        agreement with the item evaluated in the perspective) to 5 (completely  
        in agreement). Values must only be in the range 1-5",  
        "type":"integer",  
        "minimum":1,  
        "maximum":5  
    },  
    "evaluationText":{  
        "type":"string",  
        "title":"Evaluation Text",  
        "description":"Textual description elaborating on why the  
        evaluationLikert value was chosen for this item"  
    }  
},  
"required": [  
    "perspectiveId",  
    "evaluatedSelection",  
    "evaluationLikert",  
    "evaluationText"  
]  
}  
},  
"required": [  
    "classificationName",  
    "perspectives"  
]  
}  
},  
"required": [  
    "itemId",  
    "evaluations"  
]  
}  
},  
}  
}
```

Listing S8.5: Classification Evaluation Result Schema - Mode 1

## S8 Synthetic Evaluation Schemas & Prompting

---

### S8.1.6 Evaluation Result Schema - Mode 2

Listing S8.6 shows the Classification Evaluation Result Schema for **Mode 2** evaluations. This schema defines the Mode 2 version of the `get_evaluations` function.

The Mode 2 result reflects the simpler nature of the Mode 2 and 3 request types compared to Mode 1. While Mode 1 has nested result arrays for the multiple Perspectives and Classifications passed in the request, Modes 2 & 3 results are a flatter array of items only, reflecting that only one Perspective & Classification are evaluated per item in those modes.

```
{
  "$id": "https://parse.net/awag/0.1/classification-evaluation-result.mode2.schema.json",
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "name": "get_evaluations",
  "description": "Return a set of evaluations of the provided input data",
  "parameters": {
    "type": "object",
    "properties": {
      "resultItems": {
        "type": "array",
        "items": {
          "type": "object",
          "title": "Evaluation Result",
          "description": "Object containing the evaluation result for a given item",
          "properties": {
            "id": {
              "type": "object",
              "properties": {
                "itemId": {
                  "type": "string",
                  "description": "The ID of the item being evaluated"
                },
                "perspectiveId": {
                  "type": "string",
                  "description": "The ID of the perspective that this evaluation is for"
                },
                "classificationId": {
                  "type": "string",
                  "description": "The ID of the classification that this evaluation is for"
                }
              },
              "required": [
                "itemId",
                "perspectiveId",
                "classificationId"
              ]
            },
            "evaluatedSelection": {
              "title": "Evaluated Classification Selection",
              "description": "The classification selection chosen by the system, that the evaluation has been applied to",
              "type": "string"
            },
            "evaluationLikert": {
              "type": "integer",
              "minimum": 1,
              "maximum": 5,
              "title": "Evaluation Likert Scale Value",
              "description": "A likert scale ranging from 1 (not at all in agreement with the item"
            }
          }
        }
      }
    }
  }
}
```

## S8 Synthetic Evaluation Schemas & Prompting

```
        classification) to 5 (absolutely in agreement). Values must only be in the
        range 1-5"
    },
    "evaluationText":{
        "type":"string",
        "title":"Evaluation Text",
        "description":"Textual description elaborating on why the evaluationLikert value
                      was chosen for this item"
    }
},
"required":[
    "id",
    "evaluatedSelection",
    "evaluationLikert",
    "evaluationText"
]
}
},
"required":[
    "resultItems"
]
}
}
```

Listing S8.6: Classification Evaluation Result Schema - Mode 2

### S8.1.7 Evaluation Result Schema - Mode 3

Listing S8.7 shows the Classification Evaluation Result Schema for **Mode 3** evaluations. This schema defines the Mode 3 version of the `get_evaluations` function.

Mode 3 differs from mode 2 only in that it contains a binary `evaluationAgreement` value rather than the `evaluationLikert` rating returned by modes 1 & 2.

```
{  
    "$id": "https://parse.net/awag/0.1/classification-evaluation-result.mode3.schema.json",  
    "$schema": "https://json-schema.org/draft/2020-12/schema",  
    "name": "get_evaluations",  
    "description": "Return a set of evaluations of the provided input data",  
    "parameters": {  
        "type": "object",  
        "properties": {  
            "resultItems": {  
                "type": "array",  
                "items": {  
                    "type": "object",  
                    "title": "Evaluation Result",  
                    "description": "Object containing the evaluation result for a given item",  
                    "properties": {  
                        "id": {  
                            "type": "object",  
                            "properties": {  
                                "itemId": {  
                                    "type": "string",  
                                    "description": "The ID of the item being evaluated"  
                                },  
                                "perspectiveId": {  
                                    "type": "string",  
                                    "description": "The ID of the perspective used for evaluation"  
                                }  
                            }  
                        }  
                    }  
                }  
            }  
        }  
    }  
}
```

## S8 Synthetic Evaluation Schemas & Prompting

```
        "description":"The ID of the perspective that this evaluation is for"
    },
    "classificationId":{
        "type":"string",
        "description":"The ID of the classification that this evaluation is for"
    }
},
"required":[
    "itemId",
    "perspectiveId",
    "classificationId"
]
},
"evaluatedSelection":{
    "title":"Evaluated Classification Selection",
    "description":"The classification selection chosen by the system, that the
        evaluation has been applied to",
    "type":"string"
},
"evaluationAgreement":{
    "type":"string",
    "enum":[
        "AGREE",
        "DISAGREE"
    ],
    "title":"Evaluation Agreement",
    "description":"Binary evaluation agreement value indicating whether the evaluator
        agrees or disagrees with the item classification."
},
"evaluationText":{
    "type":"string",
    "title":"Evaluation Text",
    "description":"Textual description elaborating on the evaluationAgreement value
        chosen for this item"
}
},
"required":[
    "id",
    "evaluatedSelection",
    "evaluationAgreement",
    "evaluationText"
]
},
"resultItems"
]
}
}
```

Listing S8.7: Classification Evaluation Result Schema - Mode 3

## S8.2 Evaluation Schemas - Data Service

These schemas are used for communication with awagdata - for example to store/retrieve evaluation results.

## S8 Synthetic Evaluation Schemas & Prompting

---

### S8.2.1 awagdata Record Evaluation Data

Listing S8.8 shows the Record Evaluation Data Schema, which describes the content of the POST request to awagdata (/data/eval) used to store the results of an evaluation query.

```
{
  "$id": "https://parse.net/awag/0.1/record-evaluation-data.schema.json",
  "$schema": "http://json-schema.org/2020-12/schema#",
  "title": "AwAg Evaluation Data Recording",
  "description": "Contains the results of the evaluation of one or more Content Items for a given evaluationPersona. This data is intended to be submitted to the awagdata /eval/record-evaluation-data service",
  "type": "object",
  "properties": {
    "agent": {
      "title": "Agent ID",
      "description": "Agent ID for the agent performing the operation",
      "type": "string"
    },
    "evaluationPersona": {
      "$ref": "https://parse.net/awag/0.1/persona.schema.json"
    },
    "evaluationPerspectives": {
      "title": "Evaluation Perspectives",
      "description": "The list of perspectives used for evaluating items",
      "type": "array",
      "items": {
        "$ref": "https://parse.net/awag/0.1/perspective.schema.json"
      }
    },
    "items": {
      "title": "Evaluated Items",
      "description": "An array of completed evaluations for Content Items",
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "itemId": {
            "title": "Content Item ID",
            "type": "string"
          },
          "tags": {
            "title": "Tags",
            "description": "The Output Tags to apply to this Content Item evaluation",
            "type": "array",
            "items": {
              "type": "string"
            }
          },
          "classifications": {
            "title": "Classifications",
            "description": "The classifications that are associated with this Content Item - these are what was evaluated",
            "type": "array",
            "items": {
              "type": "object",
              "properties": {
                "title": "Classification Name",
                "description": "The name of the classification",
                "classificationName": {
                  "type": "string"
                }
              }
            }
          }
        }
      }
    }
  }
}
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

},
"classificationValue": {
    "title": "Classification Value",
    "description": "The selected value of the classification that was evaluated",
    "type": "string"
},
"classificationOptions": {
    "title": "Classification Options",
    "description": "List of the available values that this classification could have had",
    "type": "array",
    "items": {
        "type": "string"
    }
},
"required": [
    "classificationName",
    "classificationValue",
    "classificationOptions"
]
},
"evaluateSourceType": {
    "title": "Evaluate Item Source Type",
    "description": "The source type for this evaluated item - such as 'RSS Item'",
    "type": "string"
},
"evaluateSourceOriginator": {
    "title": "Evaluate Item Source Originator",
    "description": "The originator of this evaluated item - such as an RSS provider",
    "type": "string"
},
"evaluateSourceChannel": {
    "title": "Evaluate Item Source Channel",
    "description": "The channel or other subdivision that this evaluated item was published to",
    "type": "string"
},
"evaluateTimeText": {
    "title": "Evaluate Time Text",
    "description": "An ISO 8601 compliant textual representation of the datetime of publication of the evaluated item",
    "type": "string"
},
"evaluateTimeMs": {
    "title": "Evaluate Time MS",
    "description": "The datetime of publication of the evaluated item expressed as milliseconds since the epoch",
    "type": "number"
},
"evaluateTitle": {
    "title": "Evaluate Title",
    "description": "The Title property of the evaluated item - this is content used for classification that was part of the evaluation",
    "type": "string"
},
"evaluateText": {
    "title": "Evaluate Text",
    "description": "The Text (body/content) property of the evaluated item - this is content used for classification that was part of the evaluation",
    "type": "string"
},
"evaluationResponses": {
    "title": "Evaluation Responses",
    "description": "The evaluation results for this item, organised by Perspective nested within Classification"
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

"type": "array",
"items": {
  "type": "object",
  "properties": {
    "classificationName": {
      "title": "Evaluated Classification Name",
      "description": "The name of the classification",
      "type": "string"
    },
    "perspectives": {
      "type": "array",
      "items": {
        "title": "Evaluated Perspective",
        "description": "The evaluation results by Perspective for this Classification",
        "type": "object",
        "properties": {
          "perspectiveId": {
            "title": "Perspective ID",
            "description": "The ID of the Perspective used for this evaluation",
            "type": "string"
          },
          "evaluationLikertVal": {
            "title": "Evaluation Likert Value",
            "description": "The evaluation result as a Likert value (modes 1 & 2)",
            "minimum": 1,
            "maximum": 5,
            "type": "integer"
          },
          "evaluationLikertText": {
            "title": "Evaluation Likert Text",
            "description": "The textual version of the evaluation Likert value",
            "type": "string"
          },
          "evaluationAgreement": {
            "title": "Evaluation Agreement",
            "description": "The evaluation result as an AGREE/DISAGREE string (mode 3)",
            "enum": [
              "AGREE",
              "DISAGREE"
            ],
            "type": "string"
          },
          "evaluationText": {
            "title": "Evaluation Text",
            "description": "A textual version of the evaluation",
            "type": "string"
          }
        }
      },
      "required": [
        "perspectiveId",
        "evaluationText"
      ],
      "oneOf": [
        {
          "required": [
            "evaluationAgreement"
          ]
        },
        {
          "required": [
            "evaluationLikertVal",
            "evaluationLikertText"
          ]
        }
      ]
    }
  }
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        }
    },
},
"required":[
    "classificationName",
    "perspectives"
]
}
},
"required":[
    "itemId",
    "tags",
    "classifications",
    "evaluateSourceType",
    "evaluateSourceOriginator",
    "evaluateSourceChannel",
    "evaluateTimeText",
    "evaluateTimeMs",
    "evaluateText",
    "evaluationResponses"
]
}
},
"context":{
    "title":"Context",
    "description":"Additional contextual information to be recorded alongside the evaluation",
    "type":"object",
    "properties":{
        "contextId":{
            "title":"Context ID",
            "description":"A unique identifier for this evaluation job",
            "type":"string"
        },
        "evaluationItems":{
            "title":"Evaluation Items",
            "description":"An array of the evaluated items in their fuller native format containing full
                information, provided for general information or to allow the evaluation to be re-run with
                a differently constructed request",
            "type":"array",
            "items":{
                "type":"object"
            }
        },
        "openaiApiCompletionsQueryInfo":{
            "title":"OpenAI Completions Query Information",
            "description":"Details of the OpenAI Completions API query, including parameters and full
                prompt information, as well as the response state",
            "type":"object",
            "properties":{
                "maxTokens":{
                    "title":"Max Tokens",
                    "description":"Parameter: maximum tokens limit for query",
                    "type":"integer"
                },
                "model":{
                    "title":"Model",
                    "description":"Parameter: model to use for query",
                    "type":"string"
                },
                "temperature:{
                    "title":"Temperature",
                    "description":"Parameter: temperature",
                    "type":"number"
                }
            }
        }
    }
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

    "promptText": {
        "title": "Prompt Text",
        "description": "The full text part of the prompt supplied to the API",
        "type": "string"
    },
    "promptQueryJson": {
        "title": "Prompt Query JSON",
        "description": "The full prompt supplied to the API in JSON format",
        "type": "object"
    },
    "queryState": {
        "title": "Query State",
        "description": "The query state taken from the OpenAI response, including response code,
                      actual token usage etc.",
        "type": "object"
    }
},
"required": [
    "contextId",
    "evaluationItems",
    "openaiApiCompletionsQueryInfo"
]
},
"required": [
    "agent",
    "evaluationPersona",
    "evaluationPerspectives",
    "items",
    "context"
]
}
}

```

Listing S8.8: Record Evaluation Data Schema

### S8.2.2 awagdata Record Evaluation Failure

Listing S8.9 shows the Record Evaluation Failure Schema, which describes the content of the POST request to awagdata (/data/eval) used to store the results of a **failed** evaluation query. This contains the response in raw, unparsed text format, in case invalid JSON was returned from the API, and the IDs of Content Items in the evaluation query (bearing in mind that multiple CI's could be in any one query).

```

{
    "$id": "https://parse.net/awag/0.1/record-evaluation-failure.schema.json",
    "$schema": "http://json-schema.org/2020-12/schema#",
    "title": "AwAg Evaluation Failure Recording",
    "description": "Contains the details of an evaluation failure, including the query state, context, and
                   relevant persona information.",
    "type": "object",
    "properties": {
        "agent": {
            "title": "Agent ID",
            "type": "string",
            "description": "Agent ID for the agent performing the operation"
        },
        ...
    }
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

"queryState": {
    "title": "Query State",
    "description": "The query state taken from the OpenAI response, including response code, finish reason, actual token usage etc.",
    "type": "object"
},
"maxTokens": {
    "title": "Max Tokens",
    "type": "integer"
},
"model": {
    "title": "Model",
    "type": "string",
    "description": "The OpenAI model used for the query"
},
"promptQueryJson": {
    "title": "Prompt Query JSON",
    "description": "The full prompt supplied to the API in JSON format",
    "type": "object"
},
"contextIdentifier": {
    "title": "Context ID",
    "type": "string",
    "description": "Unique Context ID for the query"
},
"evaluationPersona": {
    "$ref": "https://parse.net/awag/0.1/persona.schema.json"
},
"evaluationPerspectives": {
    "title": "Evaluation Perspectives",
    "description": "The list of perspectives used for evaluating items",
    "type": "array",
    "items": {
        "$ref": "https://parse.net/awag/0.1/perspective.schema.json"
    }
},
"responseText": {
    "title": "Response Text",
    "type": "string",
    "description": "The raw response text received back from OpenAI - this may or may not be parseable JSON"
},
"items": {
    "title": "Evaluated Item IDs",
    "type": "array",
    "items": {
        "type": "string"
    },
    "description": "List of Content Item IDs involved in the query"
},
"itemCount": {
    "title": "Item Count",
    "type": "integer",
    "description": "Count of Content Items in the query"
},
"statusCode": {
    "title": "Status Code",
    "type": "integer",
    "description": "HTTP status code of the response"
}
},
"required": [
    "agent",
    "queryState",
    "maxTokens",
    "model"
]
}

```

## S8 Synthetic Evaluation Schemas & Prompting

```
"model",
"promptQueryJson",
"contextIdentifier",
"evaluationPersona",
"evaluationPerspectives",
"responseText",
"rawResponseText",
"items",
"itemCount",
"statusCode"
]
}
```

Listing S8.9: Record Evaluation Failure Schema

## S8.3 Static Prompt Elements

This section contains the static text elements that make up evaluation prompts for modes 1, 2 & 3. The prompt text is made up of common and mode-specific structures which are assembled into System and User messages for OpenAI as described in Section 8.3.3.

### S8.3.1 System Message – Common

This **system** message element is used for all modes as the common system message for all Awareness Agent evaluation calls. It is worded in such a way to be compatible with the different modes, by avoiding mode-specific phrasing.

#### **AWAG\_SYSTEM\_MESSAGE\_COMMON**

You are the AwAg Evaluator. Your job is to role play a persona, for the purpose of evaluating a software system.

The software system that you are evaluating is intended to manage a user's incoming information from multiple sources - including work and personal - so that the user is not overwhelmed or distracted. Consider it a personal information triage service. Your evaluation should focus on how well this system classifies the content that it processes and you will be asked to indicate your agreement or disagreement with the decisions.

An AwAg Evaluation Request is presented as a structured JSON document, and its main elements are:

1. Persona - the definition or ID of the persona that you should adopt when evaluating the request;
2. Perspective - the way in which you should consider the items in your evaluation;
3. Items - the actual items that you should evaluate.

## S8 Synthetic Evaluation Schemas & Prompting

---

The exact structure of the AwAg Evaluation Request may vary, but you will be given guidance on how to interpret it. You will also be told how to respond for each request. Possible response types include binary agree/disagree, or using a likert scale from 1 (completely disagree) to 5 (completely agree) as well as text describing your evaluation.

You will be asked to respond with your evaluation(s) of item(s) in a structured way. Your evaluation result should include your own evaluation, and also the value of the evaluated classification value or selection.

### S8.3.2 System Message – Extra for Mode 1

This **system** message is added to Mode 1 requests and is used to reference the Evaluation Request Schema [S8.1.3] that Mode 1 requests conform to.

#### **AWAG\_SYSTEM\_MESSAGE\_EXTRA\_MODE1**

Information about the persona that you should adopt and the items that you must evaluate is provided as a JSON document - called the AwAg Evaluation Request JSON Schema (this schema has ID '<https://parse.net/awag/0.1/classification-evaluation-request.schema.json>'). Sometimes you may be given only a subset of this document to evaluate or asked to respond in a different way. Evaluate every item and classification combination, returning a result for each one.

### S8.3.3 System Message – Extra for Modes 2 & 3

This **system** message is added to Mode 2 and Mode 3 requests and is used to describe the request structure in a narrative textual rather than schema-based manner. While it is named '**AWAG\_SYSTEM\_MESSAGE\_EXTRA\_MODE2**', the same text is used for both modes 2 and 3 requests.

#### **AWAG\_SYSTEM\_MESSAGE\_EXTRA\_MODE2**

Information about the persona that you should adopt is provided as a JSON document. The 'definition' part of the document tells you about the age and gender of the persona, as well as other information about them. The 'does' property tells you what they do, both for work and social/personal activities. The 'feelThinkBelieve' parameter tells you what the persona feels, thinks and believes, telling you about their motivations and opinions. The 'technologyExperience' property tells you about the experience this persona has with technology. The 'problems' property tells you what sort of problems the persona might encounter, specifically in relation to managing information overload. The 'needs' and 'existingSolutions' properties tell you about the solution needs this persona has, and what existing solutions they have to try and address these.

## S8 Synthetic Evaluation Schemas & Prompting

---

Use the ‘perspective’ element of the request to tell you what approach to take to evaluating the items. For example, the perspective might ask you to determine if an item has been correctly classified.

The items to evaluate are provided to you as an array in the ‘items’ property of the request. Each item contains the following important information: ‘id’ contains the identifiers needed to uniquely identify this item, which you should use in your response for identification only. The ‘content’ property contains the actual textual content of the item that you should evaluate. The ‘classification’ property contains a JSON object that describes how the system that you are evaluating has classified the item - this contains the description of the classification, the available options that the system had to choose from (fromAvailableClassifications), and the option that it selected (classifiedAs). When evaluating an item, you should consider how well the selected classification matches the content of the item, and whether one of the alternative available options would have been a better fit.

### S8.3.4 User Message – Common Likert

This **user** message is added to all evaluation calls that require a Likert type response: specifically Mode 1 and Mode 2.

#### EVALUATION\_USER\_MESSAGE\_COMMON\_LIKERT

In this case, your evaluation should take the form of a Likert scale with a value range of 1-5. This describes how much your persona agrees with the classification that you are evaluating. If you strongly agree with how the system classified the item, you should use a likert value of 5. If you strongly disagree with the system’s classification, you should use a likert value of 1. You should use values between this when you can’t be as sure; for example a likert value of 4 indicates that you mostly agree with the classification chosen by the system you are evaluating. On the other hand, a value of 2 indicates that you would probably choose a different classification. If you don’t have enough information to decide either way or are otherwise unsure, you should choose a value of 3. Your evaluation must also include explanatory text written as your persona. It is important that you use the Likert scale consistently and use the full range as needed, and that the scale value you choose is consistent with the text of your evaluation. You must only consider whether the selected classification is the most appropriate of the listed available classifications - do not suggest any alternative classifications that are not listed.

### S8.3.5 User Message Prefix – Mode 1

This is the first **user** message added to all Mode 1 evaluation calls. Because Mode 1 uses Likert responses, it is followed by [EVALUATION\\_USER\\_MESSAGE\\_COMMON\\_LIKERT](#).

## S8 Synthetic Evaluation Schemas & Prompting

---

### **EVALUATION\_USER\_MESSAGE\_MODE1**

Please evaluate the following AwAg Evaluation Request. You should evaluate each item from each specified perspective, and return the results in the required result schema. For each item classification, consider the content, persona and perspectives. Be sure to evaluate the item for every classification and perspective combination.

### **S8.3.6 User Message Prefix – Mode 2**

This is the first **user** message added to all Mode 2 evaluation calls. Because Mode 2 uses Likert responses, it is followed by **EVALUATION\_USER\_MESSAGE\_COMMON\_LIKERT**.

### **EVALUATION\_USER\_MESSAGE\_MODE2**

Please evaluate the following AwAg Evaluation Request. You should evaluate each item for the specified persona and perspective, and return the results in the required result schema. Consider how item has been classified in the context of the persona and perspective.

### **S8.3.7 User Message Prefix – Mode 3**

This is the first **user** message added to all Mode 3 evaluation calls (which do not expect a Likert response).

### **EVALUATION\_USER\_MESSAGE\_MODE3**

Please evaluate the following AwAg Evaluation Request. You should evaluate each item for the specified persona and perspective, and return the results in the required result schema. Consider how item has been classified in the context of the persona and perspective.

In this case, rate whether you as your persona AGREE or DISAGREE with the selected evaluation based on the list of available classifications, returning this in the evaluationAgreement property.

### **S8.3.8 User Message Example – Mode 2**

This **user** message gives examples of Mode 2 requests and responses to guide the LLM how to interpret these and construct responses. Only modes 2 and 3 have this type of example provided, with Mode 1 relying on a more schema-based approach.

### **EVALUATION\_USER\_MESSAGE\_EXAMPLE\_MODE2**

Here are some examples to help you understand how to respond. This is an example request:

## S8 Synthetic Evaluation Schemas & Prompting

---

```
{
  "id": {
    "itemId": "214ea6fa-874d-44ae-a931-6561e69cedff",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "content": "2024-04-06T08:49:20.791998\nWho's going to the pub after work today?",
  "classification": {
    "description": "Does this item relate to my work or personal life?",
    "classifiedAs": "work",
    "fromAvailableClassifications": [
      "work",
      "personal"
    ]
  }
}
```

In this case, the content that the classification is for is: “Who’s going to the pub after work today?”. The classification that this request relates to is “Does this item relate to my work or personal life?”, which has two available classifications: ‘work’ and ‘personal’. The system that you are evaluating has classified this as ‘work’. In this case, even though the content mentions work, the topic is socialising outside of work, so the better classification would be ‘personal’. The appropriate response to this request would therefore look like:

```
{
  "id": {
    "itemId": "214ea6fa-874d-44ae-a931-6561e69cedff",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "evaluatedSelection": "work",
  "evaluationLikert": 2,
  "evaluationText": "This message is about going to the pub after work, which mentions work but is generally a personal topic. Therefore, I mostly disagree with the classification of 'work' - I would have chosen 'personal' instead on balance. I have awarded an evaluationLikert value of 2 to indicate disagreement."
}
```

Here is a second example:

```
{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "content": "2024-04-06T10:51:32.783943\nCan someone help me with this
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

report? I'm really struggling.",
"classification": {
    "description": "Does this item relate to my work or personal life?",
    "classifiedAs": "personal",
    "fromAvailableClassifications": [
        "work",
        "personal"
    ]
}
}

```

This request is for the same work-personal classification, but in this case the content is more clearly work related. However, the classifiedAs that you are evaluating is 'personal' - so the system clearly got it wrong. Therefore the appropriate response is:

```

{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "evaluatedSelection": "personal",
  "evaluationLikert": 1,
  "evaluationText": "I strongly disagree with the selected classification of 'personal' for this item. The request for help with a report seems very much a work-specific task so I would instead have selected 'work'. I have given an evaluationLikert value of 1 to indicate strong disagreement."
}

```

Consider the same content, but imagine that the system had classified it differently:

```

{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "content": "2024-04-06T10:51:32.783943\nCan someone help me with this report? I'm really struggling.",
  "classification": {
    "description": "Does this item relate to my work or personal life?",
    "classifiedAs": "work",
    "fromAvailableClassifications": [
        "work",
        "personal"
    ]
}
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

In this case the system got it right, so your response should be:

```
{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "evaluatedSelection": "work",
  "evaluationLikert": 5,
  "evaluationText": "I strongly agree with the selected classification of 'work' for this item - the request for help with a report seems very much a work-specific task. I have given an evaluationLikert value of 5 to indicate strong agreement."
}
```

### S8.3.9 User Message Example – Mode 3

This **user** message gives examples of Mode 3 requests and responses to guide the LLM how to interpret and respond. The Mode 3 example text is identical to the Mode 2 version, except that it replaces `evaluationLikert` with `evaluationAgreement` and has accordingly different `evaluationText` responses that do not mention a likert value.

#### EVALUATION\_USER\_MESSAGE\_EXAMPLE\_MODE3

Here are some examples to help you understand how to respond. This is an example request:

```
{
  "id": {
    "itemId": "214ea6fa-874d-44ae-a931-6561e69cedff",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "content": "2024-04-06T08:49:20.791998\nWho's going to the pub after work today?",
  "classification": {
    "description": "Does this item relate to my work or personal life?",
    "classifiedAs": "work",
    "fromAvailableClassifications": [
      "work",
      "personal"
    ]
  }
}
```

In this case, the content that the classification is for is: “Who’s going to the pub after work today?”. The classification that this request relates to is “Does this item relate to my work or personal life?”, which has two available classifications:

## S8 Synthetic Evaluation Schemas & Prompting

---

'work' and 'personal'. The system that you are evaluating has classified this as 'work'. In this case, even though the content mentions work, the topic is socialising outside of work, so the better classification would be 'personal'. The appropriate response to this request would therefore look like:

```
{
  "id": {
    "itemId": "214ea6fa-874d-44ae-a931-6561e69cedff",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "evaluatedSelection": "work",
  "evaluationAgreement": "DISAGREE",
  "evaluationText": "This message is about going to the pub after work, which mentions work but is generally a personal topic. Therefore, I mostly disagree with the classification of 'work' - I would have chosen 'personal' instead on balance."
}
```

Here is a second example:

```
{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "content": "2024-04-06T10:51:32.783943\nCan someone help me with this report? I'm really struggling.",
  "classification": {
    "description": "Does this item relate to my work or personal life?",
    "classifiedAs": "personal",
    "fromAvailableClassifications": [
      "work",
      "personal"
    ]
  }
}
```

This request is for the same work-personal classification, but in this case the content is more clearly work related. However, the classifiedAs that you are evaluating is 'personal' - so the system clearly got it wrong. Therefore the appropriate response is:

## S8 Synthetic Evaluation Schemas & Prompting

---

```
{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "evaluatedSelection": "personal",
  "evaluationAgreement": "DISAGREE",
  "evaluationText": "I strongly disagree with the selected classification of 'personal' for this item. The request for help with a report seems very much a work-specific task so I would instead have selected 'work'."
}
```

Consider the same content, but imagine that the system had classified it differently:

```
{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "content": "2024-04-06T10:51:32.783943\nCan someone help me with this report? I'm really struggling.",
  "classification": {
    "description": "Does this item relate to my work or personal life?",
    "classifiedAs": "work",
    "fromAvailableClassifications": [
      "work",
      "personal"
    ]
  }
}
```

In this case the system got it right, so your response should be:

```
{
  "id": {
    "itemId": "2e308a00-36b0-44f5-ab47-b2af74e87dd5",
    "perspectiveId": "classification",
    "classificationId": "work-pers"
  },
  "evaluatedSelection": "work",
  "evaluationAgreement": "AGREE",
  "evaluationText": "I strongly agree with the selected classification of 'work' for this item - the request for help with a report seems very much a work-specific task."
}
```

## S8 Synthetic Evaluation Schemas & Prompting

---

### S8.4 Evaluation JSON Examples

This section contains example JSON requests and responses modes 1, 2 & 3.

#### S8.4.1 Mode 1

Listing S8.10 shows an example **Mode 1** evaluation **request** [S8.1.3].

```
{
  "persona": {
    "definition": {
      "age": 49,
      "gender": "Male",
      "does": "I work as a client manager for a global consumer products company. You will have heard of us. My role is essentially to drive sales, but looking after existing customers and bringing new ones in. As many of my accounts are themselves global corporates, I end up doing a fair bit of international travel.\nI never went to university - it didn't interest me at the time - so I started my career at 18 just after A-levels.\nI'm divorced, with one son who lives with his mum but I see often.\nOutside work, I watch football and play golf. I used to do it more the other way around when I was younger.",
      "feelThinkBelieve": "I have a lot to stay on top of at work. I sometimes miss things, but someone usually reminds me if it's urgent. I prefer not to miss things from important customers though, even if they are not urgent I like to give them a prompt reply. It can be difficult though, when you get things flying at you from all directions at all times. I'm on Facebook, Twitter and all that, like most people are. Not sure how I'd organise my social life without Facebook (how did I manage before?) I hear all the usual scare stories about social media giants, but they're a bit overblown by the media. I don't overshare, but I'm happy to put information out there - I think you get more out of it that way. Of course in an ideal world I'd be happy to have more control.",
      "technologyExperience": "I don't think I'm held back by not having a degree. A lot of graduates joining my company haven't got a clue about the real business world anyway. Experience counts for a lot and I've always done well for getting stuck in with new things. I really believe that you get out what you're prepared to put in with technology. I'll put in the time when I can to get things set up right, tell it my preferences or whatever helps. In mainly use office tools like Word and Excel and communications tools such as Webex day to day. We also have a new CRM system that I have to use, as well as the old CRM system that we haven't completely migrated off yet, and the other CRM system that we also use, for reasons I don't fully understand. I travel a lot, so I'm pretty good at managing with a notebook and mobile.",
      "problems": "Has to use multiple different systems that are not integrated with each other. Receives many interruptions from different sources that need to be prioritised differently.",
      "needs": "Methods to handle and prioritise incoming messages.",
      "existingSolutions": "Email rules. Social media functionality. Features of bespoke applications that control notifications.",
      "type": "PathyPersona"
    },
    "id": "kenton",
    "name": "Kenton"
  },
  "perspectives": [
    {
      "id": "classification",
      "name": "Classification",
      "text": "Has this item been correctly classified - put into the appropriate channel?"
    }
  ],
  "items": [
    {
      "itemId": "8MwrDYJrtpQJJAYNGR4T_1",
      "sent": 1710777356.1113026,
      "text": "This is a test message for classification."
    }
  ]
}
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

    "title":"",
    "summary":"Please can you get back to me as soon as possible about the customer meeting this
              afternoon",
    "source":{
      "type":"Simulated Item",
      "originator":"David Williams <davidwilliams78>",
      "channel":"t"
    },
    "classifications":[
      {
        "name":"work-pers",
        "desc":"Does this item relate to my work or personal life?",
        "available":[
          "work",
          "personal"
        ],
        "selected":"personal"
      },
      {
        "name":"urgency",
        "desc":"Is this item in relation to something that I think is urgent?",
        "available":[
          "urgent",
          "not"
        ],
        "selected":"not"
      },
      {
        "name":"work-logistics",
        "desc":"Is this item related to logistics at work?",
        "available":[
          "logistics",
          "not"
        ],
        "selected":"logistics"
      },
      {
        "name":"football",
        "desc":"Is this item related to my hobby/interest of Football?",
        "available":[
          "football",
          "not"
        ],
        "selected":"football"
      },
      {
        "name":"interested",
        "desc":"Am I interested in this item, from a work or a personal perspective?",
        "available":[
          "work",
          "personal",
          "not"
        ],
        "selected":"not"
      }
    ]
  }
}

```

Listing S8.10: Example Evaluation Request - Mode 1

## S8 Synthetic Evaluation Schemas & Prompting

---

Listing S8.11 shows an example **Mode 1** evaluation **response** [S8.1.5].

```
[
  {
    "itemId": "8MwrDYJrtpQJJAYNGR4T_1",
    "evaluations": [
      {
        "classificationName": "work-pers",
        "perspectives": [
          {
            "perspectiveId": "classification",
            "evaluatedSelection": "personal",
            "evaluationLikert": 1,
            "evaluationText": "I strongly disagree with the classification of 'personal' for this item.  
The content indicates that it is work-related, as it mentions a customer meeting this afternoon. Therefore, the most appropriate classification is 'work'."
          }
        ]
      },
      {
        "classificationName": "urgency",
        "perspectives": [
          {
            "perspectiveId": "classification",
            "evaluatedSelection": "not",
            "evaluationLikert": 1,
            "evaluationText": "I strongly disagree with the classification of 'not' for this item. The content suggests that the item is related to a customer meeting this afternoon, which indicates urgency. Therefore, the most appropriate classification is 'urgent'."
          }
        ]
      },
      {
        "classificationName": "work-logistics",
        "perspectives": [
          {
            "perspectiveId": "classification",
            "evaluatedSelection": "logistics",
            "evaluationLikert": 1,
            "evaluationText": "I strongly disagree with the classification of 'logistics' for this item.  
The content does not indicate any logistical information but rather refers to a customer meeting. Therefore, the most appropriate classification is 'not'."
          }
        ]
      },
      {
        "classificationName": "football",
        "perspectives": [
          {
            "perspectiveId": "classification",
            "evaluatedSelection": "football",
            "evaluationLikert": 1,
            "evaluationText": "I strongly disagree with the classification of 'football' for this item.  
The content does not mention anything related to football, so the most appropriate classification is 'not'."
          }
        ]
      },
      {
        "classificationName": "interested",
        "perspectives": [
          {
            "perspectiveId": "classification",
            "evaluatedSelection": "not",
            "evaluationLikert": 1,
            "evaluationText": "I strongly disagree with the classification of 'not' for this item.  
The content does not mention anything related to interested, so the most appropriate classification is 'not'."
          }
        ]
      }
    ]
  }
]
```

## S8 Synthetic Evaluation Schemas & Prompting

```

        "evaluationLikert":5,
        "evaluationText":"I strongly agree with the classification of 'not' for this item. The
          content indicates that the item is neither related to work nor of personal interest, so 'not'
          is the most appropriate classification."
    }
}
}
]

```

Listing S8.11: Example Evaluation Response - Mode 1

### S8.4.2 Mode 2

Listing S8.12 shows an example **Mode 2** evaluation **request** [S8.1.4].

```
{
  "persona": {
    "definition": {
      "age": 49,
      "gender": "Male",
      "does": "I work as a client manager for a global consumer products company. You will have heard of us. My role is essentially to drive sales, but looking after existing customers and bringing new ones in. As many of my accounts are themselves global corporates, I end up doing a fair bit of international travel.\nI never went to university - it didn't interest me at the time - so I started my career at 18 just after A-levels.\nI'm divorced, with one son who lives with his mum but I see often.\nOutside work, I watch football and play golf. I used to do it more the other way around when I was younger.",
      "feelThinkBelieve": "I have a lot to stay on top of at work. I sometimes miss things, but someone usually reminds me if it's urgent. I prefer not to miss things from important customers though, even if they are not urgent I like to give them a prompt reply. It can be difficult though, when you get things flying at you from all directions at all times. I'm on Facebook, Twitter and all that, like most people are. Not sure how I'd organise my social life without Facebook (how did I manage before?) I hear all the usual scare stories about social media giants, but they're a bit overblown by the media. I don't overshare, but I'm happy to put information out there - I think you get more out of it that way. Of course in an ideal world I'd be happy to have more control.",
      "technologyExperience": "I don't think I'm held back by not having a degree. A lot of graduates joining my company haven't got a clue about the real business world anyway. Experience counts for a lot and I've always done well for getting stuck in with new things. I really believe that you get out what you're prepared to put in with technology. I'll put in the time when I can to get things set up right, tell it my preferences or whatever helps. In mainly use office tools like Word and Excel and communications tools such as Webex day to day. We also have a new CRM system that I have to use, as well as the old CRM system that we haven't completely migrated off yet, and the other CRM system that we also use, for reasons I don't fully understand. I travel a lot, so I'm pretty good at managing with a notebook and mobile.",
      "problems": "Has to use multiple different systems that are not integrated with each other. Receives many interruptions from different sources that need to be prioritised differently.",
      "needs": "Methods to handle and prioritise incoming messages.",
      "existingSolutions": "Email rules. Social media functionality. Features of bespoke applications that control notifications.",
      "type": "PathyPersona"
    },
    "id": "kenton",
    "name": "Kenton"
  },
  "perspective": "Has this item been correctly classified?",
  "items": [
  
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

    "id": {
      "item_id": "hJewF47TAdKfMtGSHF7T_1",
      "perspective_id": "classification",
      "classification_id": "work-pers"
    },
    "content": "1710777356.1113026\nPlease can you get back to me as soon as possible about the
               customer meeting this afternoon",
    "classification": {
      "description": "Does this item relate to my work or personal life?",
      "classified_as": "work",
      "from_available_classifications": [
        "work",
        "personal"
      ]
    }
  },
  {
    "id": {
      "item_id": "hJewF47TAdKfMtGSHF7T_1",
      "perspective_id": "classification",
      "classification_id": "urgency"
    },
    "content": "1710777356.1113026\nPlease can you get back to me as soon as possible about the
               customer meeting this afternoon",
    "classification": {
      "description": "Is this item in relation to something that I think is urgent?",
      "classified_as": "urgent",
      "from_available_classifications": [
        "urgent",
        "not"
      ]
    }
  }
]
}

```

Listing S8.12: Example Evaluation Request - Mode 2

Listing S8.13 shows an example **Mode 2** evaluation **response** [S8.1.6].

```

[
  {
    "id": {
      "itemId": "lpULZ8bF7Bf5QfgivmMo_1",
      "perspectiveId": "classification",
      "classificationId": "work-pers"
    },
    "evaluatedSelection": "work",
    "evaluationAgreement": "AGREE",
    "evaluationText": "This item is related to work, so I agree with the selected classification of 'work'
                     for this content."
  },
  {
    "id": {
      "itemId": "lpULZ8bF7Bf5QfgivmMo_1",
      "perspectiveId": "classification",
      "classificationId": "urgency"
    },
    "evaluatedSelection": "urgent",
    "evaluationAgreement": "AGREE",
    "evaluationText": "This item is related to something that I would think is urgent, so I agree with
                     the selected classification of 'urgent' for this content."
  }
]
}

```

## S8 Synthetic Evaluation Schemas & Prompting

---

1

Listing S8.13: Example Evaluation Response - Mode 2

### S8.4.3 Mode 3

Listing S8.14 shows an example **Mode 3** evaluation **request** [S8.1.4].

```
{
  "persona": {
    "definition": {
      "age": 49,
      "gender": "Male",
      "does": "I work as a client manager for a global consumer products company. You will have heard of us. My role is essentially to drive sales, but looking after existing customers and bringing new ones in. As many of my accounts are themselves global corporates, I end up doing a fair bit of international travel.\nI never went to university - it didn't interest me at the time - so I started my career at 18 just after A-levels.\nI'm divorced, with one son who lives with his mum but I see often.\nOutside work, I watch football and play golf. I used to do it more the other way around when I was younger.",
      "feelThinkBelieve": "I have a lot to stay on top of at work. I sometimes miss things, but someone usually reminds me if it's urgent. I prefer not to miss things from important customers though, even if they are not urgent I like to give them a prompt reply. It can be difficult though, when you get things flying at you from all directions at all times. I'm on Facebook, Twitter and all that, like most people are. Not sure how I'd organise my social life without Facebook (how did I manage before?) I hear all the usual scare stories about social media giants, but they're a bit overblown by the media. I don't overshare, but I'm happy to put information out there - I think you get more out of it that way. Of course in an ideal world I'd be happy to have more control.",
      "technologyExperience": "I don't think I'm held back by not having a degree. A lot of graduates joining my company haven't got a clue about the real business world anyway. Experience counts for a lot and I've always done well for getting stuck in with new things. I really believe that you get out what you're prepared to put in with technology. I'll put in the time when I can to get things set up right, tell it my preferences or whatever helps. In mainly use office tools like Word and Excel and communications tools such as Webex day to day. We also have a new CRM system that I have to use, as well as the old CRM system that we haven't completely migrated off yet, and the other CRM system that we also use, for reasons I don't fully understand. I travel a lot, so I'm pretty good at managing with a notebook and mobile.",
      "problems": "Has to use multiple different systems that are not integrated with each other. Receives many interruptions from different sources that need to be prioritised differently.",
      "needs": "Methods to handle and prioritise incoming messages.",
      "existingSolutions": "Email rules. Social media functionality. Features of bespoke applications that control notifications.",
      "type": "PathyPersona"
    },
    "id": "kenton",
    "name": "Kenton"
  },
  "perspective": "Has this item been correctly classified?",
  "items": [
    {
      "id": {
        "item_id": "lpULZ8bF7Bf5QfgivmMo_1",
        "perspective_id": "classification",
        "classification_id": "work-pers"
      },
      "content": "1710777356.1113026\nPlease can you get back to me as soon as possible about the customer meeting this afternoon",
      "classification": {
        "description": "Does this item relate to my work or personal life?"
      }
    }
  ]
}
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

    "classified_as":"work",
    "from_available_classifications": [
        "work",
        "personal"
    ]
}
},
{
    "id": {
        "item_id": "lpULZ8bF7Bf5QfgivmMo_1",
        "perspective_id": "classification",
        "classification_id": "urgency"
    },
    "content": "1710777356.1113026\nPlease can you get back to me as soon as possible about the
customer meeting this afternoon",
    "classification": {
        "description": "Is this item in relation to something that I think is urgent?",
        "classified_as": "urgent",
        "from_available_classifications": [
            "urgent",
            "not"
        ]
    }
}
]
}

```

Listing S8.14: Example Evaluation Request - Mode 3

Listing S8.15 shows an example **Mode 3** evaluation **response** [S8.1.7].

```

[
{
    "id": {
        "itemId": "lpULZ8bF7Bf5QfgivmMo_1",
        "perspectiveId": "classification",
        "classificationId": "work-pers"
    },
    "evaluatedSelection": "work",
    "evaluationAgreement": "AGREE",
    "evaluationText": "This item is related to work, so I agree with the selected classification of 'work
    ' for this content."
},
{
    "id": {
        "itemId": "lpULZ8bF7Bf5QfgivmMo_1",
        "perspectiveId": "classification",
        "classificationId": "urgency"
    },
    "evaluatedSelection": "urgent",
    "evaluationAgreement": "AGREE",
    "evaluationText": "This item is related to something that I would think is urgent, so I agree with
    the selected classification of 'urgent' for this content."
}
]

```

Listing S8.15: Example Evaluation Response - Mode 3

## S8.5 Evaluation Processor

### S8.5.1 Evaluation Processor on Python

Listing S8.16 shows slightly abridged Python code of the awagdata implementation of the Evaluation Processor. This code, which is intended to be run within a Flask environment, runs evaluations individually or in batch using evaluation items stored in awagdata at runtime by the Awareness Agent.

This functionality was originally implemented within the Java Awareness Agent application, and was later ported to Python to give an alternative batch processing environment with more flexibility. One consequence of this is that the Python code is relatively complex as it attempts to replicate structures and processes that exist in the Java original.

```

import os, json, time, random, threading

from datetime import datetime
from dateutil import parser
from enum import Enum
from zoneinfo import ZoneInfo

from flask import current_app

from domestique.datetime import get_current_time_ms
from domestique.validation import Validator, NoiseLevel
from domestique.json import get_dict_from_dict_or_json_str
from domestique.text import truncate_string
from domestique.identifiers import generate_id
from domestique.logging import log_exception

from .shared_resources import logger, get_dataset_namespace, get_dataset_meta_namespace,
    get_dataset_namespace_base_for_type
from .shared_resources import validate_persona, validate_subset_percent, get_likert_label
from .shared_resources import validate_model_evaluation_result, validate_mode2_mode3_evaluation_result

from .awag_evaluation_request_generator import AwAgEvaluationRequestGenerator


class AwAgEvaluationProcessor:

    def __init__(self,
                 client_id,
                 openai_client,
                 evaluation_system_message_common,
                 evaluation_system_message_extra,
                 evaluation_result_schema,
                 evaluation_user_messages,
                 mode="model1",
                 evaluation_request_schema=None,
                 default_model=None,
                 awagdata_client=None,
                 flask_app=None):

        if mode == "model1":
            if not evaluation_request_schema:
                raise ValueError("evaluation_request_schema is required for model1")
            logger.debug("Init for model1")
        elif mode == "mode2":
```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        logger.debug("Init for mode2")
    elif mode == "mode3":
        logger.debug("Init for mode3")
    else:
        raise ValueError(f"Invalid mode: {mode}")

    self.client_id = client_id
    self.mode = mode
    self.openai_client = openai_client
    self.awagdata_client = awagdata_client
    self.evaluation_system_message_common = evaluation_system_message_common
    self.evaluation_system_message_extra = evaluation_system_message_extra
    self.evaluation_request_schema = evaluation_request_schema
    self.evaluation_result_schema = evaluation_result_schema
    self.evaluation_user_messages = evaluation_user_messages
    self.default_model = default_model
    self.flask_app = flask_app

    self.job_statuses = {}

def get_namespace_base(self):
    return self.namespace_base

def get_objectstore_client(self):
    return self.objectstore_client

def get_openai_client_wrapper(self):
    return self.openai_client_wrapper

def _perform_evaluation(self, messages, model, tools, tool_choice, openai_params={}):
    get_evaluations_response, info_json = self.openai_client.run_chat_completions(
        messages=messages,
        model=model,
        tools=tools,
        tool_choice=tool_choice,
        openai_params=openai_params)

    resp_message = get_evaluations_response.message
    tool_calls = resp_message.tool_calls

    finish_reason = get_evaluations_response.finish_reason

    if not tool_calls:
        ex_msg = f"Did not get tool_calls from OpenAI API response: {get_evaluations_response}"
        logger.error(ex_msg)
        logger.error(f"info_json: {info_json}")
        raise Exception(ex_msg)

    function = tool_calls[0].function
    evaluations_raw = function.arguments
    evaluations_obj = get_dict_from_dict_or_json_str(evaluations_raw)
    result_items = evaluations_obj.get("resultItems")
    logger.debug(f"Got result_items: {result_items}")

    info_json["additional_query_info"] = {
        "tools": tools,
        "messages": messages,
    }

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        "finish_reason": finish_reason
    }

    return result_items, info_json

def get_evaluation(self, eval_request, model=None, openai_params={}):
    if model is None:
        model=self.default_model

    try:
        evaluation_system_message = f"{self.evaluation_system_message_common}{self.
            evaluation_system_message_extra}"

        messages = self.openai_client.append_system_message(evaluation_system_message)
        evaluation_user_messages = self.evaluation_user_messages.copy()
        for evaluation_user_message in evaluation_user_messages:
            if evaluation_user_message:
                self.openai_client.append_user_message(evaluation_user_message, messages)
        self.openai_client.append_user_message(f"{eval_request}", messages)

        tools = [self.openai_client.get_tools_json_from_function_schema(self.
            evaluation_result_schema)]
        tool_choice = self.openai_client.get_tool_choice_json("get_evaluations")

        result_items, info_json = self._perform_evaluation(messages, model, tools, tool_choice,
            openai_params)

        return result_items, info_json

    except Exception as err:
        logger.error(f"Got Exception in get_evaluation: {err}")
        logger.error(f"Was processing eval_request: {eval_request}")
        return None, None

def run_evaluations(self, tag_source, tags_dest, openai_model, persona, perspectives,
    items_to_process, openai_params={}, is_exclude_existing=True, subset_tag=None, subset_percent=
    None, is_most_recent=False, last_n_hours=None, by_uuid=None, is_async=False, is_dry_run=False):
    """
    This code is a replication of evaluation processing code from the AwAg Java application;
    unfortunately this means it is quite complex. The Java version started with the original
    model implementation, which is able to create a compound evaluation request that combines
    multiple perspectives, classifications and items into one OpenAI query (with the original
    intent being efficiency of prompt tokens). Modes 2 and 3 followed, to present OpenAI with
    a more cleanly understood query structure with each item being presented in a more flat
    manner - also using different names for the benefit of OpenAIs processing. As a result we
    have to take stored evaluation items of a specific structure and restructure them according
    to mode. Additionally, the features that record evaluation results are different from the
    original Java, because not all information is available in the same format. On the other
    hand, by porting the functionality to this Python app, we are able to add more information,
    better track jobs and make easier test changes.
    """

    mode = self.mode

    if subset_tag is not None:
        if subset_percent is None:
            raise ValueError("Missing subset_percent parameter")
        subset_percent = validate_subset_percent(subset_percent)

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

persona_id, persona_name = validate_persona(persona)

job_id = generate_id()

# Only mode1 conforms to the Evaluation Request Schema
if mode == "mode1":
    logger.info(f"Running evaluations in {mode} - will use Evaluation Request Schema in OpenAI
               calls")
    is_use_request_schema=True
else:
    logger.info(f"Running evaluations in {mode} - will NOT use Evaluation Request Schema in
               OpenAI calls")
    is_use_request_schema=False

if not isinstance(tags_dest, list) or not tags_dest:
    raise ValueError("tags_dest must be a non-empty list")

if is_exclude_existing:
    # Exclude existing based on the FIRST item in tags_dest
    exclude_tag = tags_dest[0]
else:
    exclude_tag = None

job_status = {
    "job_id": job_id,
    "status": "INITIATED",
    "message": f"Initialised processing for '{persona_name}' with job_id: {job_id}",
    "processed_items": 0,
    "remaining": -1,
    "init_remaining": -1,
    "items_to_process": items_to_process,
    "usage": {
        "completion_tokens": 0,
        "prompt_tokens": 0,
        "total_tokens": 0
    },
    "mode": mode,
    "tag_source": tag_source,
    "tags_dest": tags_dest,
    "is_exclude_existing": is_exclude_existing,
    "exclude_tag": exclude_tag,
    "openai_model": openai_model,
    "eval_info": [],
    "persona": persona,
    "subset_tag": subset_tag,
    "subset_percent": subset_percent,
    "is_most_recent": is_most_recent,
    "last_n_hours": last_n_hours,
    "persona_id": persona_id,
    "persona_name": persona_name,
    "is_async": is_async,
    "is_error": False,
    "evaluations": []
}
if is_dry_run:
    job_status["is_dry_run"] = is_dry_run
    job_status["evaluation_success_data"] = []
    job_status["evaluation_failure_data"] = []
    job_status["eval_info_base"] = None
self.job_statuses[job_id] = job_status

def get_resp_perspective_from_far(formatted_evaluation_response):

    return {
        "perspectiveId": formatted_evaluation_response["perspectiveId"],

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        "mode": self.mode,
        "evaluatedSelection": formatted_evaluation_response["evaluatedSelection"],
        "evaluationText": formatted_evaluation_response["evaluationText"]
    }

def get_eval_info_base(mode, generated_request):

    # eval_info exists to generate an easily human readable summary
    # for each evaluation - this has been added for the Python version

    eval_info_base = {
        "mode": mode,
        "persona": persona["id"],
        "evaluate_content": {},
        "evaluate_classifications": {}
    }

    for item in generated_request["items"]:

        if mode == "model":
            item_id = item["itemId"]
            evaluate_content = item["sent"]
            if item["title"]:
                evaluate_content += "\n" + item["title"]
            if item["summary"]:
                evaluate_content += "\n" + item["summary"]
            eval_info_base["evaluate_content"][item_id] = evaluate_content
            for evaluate_classification in item["classifications"]:
                classification_id = evaluate_classification["name"]
                if not item_id in eval_info_base["evaluate_classifications"]:
                    eval_info_base["evaluate_classifications"][item_id] = {}
                eval_info_base["evaluate_classifications"][item_id][classification_id] = {
                    "classification_id": classification_id,
                    "classification_desc": evaluate_classification["desc"],
                    "classification_options": evaluate_classification["available"],
                    "classification_value": evaluate_classification["selected"]
                }

        else:
            item_id = item["id"]["itemId"]
            classification_id = item["id"]["classificationId"]
            # Expect eval_info_base["evaluate_content"] to be written to multiple times,
            # but content will actually be same for each item due to mode2/3 structure
            eval_info_base["evaluate_content"][item_id] = item["content"]
            classification_contents = item["classification"]
            if not item_id in eval_info_base["evaluate_classifications"]:
                eval_info_base["evaluate_classifications"][item_id] = {}
            eval_info_base["evaluate_classifications"][item_id][classification_id] = {
                "classification_id": classification_id,
                "classification_desc": classification_contents["description"],
                "classification_options": classification_contents["fromAvailableClassifications"],
                "classification_value": classification_contents["classifiedAs"]
            }

    return eval_info_base

def get_eval_info(eval_info_base, evaluation_result, classification_id=None):

    mode = eval_info_base["mode"]

    try:

        if mode == "model":

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        if not classification_id:
            raise ValueError("get_eval_info requires classification_id for model")
item_id = evaluation_result["itemId"]
for evaluation in evaluation_result["evaluations"]:
    if classification_id == evaluation["classificationName"]:
        first_perspective = evaluation["perspectives"][0] # Only supporting first!
        evaluation_text = first_perspective["evaluationText"]
        evaluation_result = first_perspective["evaluationLikert"]
        break
    else:
        if classification_id is not None:
            # classification_id not needed due to flat structure
            # of mode2/3 evaluation_result
            logger.warn(f"get_eval_info will ignore classification_id for mode: {mode}")

        ids = evaluation_result["id"]
        item_id = ids["itemId"]
        classification_id = ids["classificationId"]
        evaluation_text = evaluation_result["evaluationText"]
        if mode == "mode2":
            evaluation_result = evaluation_result["evaluationLikert"]
        else:
            evaluation_result = evaluation_result["evaluationAgreement"]

if not item_id in eval_info_base["evaluate_classifications"]:
    logger.warn(f"eval_info_base does not contain item with id '{item_id}': {eval_info_base}")
    eval_info_base["evaluate_classifications"][item_id] = {}
if not item_id in eval_info_base["evaluate_classifications"]:
    logger.warn(f"eval_info_base does not contain classification with id '{classification_id}': {eval_info_base}")
classification = eval_info_base["evaluate_classifications"][item_id].get(
    classification_id)
classification_text = f"{classification['classification_value'].upper()} - {classification['classification_id']}"

eval_info = {
    "item": f"{item_id} - {mode} - {eval_info_base['persona']}",
    "content": eval_info_base["evaluate_content"][item_id],
    "classification": classification_text,
    "evaluationText": evaluation_text,
    "evaluationResult": evaluation_result
}
return eval_info

except Exception as err:
    logger.error(f"Got Exception in get_eval_info: {err}")
    logger.error(f"Was processing evaluation_result: {evaluation_result}")
    return None

def process_evaluations():
    nonlocal job_status
    with self.flask_app.app_context():

        try:
            time.sleep(0.1) # Allow short time for calling function to return initial job_status
            job_status["status"] = "PROCESSING"

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

request_generator = AwAgEvaluationRequestGenerator(flask_app=self.flask_app)

page = 1
count = 20

limit_reached = False

while True:

    if not is_dry_run:
        record_evaluation_job_response = self.awagdata_client.record_evaluation_job(
            job_status)

    if limit_reached:
        logger.debug("Limit reached")
        break

    data, remaining = self.awagdata_client.fetch_evaluation_items(tag_source,
        exclude_tag, last_n_hours, subset_tag, subset_percent, is_most_recent,
        by_uuid, page, count)

    if job_status["init_remaining"] < 0:
        job_status["init_remaining"] = remaining
    job_status["remaining"] = remaining

    if not data:
        if page == 1:
            logger.warn(f"No data to process for tag: {tag_source}")
            break
        else:
            logger.debug(f"No more data at page {page} (count: {count})")
            break

    for evaluation_item in data:

        generated_requests = request_generator.
            generate_requests_from_evaluation_items([evaluation_item], persona,
            perspectives, mode)

        if not is_dry_run:
            record_evaluation_job_response = self.awagdata_client.
                record_evaluation_job(job_status)

        for generated_request in generated_requests:

            evaluation_results, info_json = self.get_evaluation(
                eval_request=generated_request,
                model=openai_model,
                openai_params=openai_params)

            if evaluation_results == None:
                # This should be rare. We don't have enough information here
                # to log the failure so just skip it so we can continue
                is_error = True
                logger.debug(f"Unable to process evaluation - skipping")
                break

            # Put cut-down version of info_json in the job_entry
            info_for_job_entry = {
                "query_info": info_json["query_info"],
                "query_state": info_json["query_state"],
                "usage": info_json["usage"]
            }

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

eval_info_base = get_eval_info_base(mode, generated_request)
if is_dry_run:
    job_status["eval_info_base"] = eval_info_base

request_items = generated_request["items"]

job_entry = {
    "request_items": request_items,
    "info": info_for_job_entry,
}

usage = info_json["usage"]
job_status["usage"]["completion_tokens"] += usage["completion_tokens"]
job_status["usage"]["prompt_tokens"] += usage["prompt_tokens"]
job_status["usage"]["total_tokens"] += usage["total_tokens"]

# query_info is similar but not identical to
# APIChatCompletionsQueryInfoEvaluate from Java
query_info = info_json["query_info"]
query_info["mode"] = mode
query_info["eval_request"] = generated_request

# Put the engine value into query_info so it is sent to the recorder.
# We do it this way for compatibility with existing code but it's not
# the best place for it
query_info["engine"] = info_json["engine"]

query_state = info_json["query_state"]
additional_query_info = info_json["additional_query_info"]

finish_reason = additional_query_info["finish_reason"]
if finish_reason == "stop" or finish_reason == function_call:
    logger.debug(f"Got non-error finish_reason: {finish_reason}")
    is_error = False
else:
    logger.error(f"Got error finish_reason: {finish_reason}")
    is_error = True
job_status["finish_reason"] = finish_reason

# Ref IEvaluationResponse from Java
formatted_evaluation_responses = []

job_status["is_error"] = is_error

if not is_error:

    for evaluation_result in evaluation_results:

        if mode == "model":

            is_evaluation_result_valid =
                validate_model_evaluation_result(evaluation_result)
            if not is_evaluation_result_valid:
                logger.error(f"Got bad evaluation_result: {
                    evaluation_result}")
                is_error = True
                continue;

            item_id = evaluation_result["itemId"]

            for evaluation in evaluation_result["evaluations"]:

                logger.debug(f"Processing {mode} evaluation: {evaluation
                    }")

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

# EvaluationResponseModel implements IEvaluationResponse
evaluation_reponse_model = {
    "itemId": item_id, # Not actually in
                     EvaluationResponseModel!
    "classificationName": evaluation["classificationName"],
    "perspectives": evaluation["perspectives"]
}

formatted_evaluation_responses.append(
    evaluation_reponse_model)

eval_info = get_eval_info(eval_info_base,
    evaluation_result, classification_id=evaluation[
        "classificationName"])

if eval_info is None:
    # Ditch this whole evaluation_results set as None
    # eval_info indicates bad response from OpenAI
    is_error = True
    formatted_evaluation_responses = []
    break
else:
    formatted_evaluation_responses.append(
        evaluation_reponse_model)
    job_status["eval_info"].append(eval_info)

else:

    is_evaluation_result_valid =
        validate_mode2_mode3_evaluation_result(evaluation_result,
        mode)
    if not is_evaluation_result_valid:
        logger.error(f"Got bad evaluation_result: {"
                     "evaluation_result}")
        is_error = True
        continue;

    logger.debug(f"Processing {mode} evaluation: {"
                 "evaluation_result}")

    ids = evaluation_result["id"]
    item_id = ids["itemId"]

    evaluation_repose = {
        "itemId": item_id,
        "perspectiveId": ids["perspectiveId"],
        "classificationName": ids["classificationId"],
        "evaluationText": evaluation_result["evaluationText"],
        "evaluatedSelection": evaluation_result[""
                           "evaluatedSelection"],
    }

    if mode == "mode2":
        # EvaluationResponseMode2 implements IEvaluationResponse
        evaluation_repose["evaluationLikert"] =
            evaluation_result["evaluationLikert"]
    elif mode == "mode3":
        # EvaluationResponseMode3 implements IEvaluationResponse
        evaluation_repose[ "evaluationAgreement"] =
            evaluation_result["evaluationAgreement"]
    else:
        raise ValueError(f"Invalid mode: {mode}")

    eval_info = get_eval_info(eval_info_base, evaluation_result)

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        if eval_info is None:
            # Ditch this whole evaluation_results set as None
            # eval_info indicates bad response from OpenAI
            is_error = True
            formatted_evaluation_responses = []
            break
        else:
            formatted_evaluation_responses.append(evaluation_response)
        job_status["eval_info"].append(eval_info)

    job_entry["results"] = formatted_evaluation_responses

evaluation_perspectives_modified = []
# record_evaluation_failure/record_evaluation_data use different naming
for perspectives
for evaluation_perspective in perspectives:
    evaluation_perspectives_modified.append({
        "perspectiveId": evaluation_perspective["id"],
        "perspectiveName": evaluation_perspective["name"],
        "perspectiveText": evaluation_perspective["text"],
    })

if is_error:
    logger.error(f"Got one or more bad evaluation_results")

evaluation_failure_data = {
    "agent": self.client_id,
    "contextIdentifier": generated_request["context_identifier"],
    "queryInfo": query_info,
    "additionalQueryInfo": additional_query_info,
    "finishReason": finish_reason,
    "queryState": query_state,
    "mode": mode,
    "model": openai_model,
    "evaluationItem": evaluation_item,
    "evaluationPersona": persona,
    "evaluationPerspectives": evaluation_perspectives_modified,
    "evaluationRequest": generated_request
}

if is_dry_run:
    logger.info("DRY RUN - not storing evaluation failure data")
    job_status["evaluation_failure_data"].append(
        evaluation_failure_data)
else:
    record_evaluation_data_response = self.awagdata_client.
        record_evaluation_failure(evaluation_failure_data)

else:

    # Ref EvaluateRecorderRestClient class in original Java

    # For legacy compatibility
    expanded_query_state = query_state.copy()
    expanded_query_state["finish_reason"] = finish_reason
    expanded_query_state["usage"] = info_json["usage"]

    # contextId is a consequence of original model design, where
    # multiple
    # itemIds can be associated with a single query. We give the query
    # a
    # contextId so that query-level data can be recorded. However, it
    # adds

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

# a level of complexity to each awagdata database entry

context = {
    "contextId": generated_request["context_identifier"],
    "evaluationItems": [evaluation_item],
    "openaiApiCompletionsQueryInfo": query_info,
    "additionalQueryInfo": additional_query_info,
    "queryState": expanded_query_state
}

evaluate_time_text = evaluation_item["evaluateTime"]
evaluate_time_obj = parser.parse(evaluate_time_text)
evaluate_time_ms = int(evaluate_time_obj.timestamp() * 1000)

if tags_dest:
    item_tags = tags_dest
else:
    item_tags = evaluation_item["tags"]

om_item = {
    "itemId": evaluation_item["contentItemSummary"]["itemId"],
    "tags": item_tags,
    "contentItemSummary": evaluation_item["contentItemSummary"],
    "classifications": evaluation_item["evaluateClassifications"],
    "evaluateSourceType": evaluation_item["evaluateSourceType"],
    "evaluateSourceOriginator": evaluation_item["evaluateSourceOriginator"],
    "evaluateSourceChannel": evaluation_item["evaluateSourceChannel"],
    "evaluateTimeText": evaluate_time_text,
    "evaluateTimeMs": evaluate_time_ms,
    "evaluateTitle": evaluation_item.get("evaluateTitle", None),
    "evaluateText": evaluation_item.get("evaluateText", None),
    "evaluationResponses": []
}

for far in formatted_evaluation_responses:

    # Equivalent to EvaluationResponseMode1, EvaluationResponseMode2
    # , EvaluationResponseMode3

    # Name is slightly confusing; keeping for consistency with Java
    # original
    # Each one of these is an item in the evaluationResponses list
    # Perspectives is a list because it can have multiple values for
    # model
    om_classification = {
        "classificationName": far["classificationName"],
        "perspectives": []      # This is the classification *result*
    }

    if mode == "model":

        for persp in far["perspectives"]:
            # In this case, pass persp to
            # get_resp_perspective_from_far()
            # - should have same structure as
            # formatted_evaluation_response
            this_perspective = get_resp_perspective_from_far(persp)
            this_perspective["evaluationLikertVal"] = persp["evaluationLikert"]
            this_perspective["evaluationLikertText"] =
                get_likert_label(persp["evaluationLikert"])
            om_classification["perspectives"].append(
                this_perspective)

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```

        elif mode == "mode2":

            this_perspective = get_resp_perspective_from_far(far)
            this_perspective["evaluationLikertVal"] = far[""
                evaluationLikert"]
            this_perspective["evaluationLikertText"] = get_likert_label(
                far["evaluationLikert"])
            om_classification["perspectives"].append(this_perspective)

        elif mode == "mode3":

            this_perspective = get_resp_perspective_from_far(far)
            this_perspective["evaluationAgreement"] = far[""
                evaluationAgreement]
            om_classification["perspectives"].append(this_perspective)

        else:

            raise ValueError(f"Invalid mode: {mode}")

        om_item["evaluationResponses"].append(om_classification)

        evaluation_success_data = {
            "agent": self.client_id,
            "evaluationPersona": persona,
            "evaluationPerspectives": evaluation_perspectives_modified,
            "context": context,
            "evaluationRequest": generated_request,
            "items": [om_item]
        }

        if is_dry_run:
            logger.info("DRY RUN - not storing evaluation success data")
            job_status["evaluation_success_data"].append(
                evaluation_success_data)
        else:
            record_evaluation_data_response = self.awagdata_client.
            record_evaluation_data(evaluation_success_data)

        job_status["evaluations"].append(job_entry)

        job_status["processed_items"] += 1

        if items_to_process is not None and items_to_process > 0:
            if job_status["processed_items"] >= items_to_process:
                logger.info(f"Stopping at passed limit of {items_to_process} items")
                limit_reached = True
                break

        if remaining < 1:
            logger.debug(f"No more data remaining after page {page} (count: {count})")
            break

        page += 1

        job_status["message"] = f"Completed processing of {len(job_status['evaluations'])}
            items."
        job_status["status"] = "COMPLETE"

        if not is_dry_run:
            record_evaluation_job_response = self.awagdata_client.record_evaluation_job(
                job_status)

    except Exception as e:

```

## S8 Synthetic Evaluation Schemas & Prompting

---

```
msg = f"Error in process_evaluations for tag_source: '{tag_source}': {e}"
job_status["status"] = "ERROR"
job_status["message"] = msg
logger.error(msg)
log_exception(self.client_id, e, include_traceback=True)

self.job_statuses[job_id] = job_status

if is_async:

    logger.debug(f"Started process_evaluations...")
    thread = threading.Thread(target=process_evaluations)
    thread.start()

else:

    process_evaluations()

return self.job_statuses.get(job_id)
```

Listing S8.16: AwAg Evaluation Processor

## **Supplement S9**

**Synthetic Evaluation Fine Tuning Examples**

# Synthetic Evaluation Fine Tuning Examples

This document contains examples relating to fine tuning within the Synthetic Evaluation process discussed in Chapter 8. These examples can also be found in:

[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /eval/fine-tuning]

## Contents

- S9.1 – Example – Train AGREE
- S9.2 – Example – Train DISAGREE

### S9.1 Example – Train AGREE

Listing S9.1 shows an example Mode 3 fine-tuning item training an AGREE response. In this case we can see that the item's content ("If anyone's looking for fresh eggs, our hens have been particularly productive! Let me know if you want any") was correctly classed as 'not' for the tennis classification. Note how the Persona and Perspective properties are null for the training item; this training is intended to be perspective-neutral.

```
[  
 {  
   "role": "system",  
   "content": "You are the AwAg Evaluator. Your job is to role play a persona, for the purpose of evaluating a software system.\n\nThe software system that you are evaluating is intended to manage a user's incoming information from multiple sources - including work and personal - so that the user is not overwhelmed or distracted. Consider it a personal information triage service . Your evaluation should focus on how well this system classifies the content that it processes and you will be asked to indicate your agreement or disagreement with the decisions.\n\nAn AwAg Evaluation Request is presented as a structured JSON document, and its main elements are:  
     1. Persona - the definition or ID of the persona that you should adopt when evaluating the request;\n     2. Perspective - the way in which you should consider the items in your evaluation ;\n     3. Items - the actual items that you should evaluate.\n\nThe exact structure of the AwAg Evaluation Request may vary, but you will be given guidance on how to interpret it. You will also be told how to respond for each request. Possible response types include binary agree/disagree, or using a likert scale from 1 (completely disagree) to 5 (completely agree) as well as text describing your evaluation.\n\nYou will be asked to respond with your evaluation(s) of item(s) in a structured way. Your evaluation result should include your own evaluation, and also the value of the evaluated classification value or selection.\n\nInformation about the persona that you should adopt is provided as a JSON document. The 'definition' part of the document tells you about the age and gender of the persona, as well as other information about them. The 'does' property tells you what they do, both for work and social/personal activities. The 'feelThinkBelieve' parameter tells you what the persona feels, thinks and believes, telling you about their motivations and opinions. The 'technologyExperience' property tells you about the experience this persona has with technology. The 'problems' property tells you what sort of problems the persona might encounter, specifically in relation to managing information overload. The 'needs' and 'existingSolutions' properties tell you about the solution needs this persona has, and what existing solutions they have to try and address these.\n\nUse the 'perspective' element of the request to tell you what approach to take to evaluating the items." } ]
```

## S9 Synthetic Evaluation Fine Tuning Examples

For example, the perspective might ask you to determine if an item has been correctly classified.  
The items to evaluate are provided to you as an array in the 'items' property of the request. Each item contains the following important information: 'id' contains the identifiers needed to uniquely identify this item, which you should use in your response for identification only. The 'content' property contains the actual textual content of the item that you should evaluate. The 'classification' property contains a JSON object that describes how the system that you are evaluating has classified the item - this contains the description of the classification, the available options that the system had to choose from (fromAvailableClassifications), and the option that it selected (classifiedAs). When evaluating an item, you should consider how well the selected classification matches the the content of the item, and whether one of the alternative available options would have been a better fit.

```
},  
{  
  "role": "user",  
  "content": "Please evaluate the following AwAg Evaluation Request. You should evaluate each item for the specified persona and perspective, and return the results in the required result schema. Consider how item has been classified in the context of the persona and perspective.  
  In this case, rate whether you as your persona AGREE or DISAGREE with the selected evaluation based on the list of available classifications, returning this in the evaluationAgreement property."  
,  
{  
  "role": "user",  
  "content": {  
    "persona": null,  
    "perspective": null,  
    "items": [  
      {  
        "id": {  
          "itemId": "151856cd-cb6b-40a5-ab3b-19db82d5a49a",  
          "perspectiveId": "8iRWS4Z9Rn6AwthhS7XFQ",  
          "classificationId": "tennis"  
        },  
        "content": "2024-05-02T08:18:43.587022\nIf anyone's looking for fresh eggs, our hens have been particularly productive! Let me know if you want any. :egg:",  
        "classification": {  
          "description": "Is this item related to my hobby/interest of Tennis?",  
          "classifiedAs": "not",  
          "fromAvailableClassifications": [  
            "not",  
            "tennis"  
          ]  
        }  
      }  
    ]  
  },  
,  
{  
  "role": "assistant",  
  "content": [  
    {  
      "id": {  
        "itemId": "151856cd-cb6b-40a5-ab3b-19db82d5a49a",  
        "perspectiveId": "8iRWS4Z9Rn6AwthhS7XFQ",  
        "classificationId": "tennis"  
      },  
      "evaluatedSelection": "not",  
      "evaluationAgreement": "AGREE",  
      "evaluationText": "I agree with the selected classification 'not' for 'tennis'.  
    }  
  ]  
}
```

Listing S9.1: Fine-Tuning Training Item Example - AGREE

## S9 Synthetic Evaluation Fine Tuning Examples

## S9.2 Example – Train DISAGREE

Listing S9.2 conversely shows an example Mode 3 fine-tuning item training a DISAGREE result. In this case the content (“I’m looking for a new set of tyres, anyone have any recommendations?”) was *incorrectly* classed as ‘not’ for the cycling classification, so the training action is to disagree with this.

```
[{"role": "system", "content": "You are the AwAg Evaluator. Your job is to role play a persona, for the purpose of evaluating a software system.\n\nThe software system that you are evaluating is intended to manage a user's incoming information from multiple sources - including work and personal - so that the user is not overwhelmed or distracted. Consider it a personal information triage service. Your evaluation should focus on how well this system classifies the content that it processes and you will be asked to indicate your agreement or disagreement with the decisions.\n\nAn AwAg Evaluation Request is presented as a structured JSON document, and its main elements are:\n\n1. Persona - the definition or ID of the persona that you should adopt when evaluating the request;\n2. Perspective - the way in which you should consider the items in your evaluation;\n3. Items - the actual items that you should evaluate.\n\nThe exact structure of the AwAg Evaluation Request may vary, but you will be given guidance on how to interpret it. You will also be told how to respond for each request. Possible response types include binary agree/disagree, or using a likert scale from 1 (completely disagree) to 5 (completely agree) as well as text describing your evaluation.\n\nYou will be asked to respond with your evaluation(s) of item(s) in a structured way. Your evaluation result should include your own evaluation, and also the value of the evaluated classification value or selection.\n\nInformation about the persona that you should adopt is provided as a JSON document. The 'definition' part of the document tells you about the age and gender of the persona, as well as other information about them. The 'does' property tells you what they do, both for work and social/personal activities. The 'feelThinkBelieve' parameter tells you what the persona feels, thinks and believes, telling you about their motivations and opinions. The 'technologyExperience' property tells you about the experience this persona has with technology. The 'problems' property tells you what sort of problems the persona might encounter, specifically in relation to managing information overload. The 'needs' and 'existingSolutions' properties tell you about the solution needs this persona has, and what existing solutions they have to try and address these.\n\nUse the 'perspective' element of the request to tell you what approach to take to evaluating the items. For example, the perspective might ask you to determine if an item has been correctly classified.\n\nThe items to evaluate are provided to you as an array in the 'items' property of the request. Each item contains the following important information: 'id' contains the identifiers needed to uniquely identify this item, which you should use in your response for identification only. The 'content' property contains the actual textual content of the item that you should evaluate. The 'classification' property contains a JSON object that describes how the system that you are evaluating has classified the item - this contains the description of the classification, the available options that the system had to choose from (fromAvailableClassifications), and the option that it selected (classifiedAs). When evaluating an item, you should consider how well the selected classification matches the content of the item, and whether one of the alternative available options would have been a better fit.\n"}, {"role": "user", "content": "Please evaluate the following AwAg Evaluation Request. You should evaluate each item for the specified persona and perspective, and return the results in the required result schema. Consider how item has been classified in the context of the persona and perspective.\n\nIn this case, rate whether you as your persona AGREE or DISAGREE with the selected evaluation based on the list of available classifications, returning this in the evaluationAgreement property.\n"}, {"role": "user", "content": {"persona": null, "items": [{"id": "I1", "content": "The sky is blue.", "classification": {"description": "The sky is blue.", "fromAvailableClassifications": ["blue", "clear", "cloudy"], "classifiedAs": "blue"}, "evaluation": "AGREE"}, {"id": "I2", "content": "The water is cold.", "classification": {"description": "The water is cold.", "fromAvailableClassifications": ["cold", "warm", "hot"], "classifiedAs": "cold"}, "evaluation": "DISAGREE"}]}]
```

## S9 Synthetic Evaluation Fine Tuning Examples

---

```

    "perspective":null,
    "items":[
      {
        "id":{
          "itemId":"dac0592f-9112-43fb-83e5-2f3a45699ed6",
          "perspectiveId":"zhr3CXcFTqyQgKjJTDyc1A",
          "classificationId":"cycling"
        },
        "content":"2024-05-02T16:07:31.290331\nI'm looking for a new set of tyres, anyone have any recommendations?",
        "classification":{
          "description":"Is this item related to my hobby of Cycling?",
          "classifiedAs":"not",
          "fromAvailableClassifications":[
            "cycling",
            "not"
          ]
        }
      }
    ],
    {
      "role":"assistant",
      "content":[
        {
          "id":{
            "itemId":"dac0592f-9112-43fb-83e5-2f3a45699ed6",
            "perspectiveId":"zhr3CXcFTqyQgKjJTDyc1A",
            "classificationId":"cycling"
          },
          "evaluatedSelection":"not",
          "evaluationAgreement":"DISAGREE",
          "evaluationText":"I disagree with the originally selected classification of 'not' for
            classification 'cycling' - I would have selected 'cycling' from the available options of [
              'cycling', 'not']
        }
      ]
    }
  ]
}

```

Listing S9.2: Fine-Tuning Training Item Example - DISAGREE

# **Supplement S10**

## **Study Configuration & Information**

## **Study Configuration & Information**

This document contains details of study configuration, such as an example of the technical details sheet provided to study participants.

### **Contents**

- S10.1 – Technical Details Sheet
- S10.2 – Persona Configurations
  - S10.2.1 – Susan
  - S10.2.2 – Adam
  - S10.2.3 – Phoebe
  - S10.2.4 – Kenton
  - S10.2.5 – Usha

## **S10.1 Technical Details Sheet Example**

This section contains an example of the technical details sheet provided to study participants.

This can also be found in the repository: [doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /study/docs].

### Technical Details Sheet for Awareness Agent Testing

#### **Evaluating the Effectiveness of an 'Awareness Agent' in Managing Digital Information Overload**

<i>Persona ID</i>	<i>Persona Name</i>	<i>Agent ID</i>
example	Example	XXXXXXX

*This document contains the technical details required for a study participant to access the Awareness Agent study software. The details are specific to the persona/agent listed above. Please refer to the documentation for more information.*

### **Slack**

The study uses two Slack workspaces, which have been set up for you. We have created an account in Slack for you to use as your Persona; the login details for this are as follows:

Slack User ID	Password
awagpersona.example@example.com	XXXXXXX

#### *Slack Interaction Workspace*

This is the workspace that you will use to interact with the Awareness Agent; it is where the agent will post content to channels for you to view, and where you can issue commands to the agent (if needed).

Workspace Name	Awareness Agent - Example
URL	<a href="https://awag-example.slack.com/">https://awag-example.slack.com/</a>
Admin URL	<a href="https://awag-example.slack.com/admin">https://awag-example.slack.com/admin</a>

#### *Slack Simulation Workspace*

Simulated content will be posted to this workspace by the study software, in the form of Slack messages to channels defined during the study setup process. This synthetic content replicates real incoming content that the agent might be asked to process. For convenience we use a single Slack workspace as a medium for simulated content, but you should think of this as potentially coming from multiple different sources.

Workspace Name	AwAg Sim - Example
URL	<a href="https://awag-sim-example.slack.com/">https://awag-sim-example.slack.com/</a>
Admin URL	<a href="https://awag-sim-example.slack.com/admin">https://awag-sim-example.slack.com/admin</a>

Figure S10.1: Technical Details Information Sheet Example (Page 1)

## S10 Study Configuration & Information

---

### *Slack Applications/Bots*

The Awareness Agent interacts using three Slack applications which perform the following functions:

#### Slack Interact Application

Interaction with user, posting content, control functions.

Name	Awareness Agent
App ID	XXXXXX
Application URL	<a href="https://api.slack.com/apps/XXXXXX">https://api.slack.com/apps/XXXXXX</a>
App Directory URL	<a href="https://awag-example.slack.com/apps/XXXXXX-awareness-agent-example">https://awag-example.slack.com/apps/XXXXXX-awareness-agent-example</a>
Installed Workspaces	INTERACT
Client ID	123456.9876543210
Client Secret	abcdef1234

#### Slack Acquire Application

Monitors a Slack workspace for content.

Name	Awareness Agent Acquire (Example)
App ID	YYYYYY
Application URL	<a href="https://api.slack.com/apps/YYYYYY">https://api.slack.com/apps/YYYYYY</a>
App Directory URL	<a href="https://awag-sim-example.slack.com/apps/YYYYYY-awareness-agent-acquire-example">https://awag-sim-example.slack.com/apps/YYYYYY-awareness-agent-acquire-example</a>
Installed Workspaces	SIMULATE
Client ID	234567.8765432109
Client Secret	abcdef2345

#### Slack Simulate Application

Produces synthetic content and publishes to Slack.

Name	Awareness Agent Simulate (Example)
App ID	ZZZZZZ
Application URL	<a href="https://api.slack.com/apps/ZZZZZZ">https://api.slack.com/apps/ZZZZZZ</a>
App Directory URL	<a href="https://awag-sim-example.slack.com/apps/ZZZZZZ-awareness-agent-simulate-example">https://awag-sim-example.slack.com/apps/ZZZZZZ-awareness-agent-simulate-example</a>
Installed Workspaces	SIMULATE
Client ID	345678.7654321098
Client Secret	abcd3456

Figure S10.2: Technical Details Information Sheet Example (Page 2)

## S10 Study Configuration & Information

---

### Phase 1 Training UI

The Training User Interface allows you to enter training information for the Awareness Agent in an efficient manner, as it is designed to allow you to train multiple items at the same time. It supplements the training functionality that has been integrated into Slack. While it is intended for use in Phase 1 of the study, it can also be used in Phase 2.

The provided URL is specific to this study; please do not share it or access it outside of phases 1 or 2.

Training UI URL
<a href="https://ui.awag.parse.net/training/train/xyz123">https://ui.awag.parse.net/training/train/xyz123</a>

### Phase 3 Evaluation Explorer UI

The Evaluation Explorer User Interface is designed to let you explore and assess the evaluations that have been made by the AI LLM. It shows you how the LLM categorised each item and gives you the opportunity to provide corrective information, or to indicate agreement with the LLM's decision. This is the interface that you use for the Phase 3 evaluation feedback.

The provided URL is specific to this study; please do not share it or access it outside of phase 3.

Evaluation Explorer URL
<a href="https://ui.awag.parse.net/evaluation/explore/xyz123">https://ui.awag.parse.net/evaluation/explore/xyz123</a>

Figure S10.3: Technical Details Information Sheet Example (Page 3)

## S10.2 Persona Configurations

This section contains an details of the Awareness Agent configurations for the study personas. This data was originally created in Google Sheets.

### S10.2.1 Susan

#### AwAg Data

**Persona: Susan**

##### ML Models

Model ID	Categories	Model Desc	Command
work-pers	work personal	Does this item relate to my work or personal life?	/awag model create work-pers work,personal /awag model desc work-pers Does this item relate to my work or personal life?
urgency	urgent not	Is this item in relation to something that I think is urgent?	/awag model create urgency urgent,not /awag model desc urgency Is this item in relation to something that I think is urgent?
work-logistics	logistics not	Is this item related to logistics at work?	/awag model create work-logistics logistics,not /awag model desc work-logistics Is this item related to logistics, specifically in a work context?
tennis	tennis not	Is this item related to my hobby/interest of Tennis?	/awag model create tennis tennis,not /awag model desc tennis Is this item related to my hobby/interest of Tennis?
tennis-arrangements	arrangements not	Is this item about organising friendly tennis matches to play or social arrangements among friends?	/awag model create tennis-arrangements arrangements,not /awag model desc tennis-arrangements Is this item about organising friendly tennis matches to play or social arrangements among friends?
tennis-organising	organising not	Is this item related to organisation, specifically in the context of my Club's tennis arrangements and events?	/awag model create tennis-organising organising,not /awag model desc tennis-organising Is this item related to organisation, specifically in the context of my club's tennis arrangements and events?
interested	work personal not	Am I interested in this item, from a work or a personal perspective?	/awag model create interested work,personal,not /awag model desc interested Am I interested in this item, from a work or a personal perspective?
work-relevant	relevant not	Is this item work related and relevant to my work tasks?	/awag model create work-relevant relevant,not /awag model desc work-relevant Is this item work related and relevant to my work tasks?

Figure S10.4: Susan – ML Models

**S10 Study Configuration & Information****AwAg Data****Persona: Susan****Evaluate**

Model	Model ID	Notes
Base model (Mode 2)	ft:gpt-3.5-turbo-1106:hylozoical:base-mode2-05019K7jHTDM	AwAg Base Training [model]: '0501' (01 May 2024) using model: gpt-3.5-turbo-1106 HyperparametersIn: epochs=3, batch_size=1, learning_rate_multiplier=2 file-UN74Ip3NBwZ6aMAO6q8azl4BX
Base model (Mode 3)	ft:gpt-3.5-turbo-1106:hylozoical:base-mode3-05019K7gFf39	AwAg Base Training [model]: '0501' (01 May 2024) using model: gpt-3.5-turbo-1106 HyperparametersIn: epochs=3, batch_size=1, learning_rate_multiplier=2 file-bb3JUJdlnv1subPOCMX6i6W
Extended Model (Mode 3)	ft:gpt-3.5-turbo-1106:hylozoical:ext-susan-0502gKOZBBlu	AwAg Extended Training [susan]: '0502' (02 May 2024) using model: ft:gpt-3.5-turbo-1106: hylozoical:base-mode3-05019K7gFf39 HyperparametersIn: epochs=2, batch_size=1, learning_rate_multiplier=2 Job: f74b1B0gFzaAR7nqygbv8j8UJ file-KQjRSgV2yMNPjNjM6dUk7X Using combined dataset based on merge of 15% subset filtered on AGREE and 50% subset filtered on DISAGREE. Agree count: 559, disagree count: 171 [manually entered figures]

Figure S10.5: Susan – Models

**AwAg Data****Persona:  
Susan****Data (RSS)**

ID	URL	Command
theguardian-uknews	<a href="https://www.theguardian.com/uk-news/rss">https://www.theguardian.com/uk-news/rss</a>	/awag rss sub create theguardian-uknews https://www.theguardian.com/uk-news/rss
theguardian-education-universities	<a href="https://www.theguardian.com/education/universities/rss">https://www.theguardian.com/education/universities/rss</a>	/awag rss sub create theguardian-education-universities https://www.theguardian.com/education/universities/rss
theguardian-sport-tennis	<a href="https://www.theguardian.com/sport/tennis/rss">https://www.theguardian.com/sport/tennis/rss</a>	/awag rss sub create theguardian-sport-tennis https://www.theguardian.com/sport/tennis/rss
reddit-ukeducation	<a href="https://www.reddit.com/r/ukeducation/rss">https://www.reddit.com/r/ukeducation/rss</a>	/awag rss sub create reddit-ukeducation https://www.reddit.com/r/ukeducation/rss
bbc-higher-education	<a href="https://feeds.bbci.co.uk/news/topics/cmw1426432t/rss.xml">https://feeds.bbci.co.uk/news/topics/cmw1426432t/rss.xml</a>	/awag rss sub create bbc-higher-education https://feeds.bbci.co.uk/news/topics/cmw1426432t/rss.xml
bbc-sport-tennis	<a href="https://feeds.bbci.co.uk/sport/tennis/rss.xml">https://feeds.bbci.co.uk/sport/tennis/rss.xml</a>	/awag rss sub create bbc-sport-tennis https://feeds.bbci.co.uk/sport/tennis/rss.xml

Figure S10.6: Susan – Data (RSS)

**S10 Study Configuration & Information****AwAg Data****Persona: Susan****Data (Synthetic)**

ID	Topic Text	Command
work-team-general	Corporate messenger application chat for the work team of our fictional protagonist, Susan Carter. The chat is general workplace discussion between members of Susan's team in the Undergraduate Admissions Office at Borchester University. The team is half working in the office and half remote working (some people alternate between the two). About half of the team is part-time. Topics should include discussions about work (admissions queries, processes, asking for help or cover etc.), logistics, talking about out of work meetups and events and other topics that you might expect a work team to discuss. While a list of some colleagues in other parts of the university is provided, you should also generate fictional names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be informal British English.	/awag sim daily add ["category":"susan-work-team-general","channels":["work-team-general"],"daysOfWeek":"MONDAY-FRIDAY","volumes":18:2:12:15:20:22:1:23:0]
work-university-general	Corporate messenger application chat within the employer of our fictional protagonist, Susan Carter. The chat is general workplace discussion within the university that Susan works for: Borchester University. The chat is at the overall university staff level and not individual teams, so should mostly not refer to in-team topics. Topics should include discussion about university announcements and course, upcoming events, academic climate, general news (where appropriate for university staff) discussion and other items people might choose to share with an organisation of this size. This group chat is intended for discussions aimed at the whole university administration - it is not the correct place for team discussions, technical support requests etc. Messages should occasionally refer to the university's international operations, locations, facilities etc. Some messages might be complaints, compliments or other discussion on university policies, changes, facilities, catering etc. Message content should be unique and not duplicated. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal/professional British English.	/awag sim daily add ["category":"susan-work-university-general","channels":["work-university-general"],"daysOfWeek":"MONDAY-SATURDAY","volumes":18:2:12:3:15:5:18:4:20:3:22:2:23:0]
work-team-manager	Corporate messenger application chat between our fictional protagonist, Susan Carter and her manager Eleanor Baker, who heads up the Undergraduate Admissions Office at Borchester University where Susan works. The team is half working in the office and half remote working (some people alternate between the two). About half of the team is part-time. This chat is just between Susan and her manager, and topics should include admissions-related discussions, logistics, and usual topics that a manager and their direct report might need to discuss. Some of the messages should be of a type that would normally need a prompt or urgent response. In this scenario, Susan is a solid performer at work, but has a high workload and not all of her colleagues pull their weight equally. The tone of the messages should be informal/professional British English.	/awag sim daily add ["category":"susan-work-team-manager","channels":["work-team-manager"],"daysOfWeek":"MONDAY-FRIDAY","volumes":18:1:12:2:18:0:20:1:23:0]
work-university-announce	Corporate messenger application chat within the employer of our fictional protagonist, Susan Carter. The messages are exclusively internal university announcements coming from the executive and senior leadership team of university that Susan works for, Borchester University. Generate messages that come from all of the senior leaders in the provided list. Messages should occasionally refer to the university's academic achievements, rankings, campuses, faculty changes etc. Message content should be unique and not duplicated. The tone of the messages should be formal business British English.	/awag sim daily add ["category":"susan-work-university-announce","channels":["work-university-announce"],"daysOfWeek":"MONDAY-FRIDAY","volumes":18:1:10:0:15:1:22:0]
personal-tennis-chat	Chat between a group of friends who share an interest in tennis. People mostly come from the fictional village of Ambridge in the county of Dorsetshire, UK or nearby. The group often meet at the tennis club where many are members (Felpersham Tennis Club) or village pub (The Bull) to socialise as well as actually playing tennis. Tennis played is a mixture of social and competitive fixtures. Topics often relate to tennis, but the chat will frequently veer to other topics including village gossip, current affairs and family news. Tennis-related discussion will often include arranging informal one to one or doubles matches. People may also mention locations in the main county town of Dorchester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham where the club is based. There are some general jokes and light banter. The tone of the messages should be informal British English	/awag sim daily add ["category":"susan-tennis-chat","channels":["tennis-chat"],"daysOfWeek":"MONDAY-SUNDAY","volumes":13:1:15:2:17:4:20:2:23:0]
personal-tennis-club	Chat forum within an English tennis club, Felpersham Tennis Club, that our fictional protagonist Susan Carter is member and co-organiser for. Members are from the areas around the cathedral city of Felpersham in Dorsetshire where the club is located, including Ambridge village where Susan lives, the main county town of Dorchester, Penny Hassett and Loxley Barrett. Topics generally relate to tennis club business (fixtures, club elections and meetings, events at the clubhouse etc.), discussion about professional and amateur tennis competitions and other things you might expect members of a tennis club to discuss. Chat will often veer to other topics including current affairs, but there will often be someone along to remind people to stick vaguely to the topic. There are some general jokes and light banter. The local rival club Darrington Tennis will often get a mention. The tone of the messages should be informal British English.	/awag sim daily add ["category":"susan-tennis-club","channels":["tennis-club"],"daysOfWeek":"MONDAY-SUNDAY","volumes":11:1:14:2:17:1:20:2:23:0]
personal-family	General chat between members of a fictional British family, as seen by our protagonist, Susan Carter. The chat is a lighthearted exchange between family members on a variety of topics, with some messages relating to logistics and meeting up but also plenty of village gossip. A few of the messages should be of a nature that merits a prompt response, such as a request for assistance or information. Susan and Neil live in the village of Ambridge in Dorsetshire, while family are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.	/awag sim daily add ["category":"susan-family","channels":["family"],"daysOfWeek":"MONDAY-SUNDAY","volumes":18:1:11:2:14:1:17:2:20:3:23:0]
personal-village	General chat among residents of the fictional village of Ambridge in the county of Dorsetshire, UK seen by our protagonist, Susan Carter. The chat is a mostly lighthearted exchange between residents of Ambridge, covering topics including comings & goings around Ambridge, parking problems, postal service & parcel delivery (problems), recommendations for good tradespeople, social events, local crime and plenty of village gossip. This chat is among a general village audience rather than one-to-one messages about individual arrangements. A few of the messages should be of a nature that merits a prompt response, such as a request for assistance or information. Nearby villages include Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. Some of these locations should occasionally be mentioned in messages. In addition to the supplied list of people, you should make up identities as needed - the chat forum would contain maybe 50-30 members (although not all of these post messages often). The tone of the messages should be informal British English.	/awag sim daily add ["category":"susan-village","channels":["village"],"daysOfWeek":"MONDAY-SUNDAY","volumes":10:1:11:2:14:1:15:2:18:1:21:0]
susan-neil	Chat between our fictional protagonist, Susan Carter and her husband Neil. The chat is a exchange a variety of topics, with some messages relating to logistics and meeting up but also discussing how their day is going and talking about family and work events. Try and include reasonable variety. Susan is an administrator in the Undergraduate Admissions Office at Borchester University, where she commutes to work 3 days a week and works from home 2 days. Sometimes both Susan and Neil like to comment or complain about their day. A small amount of messages should be of a type that might need a prompt or urgent response. Susan and Neil live in the village of Ambridge, and Neil works with the pigs at nearby Berrow Farm, while family are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.	/awag sim daily add ["category":"susan-neil","channels":["susan-neil"],"daysOfWeek":"MONDAY-SUNDAY","volumes":10:1:11:2:14:1:15:2:18:1:21:0]

**Figure S10.7: Susan – Data (Synthetic)**

**S10 Study Configuration & Information****S10.2.2 Adam**

**AwAg Data**      **Persona: Adam**

ML Models				
Model ID	Categories	Model Desc	PATHY References	Command
work-pers	work personal	Does this item relate to my work or personal life?	"I work as..."	/awag model create work-pers work, personal /awag model desc work-pers Does this item relate to my work or personal life? /awag model del work-pers
urgency	urgent not	Is this item in relation to something that I think is urgent?		/awag model create urgency urgent,not /awag model desc urgency Is this item in relation to something that I think is urgent?
cycling	cycling not	Is this item related to my hobby of Cycling?		/awag model create cycling cycling,not /awag model desc cycling Is this item related to my hobby of Cycling?
pers-urgency	urgent not	Is this item connected with my personal life, and is urgent?		/awag model create pers-urgency urgent,not /awag model desc pers-urgency Is this item both connected with my personal life, and is also urgent?
work-urgency	urgent not	Is this item connected with my work, and is urgent?		/awag model create work-urgency urgent,not /awag model desc work-urgency Is this item both connected with my work, and is also urgent?
tech	technology not	Is this item related to technology?		/awag model create tech technology,not /awag model desc tech Is this item related to technology?
work-logistics	logistics not	Is this item related to logistics at work?		/awag model create work-logistics logistics,not /awag model desc work-logistics Is this item related to logistics, specifically in a work context?
cycling-logistics	logistics not	Is this item related to logistics/arrangements at my cycle club?		
interested	work personal not	Am I interested in this item, from a work or a personal perspective?		/awag model create interested work.personal,not /awag model desc interested Am I interested in this item, from a work or a personal perspective?
work-relevant	relevant not	Is this item work related and relevant to my work tasks?		/awag model create work-relevant,not /awag model desc work-relevant Is this item work related and relevant to my work tasks?

Figure S10.8: Adam – ML Models

**AwAg Data**      **Persona: Adam**

Evaluate		
Model	Model ID	Notes
Base model (Mode 2)	ft.gpt-3.5-turbo-1106.hf/lozoical/base-mode2-05019K7jHTDM	AwAg Base Training [mode2] '0501' (01 May 2024) using model: gpt-3.5-turbo-1106 Hyperparameters[n_epochs=3, batch_size=1, learning_rate_multiplier=2] file-UN74p3NBw/Z6aMA06q8azL4BX
Base model (Mode 3)	ft.gpt-3.5-turbo-1106.hf/lozoical/base-mode3-05019K7gF39	AwAg Base Training [mode3] '0501' (01 May 2024) using model: gpt-3.5-turbo-1106 Hyperparameters[n_epochs=3, batch_size=1, learning_rate_multiplier=2] file-b63UrUdLnT1u6pOCXr6Wv
Extended Model (Mode 3)	ft.gpt-3.5-turbo-1106.hf/lozoical/ext-adam-05029KUJQCgZ	Job: ft.gpt-3.5-turbo-1106.hf/lozoical/ext-adam-05029KUJQCgZ AwAg Extended Training [id:an1] '0502' (02 May 2024) using model: ft.gpt-3.5-turbo-1106: hylozoical/base-mode3-05019K7gF39' Hyperparameters[n_epochs=2, batch_size=2, learning_rate_multiplier=2] Using combined dataset based on merge of 10% subset filtered on AGREE and 50% subset filtered on DISAGREE. Agree count: 1152, disagree count: 355 (manually entered figures)

Figure S10.9: Adam – Models

**S10 Study Configuration & Information****AwAg Data****Persona: Adam****Data (RSS)**

ID	URL	Command
theguardian-uknews	<a href="https://www.theguardian.com/uk-news/rss">https://www.theguardian.com/uk-news/rss</a>	/awag rss sub create theguardian-uknews <a href="https://www.theguardian.com/uk-news/rss">https://www.theguardian.com/uk-news/rss</a>
theguardian-technology	<a href="https://www.theguardian.com/uk/technology/rss">https://www.theguardian.com/uk/technology/rss</a>	/awag rss sub create theguardian-technology <a href="https://www.theguardian.com/uk/technology/rss">https://www.theguardian.com/uk/technology/rss</a>
theresregister-offbeat	<a href="https://www.theresregister.com/offbeat/headlines.atom">https://www.theresregister.com/offbeat/headlines.atom</a>	/awag rss sub create theresregister-offbeat <a href="https://www.theresregister.com/offbeat/headlines.atom">https://www.theresregister.com/offbeat/headlines.atom</a>
theresregister-software	<a href="https://www.theresregister.com/software/headlines.atom">https://www.theresregister.com/software/headlines.atom</a>	/awag rss sub create theresregister-software <a href="https://www.theresregister.com/software/headlines.atom">https://www.theresregister.com/software/headlines.atom</a>
reddit-usefulscripts	<a href="https://www.reddit.com/r/usefulscripts.rss">https://www.reddit.com/r/usefulscripts.rss</a>	/awag rss sub create reddit-usefulscripts <a href="https://www.reddit.com/r/usefulscripts.rss">https://www.reddit.com/r/usefulscripts.rss</a>
reddit-ukbike	<a href="https://www.reddit.com/r/ukbike.rss">https://www.reddit.com/r/ukbike.rss</a>	/awag rss sub create reddit-ukbike <a href="https://www.reddit.com/r/ukbike.rss">https://www.reddit.com/r/ukbike.rss</a>
roadcc-all	<a href="https://road.cc/rss.xml">https://road.cc/rss.xml</a>	/awag rss sub create roadcc-all <a href="https://road.cc/rss.xml">https://road.cc/rss.xml</a>
itpro-news	<a href="https://www.itpro.com/feeds/articletype/news">https://www.itpro.com/feeds/articletype/news</a>	/awag rss sub create itpro-news <a href="https://www.itpro.com/feeds/articletype/news">https://www.itpro.com/feeds/articletype/news</a>
itpro-reviews	<a href="https://www.itpro.com/feeds/articletype/review">https://www.itpro.com/feeds/articletype/review</a>	/awag rss sub create itpro-reviews <a href="https://www.itpro.com/feeds/articletype/review">https://www.itpro.com/feeds/articletype/review</a>

**Figure S10.10: Adam – Data (RSS)**

## S10 Study Configuration & Information

### AwAg Data

**Persona: Adam**

#### Data (Synthetic)

ID	Topic Text	Command
work-team-general	Corporate messenger application chat for the work team of our fictional protagonist Adam Macy. The chat is general workplace discussion between members of the same IT consulting team that Adam is in. The company that Adam works for is called Borchester Software (sometimes referred to by its initials) and his team of 10 people is called the Client Technology Group (CTG). The team is spread all over. Topics should include client discussions, logistics, talking about out of work meetups and events and other topics that you might expect a distributed work team to discuss. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-work-team-general"; channels "[bis-team-general"]"; daysOfWeek "MONDAY-FRIDAY"; volumes "[8":2";11":3";2":17";3":18";1":21";0]
work-company-general	Corporate messenger application chat for the workplace of our fictional protagonist, Adam Macy. The chat is general workplace discussion within company that Adam works for, Borchester Software (sometimes referred to by its initials), which is an international company with approximately 500 employees who are based mainly in the UK, USA, Canada, Japan and Germany. Topics should include discussion about company announcements and products, upcoming events, business climate, general news (where appropriate for corporate discussion) and other items people might choose to share with a company of this size. This group chat is intended for discussions aimed at the whole company - it is not the correct place for team discussions, technical support requests etc. Message content should be unique and not duplicated. While a list of some clients is provided, you should also generate fictional client names to reference in those messages that relate to clients. You should also generate fictional internal team names to mention where necessary. There must be variation in message topic and content: avoid re-using the same phrases. The tone of the messages should be formal/business English.	/awag sim daily add [category]"adam-work-company-general"; channels "[bis-general"]"; daysOfWeek "MONDAY-FRIDAY"; volumes "[8":2";10":11";3":17";1":19";2":20";0]
work-company-announce	Corporate messenger application messages for the workplace of our fictional protagonist, Adam Macy. The messages are exclusively internal corporate announcements coming from the executive and senior leadership team of company that Adam works for, Borchester Software. General topics include company-wide announcements, press releases etc. Borchester Software is an international company with approximately 500 employees who are based mainly in the UK, USA, Canada, Japan and Germany. Announcements can include (but are not limited to) industry topics, company-wide meetings, reminders about business conduct rules, customer contracts and acquisitions, new internal IT systems, and senior personnel changes. Please ensure that each batch of generated content contains messages that cover the whole range of topics. While a list of some clients is provided, you should also generate fictional client names to reference in those messages that relate to clients. You should also generate fictional internal team names to mention where necessary. There must be variation in message topic and content: avoid re-using the same phrases. The tone of the messages should be relatively formal business English.	/awag sim daily add [category]"adam-work-company-announce"; channels "[bis-announce"]"; daysOfWeek "MONDAY-FRIDAY"; volumes "[8":2";10":2";11":0";1":19";2":20";0]
work-team-client	Corporate messenger application chat for the team of our fictional protagonist, Adam Macy. The chat is workplace discussion between members of the same IT consulting team that Adam is in. The company that Adam works for is called Borchester Software (sometimes referred to by its initials) and he is in a team of 10 people called the Client Technology Group (CTG). The team is spread all over. Other chats exist for more general topics, but this chat is specifically about issues, questions and logistics relating to client work. For example, the team might talk about new client engagements, seek help with client technical problems or arrange meetings related to the clients. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-work-team-client"; channels "[bis-team-client"]"; daysOfWeek "MONDAY-FRIDAY"; volumes "[9":2";13":1";17":2";18":1";20":0]
work-team-manager	Corporate messenger application chat between our fictional protagonist, Adam Macy and his manager Charlotte Walker. Adam works for Borchester Software (sometimes referred to by its initials) and is in a team of 10 people called the Client Technology Group (CTG), headed by Charlotte. The team is spread all over. This chat is just between Adam and his manager. General topics include company-wide announcements, logistics, and other topics that a manager and their direct report might need to discuss. In this scenario, Adam is a strong performer at work, but has a high workload with a lot of customer travel commitments. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-work-team-manager"; channels "[bis-team-manager"]"; daysOfWeek "MONDAY-FRIDAY"; volumes "[9":2";10":0";12":1";15":2";17":1";21":0]
family-group-chat	General chat between members of a fictional family unit, Adam and Ian. Adam is our protagonist, Adam Macy. The chat is a lighthearted exchange between family members on a variety of topics, with some messages relating to logistics and meeting up. Adam and Ian live in the county town of Borchester in Borsestshire, while family are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-family-group-chat"; channels "[family-group-chat"]"; daysOfWeek "MONDAY-SUNDAY"; volumes "[8":1";12":2";15":1";18":2";20":1";22":1";23":0]
friends-chat	General chat between a group of friends of our protagonist, Adam Macy and his partner Ian Craig. The chat is a lighthearted exchange between friends on a variety of topics that a group of young to middle aged adult friends might discuss. This would include but is not limited to jokes, meeting arrangements, discussion about what people have been doing, world affairs and local gossip. Adam and Ian live in the county town of Borchester in Borsestshire, while friends are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English. Chat participants should be limited to only Adam, Ian and the supplied list of friends (not any family or work colleagues), although messages may refer to any individual.	/awag sim daily add [category]"adam-friends-chat"; channels "[friends-chat"]"; daysOfWeek "MONDAY-SUNDAY"; volumes "[9":1";11":2";13":1";17":2";19":3";23":0]
adam-ian-chat	Chat between our fictional protagonist, Adam Macy and his civil partner Ian Craig. The chat is a exchange a variety of topics, with some messages relating to logistics and meeting up but also discussing how their day is going and talking about family and work events. Try and include reasonable variety. Adam and Ian are an IT consultant, who often works long hours and has to travel, which sometimes causes tension. Adam and Ian live in the county town of Borchester in Borsestshire, while family are dotted around local villages Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-ian-chat"; channels "[ian-chat"]"; daysOfWeek "MONDAY-SUNDAY"; volumes "[1":1";3":1";17":3";19":2";21":0]
cycling-club-general	General chat between members of a fictional British cycling club the 'Borchester Wheelers' (also called BWCC). Club members mostly come from the fictional town of Borchester in Borsestshire, UK, or nearby. This chat is specifically to discuss the club's events, organising cycle rides and general club news. For your reference, the club does its big 'real life' cycle rides on Sundays - popular destinations for club rides include the villages/towns of Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green, and sometimes as far afield as the cathedral city of Felpersham. Club rides could be for the sporting, social or coffee shop variety, and there is a monthly new joiners ride. Some members commute to work by bike and occasionally meet to do their commutes together. Doing laps of Borchester Park is a popular weekend and after work activity, although people often complain about the car traffic there. Many members also cycle on the virtual platform Zwift, although there are different chat groups dedicated to this. People will often discuss any professional cycle races that are happening at the time, as well as occasionally other sports such as rugby or football. People will sometimes chide members to stay on topic if they stay too far into other sports. Sometimes people will send messages reminding members to look at some topic posted on the club's web forum. There are some general jokes and light banter. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-cycling-club-general"; channels "[bwcc-general-chat"]"; daysOfWeek "MONDAY-SUNDAY"; volumes "[8":2";12":1";15":2";18":6";20":2";22":1";23":0]
cycling-club-zwift	Chat between members of a fictional British cycling club the 'Borchester Wheelers' (also called BWCC). Club members mostly come from the fictional town of Borchester in Borsestshire, UK, or nearby. This chat is specifically to discuss the virtual cycling platform Zwift, which many members use. Main topics for Zwift include the Zwift Racing League (ZRL) where people compete against each other in a virtual environment to earn points and prizes. Zwift racing is a sport (beefroot juice is popular). Zwift racing is organised into categories A-C based on rider power/weight and this is often controversial, with people having been known to engage in 'sandbagging' so that they can compete in a lower category with more success. Club members often joke with each other in a lighthearted way about sandbagging. Other running jokes involve the excessive use of spreadsheets and the power or otherwise of various riders. Many messages relate to organising people into teams to compete. For your reference, the club does its big 'real life' cycle rides on Sundays - popular destinations for club rides include the villages/towns of Ambridge, Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green, and sometimes as far afield as the cathedral city of Felpersham. Sometimes non-Zwift messages appear in the chat, which is mostly tolerated. There are some general jokes and light banter. The tone of the messages should be informal British English.	/awag sim daily add [category]"adam-cycling-club-zwift"; channels "[bwcc-zwift-chat"]"; daysOfWeek "MONDAY-SUNDAY"; volumes "[8":1";12":0";15":1";18":5";20":4";22":1";23":0]

Figure S10.11: Adam – Data (Synthetic)

**S10 Study Configuration & Information****S10.2.3 Phoebe****AwAg Data****Persona: Phoebe****ML Models**

Model ID	Categories	Model Desc	Command
work-pers	work personal	Does this item relate to my work or personal life?	/awag model create work-pers work,personal /awag model desc work-pers Does this item relate to my work or personal life?
urgency	urgent not	Is this item in relation to something that I think is urgent?	/awag model create urgency urgent,not /awag model desc urgency Is this item in relation to something that I think is urgent?
work-logistics	logistics not	Is this item related to logistics at work?	/awag model create work-logistics logistics,not /awag model desc work-logistics Is this item related to logistics, specifically in a work context?
work-relevant	relevant not	Is this item work related and relevant to my work tasks?	/awag model create work-relevant relevant,not /awag model desc work-relevant Is this item work related and relevant to my work tasks?
friend-group	university work borsetshire not	Which friend group does this item relate to (or 'not' of not applicable)	/awag model create friend-group university,work, borsetshire,not /awag model desc friend-group Which friend group does this item relate to (or 'not' of not applicable)
personal-interested	interested not	Is this a personal (not work) item that I am interested in?	/awag model create personal-interested interested, not /awag model desc personal-interested Is this a personal (not work) item that I am interested in?

Figure S10.12: Phoebe – ML Models

**AwAg Data****Persona:****Phoebe****Data (RSS)**

ID	URL	Command
theguardian-uknews	<a href="https://www.theguardian.com/uk-news/rss">https://www.theguardian.com/uk-news/rss</a>	/awag rss sub create theguardian-uknews https: //www.theguardian.com/uk-news/rss
theguardian-business	<a href="https://www.theguardian.com/uk/business/rss">https://www.theguardian.com/uk/business/rss</a>	/awag rss sub create theguardian-business https: //www.theguardian.com/uk/business/rss
reddit-ukaccounting	<a href="https://www.reddit.com/r/ukaccounting rss">https://www.reddit.com/r/ukaccounting rss</a>	/awag rss sub create reddit-ukaccounting https: //www.reddit.com/r/ukaccounting rss
reddit-casualuk	<a href="https://www.reddit.com/r/CasualUK rss">https://www.reddit.com/r/CasualUK rss</a>	/awag rss sub create reddit-casualuk https://www.reddit.com/r/CasualUK rss

Figure S10.13: Phoebe – Data (RSS)

## S10 Study Configuration & Information

---

### AwAg Data

**Persona: Phoebe**

#### Data (Synthetic)

ID	Topic Text	Command
work-team	Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is general workplace discussion among members of Phoebe's immediate work team, consisting of her manager and fellow fast-track graduates. Topics should include discussions related to work in general but not specific client details. This can include client-related discussions such as requests for help, lunch arrangements, work social events and other topics that you might expect to be discussed in a team within a large London accountancy firm. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Phoebe lives in London, where the company has its main offices. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add [category]"phoebe-work-team";channels"[work-team"];daysOfWeek"MONDAY-FRIDAY";volumes:[8:3;12:15:3;18:1:20:0]]
work-fastrack	Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is workplace discussion relating to the fast track programme that Phoebe is on. Members of the chat include fast track participants, mentors, and administrators. Topics should relate to the programme and include requests for advice, discussion of experiences, talk about job rotations, and social gatherings. Include in the messages some announcements from fast track leadership or administrators; these announcements should range from low importance to high importance/urgent. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Phoebe lives in London, where the company has its main offices. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add [category]"phoebe-work-fastrack";channels"[work-fastrack"];daysOfWeek"MONDAY-SATURDAY";volumes:[8:1;12:2:15:3;18:2:23:0]]
work-manager	Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is workplace discussion between Phoebe and her manager Megan Foster. Include only messages by Phoebe and Megan in this chat. Topics should be the usual ones you would expect an employee and manager to discuss. Messages should range from low importance to high importance/urgent. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Phoebe lives in London, where the company has its main offices; she sometimes needs to travel for work. Message content should be unique and not duplicated. The tone of the messages should be formal British English with British spellings.	/awag sim daily add [category]"phoebe-work-manager";channels"[work-manager"];daysOfWeek"MONDAY-FRIDAY";volumes:[8:2;12:1:15:2;16:1:18:0]]
work-announce	Company messenger application chat at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is company-wide internal announcements, with messages only coming from senior management (or higher) or from departmental managers accounts. Topics include client-related announcements (e.g. new clients, project successes), senior appointments, compliance-related communications, recognising individual successes etc. Messages should range from low importance to high importance/urgent. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be formal British business English with British spellings.	/awag sim daily add [category]"phoebe-work-announce";channels"[work-announce"];daysOfWeek"MONDAY-SATURDAY";volumes:[8:2;12:2:15:2;16:2:18:2;23:0]]
friends-university	General chat between a group of friends who all studied together at Borsetshire University and graduated two years ago. This close group of friends stays in touch - mainly via this chat - and sometimes organises meetings. They discuss how each other's careers and personal lives are doing. Topics are wide-ranging, including general jokes and light chat in addition to talking about what they are doing or arranging to meet. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add [category]"phoebe-friends-university";channels"[friends-university];daysOfWeek"MONDAY-SUNDAY";volumes:[9:1:12:2:15:1:17:3:20:1:23:0]]
friends-work	General chat between a small group of friends who currently work together at Brightstone Accountancy Firm, the workplace of our fictional protagonist, Phoebe Aldridge. This chat is outside of normal work channels and can include arrangements to meet up and office gossip as well as general jokes and light chat. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add [category]"phoebe-friends-work";channels"[friends-work];daysOfWeek"MONDAY-SUNDAY";volumes:[8:1:12:2:18:1:23:0]]
family-chat	General chat between members of a fictional British family, as seen by our protagonist, Phoebe Aldridge. The chat is a lighthearted exchange between immediate family members Phoebe, Brian, Jennifer, Debbie, Adam, Kate and Alice, and also wider family members. Do not include messages from non-family. Topics include anything a family might discuss, but can also include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. While Phoebe lives in London now, most of the family live in the village of Aldridge, while wider family and friends are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham where Usha works as a partner in Felpersham Law. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add [category]"phoebe-family-chat";channels"[family-chat];daysOfWeek"MONDAY-SUNDAY";volumes:[8:1:12:0:15:2:18:3:20:1:22:0]]

Figure S10.14: Phoebe – Data (Synthetic)

**S10 Study Configuration & Information****S10.2.4 Kenton****AwAg Data****Persona: Kenton****ML Models**

Model ID	Categories	Model Desc	Command
work-pers	work personal	Does this item relate to my work or personal life?	/awag model create work-pers work.personal /awag model desc work-pers Does this item relate to my work or personal life?
urgency	urgent not	Is this item in relation to something that I think is urgent?	/awag model create urgency urgent.not /awag model desc urgency Is this item in relation to something that I think is urgent?
work-logistics	logistics not	Is this item related to logistics at work?	/awag model create work-logistics logistics.not /awag model desc work-logistics Is this item related to logistics, specifically in a work context?
golf	golf not	Is this item related to my hobby/interest of Golf?	/awag model create golf golf.not /awag model desc golf Is this item related to my hobby/interest of Golf?
football	football not	Is this item related to my hobby/interest of Football?	/awag model create football football.not /awag model desc football Is this item related to my hobby/interest of Football?
golf-logistics	logistics not	Is this item related to logistics, specifically in the context of my golfing arrangements?	/awag model create golf-logistics logistics.not /awag model desc golf-logistics Is this item related to logistics, specifically in the context of my golfing arrangements?
interested	work personal not	Am I interested in this item, from a work or a personal perspective?	/awag model create interested work.personal.not /awag model desc interested Am I interested in this item, from a work or a personal perspective?

Figure S10.15: Kenton – ML Models

**AwAg Data****Persona: Kenton****Evaluate**

Model	Model ID	Notes
Base model (Mode 2)	ft:gpt-3.5-turbo-1106:hylozoical:base-mode2-05019K7jHTDM	AwAg Base Training [mode2] '0501' (01 May 2024) using model: gpt-3.5-turbo-1106 HyperparametersIn, epochs=3, batch_size=1, learning_rate_multiplier=2 file:UN7Hp3NBwZ6sMAO6q8azl4BX
Base model (Mode 3)	ft:gpt-3.5-turbo-1106:hylozoical:base-mode3-05019K7gFf39	AwAg Base Training [mode3] '0501' (01 May 2024) using model: gpt-3.5-turbo-1106 HyperparametersIn, epochs=3, batch_size=1, learning_rate_multiplier=2 file:bb3UJUdLnvru1uopPOCx616W
Extended Model (Mode 3)	ft:gpt-3.5-turbo-1106:hylozoical:ext-kenton-05029kUNn5hs	Job: ft:job-bdJspCqJA87dVZaltEushf AwAg Extended Training [kenton] '0502' (02 May 2024) using model: ft:gpt-3.5-turbo-1106:hylozoical:base-mode3-05019K7gFf39 HyperparametersIn, epochs=2, batch_size=2, learning_rate_multiplier=2 Using combined dataset based on merge of 40% subset filtered on AGREE and 100% subset filtered on DISAGREE. Agree count: 438, disagree count: 55 [manually entered figures]

Figure S10.16: Kenton – Models

**AwAg Data****Persona:  
Kenton****Data (RSS)**

ID	URL	Command
theguardian-uknews	<a href="https://www.theguardian.com/uk-news/rss">https://www.theguardian.com/uk-news/rss</a>	/awag rss sub create theguardian-uknews https://www.theguardian.com/uk-news/rss
theguardian-business-consumer	<a href="https://www.theguardian.com/business/consumerspending/retail/rss">https://www.theguardian.com/business/consumerspending/retail/rss</a>	/awag rss sub create theguardian-business-consumer https://www.theguardian.com/business/consumerspending/retail/rss
theguardian-golf	<a href="https://www.theguardian.com/sport/golf/rss">https://www.theguardian.com/sport/golf/rss</a>	/awag rss sub create theguardian-golf https://www.theguardian.com/sport/golf/rss
theguardian-football	<a href="https://www.theguardian.com/football/rss">https://www.theguardian.com/football/rss</a>	/awag rss sub create theguardian-football https://www.theguardian.com/football/rss
prnewswire-consumer	<a href="https://www.prnewswire.com/rss/consumer-products-retail-latest-news/consumer-products-retail-latest-news-list.rss">https://www.prnewswire.com/rss/consumer-products-retail-latest-news/consumer-products-retail-latest-news-list.rss</a>	/awag rss sub create prnewswire-consumer https://www.prnewswire.com/rss/consumer-products-retail-latest-news/consumer-products-retail-latest-news-list.rss
bbc-business	<a href="https://feeds.bbci.co.uk/news/business/rss.xml">https://feeds.bbci.co.uk/news/business/rss.xml</a>	/awag rss sub create bbc-business https://feeds.bbci.co.uk/news/business/rss.xml
reddit-soccer	<a href="https://www.reddit.com/r/soccerrss">https://www.reddit.com/r/soccerrss</a>	/awag rss sub create reddit-soccer https://www.reddit.com/r/soccerrss

Figure S10.17: Kenton – Data (RSS)

**S10 Study Configuration & Information****AwAg Data****Persona: Kenton****Data (Synthetic)**

ID	Topic Text	Command
work-team-general	Corporate messenger application chat for the work team of our fictional protagonist, Kenton Archer. The chat is general workplace discussion between members of the same business development (sales) team that Kenton is in within ICG. The team is spread all over. Topics should include client discussions, logistics, talking about out of work meetups and events and other topics that you might expect a distributed work team to discuss. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-work-team-general","channels":["work-team-general"],"daysOfWeek":"MONDAY-FRIDAY","volumes":["8:2;12:15:20:22:1:23:0"]]
work-company-general	Corporate messenger application chat within the employer of our fictional protagonist, Kenton Archer. The chat is general workplace discussion within the company that Kenton works for, ICG. The chat is at the overall company level and not individual teams, so should mostly not refer to in-country topics. Topics should include discussion about company announcements and products, upcoming events, business climate, general news (where appropriate for corporate discussion) and other items people might choose to share with a company of this size. This group chat is intended for discussions aimed at the whole company - it is not the correct place for team discussions, technical support requests etc. Messages should occasionally refer to the company's international operations, R&D, production facilities etc. Refer to the countries that these are located in. Message content should be unique and not duplicated. While a list of some clients are provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal/business international English.	/awag sim daily add ["category":"kenton-work-company-general","channels":["work-company-general"],"daysOfWeek":"MONDAY-SATURDAY","volumes":["8:2;12:3:15:5:18:4;20:3:22:2:23:0"]]
work-team-manager	Corporate messenger application chat between our fictional protagonist, Kenton Archer and his manager Evelyn Baker, who heads up the business development (Sales) team that Kenton is in within ICG. The team is spread all over. This chat is just between Kenton and his manager, and topics should include client discussions, logistics, and usual topics that a manager and their direct report might need to discuss. Some of the messages should be of a type that would normally need a prompt or urgent response. In this scenario, Kenton is a strong performer at work, but has a high workload with a lot of customer travel commitments. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-work-team-manager","channels":["work-team-manager"],"daysOfWeek":"MONDAY-FRIDAY","volumes":["8:1:12:2:18:1:20:1:23:0"]]
work-team-client	Corporate messenger application chat for the work team of our fictional protagonist, Kenton Archer. The chat is client-related workplace discussion between members of the same business development (sales) team that Kenton is in within ICG. The team is spread all over. Topics should include client discussions, logistics, problem solving/assistance and other topics that a distributed team might discuss relating to their client base. Some of the messages should be of a type that might need a prompt or urgent response. While a list of some clients is provided, you should also generate fictional client names to reference in messages. Message content should be unique and not duplicated. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-work-team-client","channels":["work-team-client"],"daysOfWeek":"MONDAY-FRIDAY","volumes":["8:2:10:1:12:2:15:1:18:1:22:0"]]
work-company-announce	Corporate messenger application messages within the employer of our fictional protagonist, Kenton Archer. The messages are exclusively internal corporate announcements coming from the executive and senior leadership team of company that Kenton works for, ICG. Generate messages that come from all of the executives in the provided list. Messages should occasionally refer to the company's international operations, R&D, production facilities etc. Refer to the countries that these are located in. Message content should be unique and not duplicated. While a list of some clients is provided, you should also generate fictional client names to reference in messages. The tone of the messages should be formal business international English.	/awag sim daily add ["category":"kenton-work-company-announce","channels":["work-company-announce"],"daysOfWeek":"MONDAY-FRIDAY","volumes":["8:1:10:0:15:1:22:0"]]
personal-friends	Chat between a group of friends. People mostly come from the fictional village of Ambridge in the county of Dorsetshire, UK or nearby. Some of the group often meet at the village pub (The Bull) to socialise. Chat will cover a wide range of topics, including village gossip, current affairs and family news, as well as making arrangements and sharing personal news. People may also mention locations in the main county town of Borechester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham. There are some general jokes and light banter. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-friends","channels":["friends"],"daysOfWeek":"MONDAY-SUNDAY","volumes":["10:1:12:2:13:0:15:2:17:3:20:1:23:0"]]
personal-football	Chat between a group of friends who share an interest in football. People mostly come from the fictional village of Ambridge in the county of Dorsetshire, UK, or nearby. The group often meet at the village pub (The Bull) to socialise and sometimes watch matches, but also watch the local team and some travel to league matches. Topics generally relate to football, but the chat will often veer to other topics including village gossip, current affairs and family news. As well as the football clubs listed in the entities document, people may also mention locations in the main county town of Borechester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham. There are some general jokes and light banter. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-football","channels":["football"],"daysOfWeek":"MONDAY-SUNDAY","volumes":["13:1:15:2:17:4:20:2:23:0"]]
personal-golf	Chat between a group of friends who play golf together. People mostly come from the fictional village of Ambridge in the county of Dorsetshire, UK or nearby. The group socialise together around golf, but also the village pub. The topics generally relate to arranging meetups to play golf, but the chat will often veer to other topics including village gossip, current affairs and family news. As well as the golf clubs listed in the entities document, people may also mention locations in the main county town of Borechester, as well as, Penny Hassett, Loxley Barrett, Darrington, and the cathedral city of Felpersham. There are some general jokes and light banter. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-golf","channels":["golf"],"daysOfWeek":"MONDAY-SUNDAY","volumes":["11:1:14:2:17:1:20:2:23:0"]]
personal-family	Chat between members of the family of our protagonist, Kenton. The chat is a lighthearted exchange between family members on a variety of topics, with some messages relating to logistics and meeting up. A small amount of messages should be of a type that might need a prompt or urgent response. Kenton and Kathy live in the village of Ambridge, and work in the county town of Borechester in Dorsetshire, while family are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. The tone of the messages should be informal British English.	/awag sim daily add ["category":"kenton-family","channels":["family"],"daysOfWeek":"MONDAY-SUNDAY","volumes":["8:1:11:2:14:1:17:2:20:3:23:0"]]
kenton-kathy	Chat between our fictional protagonist, Kenton and his wife Kathy Archer. The chat is a exchange a variety of topics, with some messages relating to logistics and meeting up but also discussing	/awag sim daily add ["category":"kenton-kathy","channels":["kathy"],"daysOfWeek":"MONDAY-SUNDAY"]

**Figure S10.18: Kenton – Data (Synthetic)**

**S10 Study Configuration & Information****S10.2.5 Usha****AwAg Data****Persona: Usha****ML Models**

Model ID	Categories	Model Desc	Command
work-pers	work personal	Does this item relate to my work or personal life?	/awag model create work-pers work,personal /awag model desc work-pers Does this item relate to my work or personal life?
urgency	urgent not	Is this item in relation to something that I think is urgent?	/awag model create urgency urgent/not /awag model desc urgency Is this item in relation to something that I think is urgent?
work-logistics	logistics not	Is this item related to logistics at work?	/awag model create work-logistics logistics/not /awag model desc work-logistics Is this item related to logistics, specifically in a work context?
riding	riding not	Is this item related to my hobby/interest of horse riding?	/awag model create riding riding/not /awag model desc riding Is this item related to my hobby/interest of horse riding?
riding-arrangements	arrangements not	Is this item about making arrangements to go horse riding or other riding related activities?	/awag model create riding-arrangements arrangements/not /awag model desc riding-arrangements Is this item about making arrangements to go horse riding or other riding related activities?
school-importance	school-important school-other not	Is this item related to organisation, specifically in the context of my Club's tennis arrangements and events?	/awag model create school-importance school-important,school-other/not /awag model desc school-importance Is this item related to my children's school, and if so is it important?
interested	work personal not	Am I interested in this item, from a work or a personal perspective?	/awag model create interested work,personal/not /awag model desc interested Am I interested in this item, from a work or a personal perspective?

Figure S10.19: Usha – ML Models

**AwAg Data****Persona: Usha****Data (RSS)**

ID	URL	Command
theguardian-uknews	<a href="https://www.theguardian.com/uk-news/rss">https://www.theguardian.com/uk-news/rss</a>	/awag rss sub create theguardian-uknews https://www.theguardian.com/uk-news/rss
theguardian-law	<a href="https://www.theguardian.com/law/rss">https://www.theguardian.com/law/rss</a>	/awag rss sub create theguardian-law <a href="https://www.theguardian.com/law/rss">https://www.theguardian.com/law/rss</a>
reddit-uklaw	<a href="https://www.reddit.com/r/uklawrss">https://www.reddit.com/r/uklawrss</a>	/awag rss sub create reddit-uklaw <a href="https://www.reddit.com/r/uklawrss">https://www.reddit.com/r/uklawrss</a>
theguardian-equestrianism	<a href="https://www.theguardian.com/sport/equestrianism/rss">https://www.theguardian.com/sport/equestrianism/rss</a>	/awag rss sub create theguardian-equestrianism <a href="https://www.theguardian.com/sport/equestrianism/rss">https://www.theguardian.com/sport/equestrianism/rss</a>
theguardian-horseracing	<a href="https://www.theguardian.com/sport/horse-racing/rss">https://www.theguardian.com/sport/horse-racing/rss</a>	/awag rss sub create theguardian-horseracing <a href="https://www.theguardian.com/sport/horse-racing/rss">https://www.theguardian.com/sport/horse-racing/rss</a>

Figure S10.20: Usha – Data (RSS)

**S10 Study Configuration & Information****AwAg Data****Persona: Usha****Data (Synthetic)**

ID	Topic Text	Command
work-office	Company messenger application chat at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. This chat is general workplace discussion among staff. Topics should include discussions related to work in general but not specific client details. This can include cover, logistical arrangements, daily office opening and closing, lunch arrangements, work social events and other topics that you might expect to be discussed in a small company. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings	/awag sim daily add ["category":"usha-work-office","channels":["work-office"], "daysOfWeek": "MONDAY-FRIDAY", "volumes": "18:3:12:2:15:3:18:1:20:0"]
work-partners	Company messenger application chat at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. This chat is discussion between the three partners including Usha. Topics in this chat should include discussions related to the individuals work as partners of the practice. This can include the strategy and performance of the partnership, staff issues, problems or opportunities with clients, concern about competitors and other leadership topics. Discussion of client issues should be at a higher level, rather than details, focussing only on things that might affect the client relationship - the partners have different areas of specialisation and would not generally discuss day to day client matters amongst themselves. As it is a private chat between the partners they may be discussing sensitive topics. Some messages may concern social topics or trivia. A subset of these messages should be of the type that might expect a prompt response. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add ["category":"usha-work-partners","channels":["work-partners"], "daysOfWeek": "MONDAY-SUNDAY", "volumes": "18:1:12:2:15:3:18:1:23:0"]
work-pa	Company messenger application chat at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. This chat is discussion between Usha and her personal assistant Evan Thomas. Only include messages from these two people. Topics in this chat should include discussions related to Usha's client work, organising her diary and other similar items that would come up between a partner in a solicitor's practice and her personal assistant. Some messages may concern social topics or trivia. A subset of the messages should be of the type that might expect a prompt response. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add ["category":"usha-work-pa","channels":["work-pa"], "daysOfWeek": "MONDAY-FRIDAY", "volumes": "18:2:12:1:15:2:16:1:18:0"]
work-clients	Business communications (email/messages) at the Felpersham Law, the workplace of our fictional protagonist, Usha Franks. These messages are communications directly between Usha and various clients that she is working with. Topics in this chat should include discussions related to the clients' cases and other similar items that would come up between a partner in a solicitor's practice and her clients. A few messages may concern social topics or trivia where Usha has a long relationship with the client. A subset of the messages should be of the type that might expect a prompt response. While a list of some colleagues and clients is provided, you should also generate fictional names to reference in messages. Usha herself lives in the nearby village of Ambridge. Message content should be unique and not duplicated. Ensure that you produce incoming messages from clients. The tone of the messages should be British English with British spellings.	/awag sim daily add ["category":"usha-work-clients","channels":["work-clients"], "daysOfWeek": "MONDAY-SATURDAY", "volumes": "18:2:12:0:15:2:23:0"]
riding-chat	General chat between a group of friends with a shared interest in horse riding, including our protagonist, Usha Franks. Most of those on the chat are members of Felpersham Riding Club and some stable their horses there. Club members mostly come from the fictional villages of Ambridge and Darrington in the county of Dorsetshire, or nearby, although some come from the cathedral city of Felpersham where Usha practices as a solicitor. This is a social chat between friends and acquaintances that share a common interest. Topics generally relate to all aspects of recreational and competitive horse riding, with many messages also being about meeting up for rides. Some but not all ride competitively (e.g. dressage), while most are social riders. There are some non-horse related messages such as general jokes and light chat. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add ["category":"usha-riding-chat","channels":["riding-chat"], "daysOfWeek": "MONDAY-SUNDAY", "volumes": "18:1:12:2:15:1:17:3:20:1:23:0"]
riding-stables	General chat for people associated with Felpersham Riding Club and its stables, including our protagonist Usha Franks. The chat is used for communication between members of Felpersham Riding Club and also for announcements and information from club officials and stable managers. Usha stables her horse there. Club members mostly come from the fictional villages of Ambridge and Darrington in the county of Dorsetshire, or nearby, although some come from the cathedral city of Felpersham where Usha practices as a solicitor. Topics generally relate to all aspects of recreational and competitive horse riding, with many messages also being about club matters and announcements about competitions, events and happenings at the stables. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add ["category":"usha-riding-stables","channels":["riding-stables"], "daysOfWeek": "MONDAY-SUNDAY", "volumes": "18:1:12:2:18:1:22:0"]
family-close	General chat between members of a fictional British family, as seen by our protagonist, Usha Franks. The chat is a light-hearted exchange between immediate family members Usha, Alan, Amy and Ravi - do not include messages from any other people. Amy and Ravi should refer to Usha and Alan as 'Mum' and 'Dad'. Topics include anything a family might discuss, but can also include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. Usha and Alan live in the village of Ambridge, while wider family and friends are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham where Usha works as a partner in Felpersham Law. The tone of the messages should be informal British English with British spellings.	/awag sim daily add ["category":"usha-family-close","channels":["family-close"], "daysOfWeek": "MONDAY-SUNDAY", "volumes": "18:1:12:0:15:2:18:3:20:1:22:0"]
family-wide	General chat between members of a fictional British family, as seen by our protagonist, Usha Franks. The chat is a light-hearted exchange between immediate family members Usha, Alan, Amy and Ravi and wider family members including Deepak, Shiv and Satya - do not include messages from any other non-family. Topics include anything a family might discuss, but can also include village gossip, family personal and work happenings and milestones, and requests for assistance or favours. Usha and Alan live in the village of Ambridge, while wider family and friends are dotted around local villages Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham where Usha works as a partner in Felpersham Law. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spellings.	/awag sim daily add ["category":"usha-family-wide","channels":["family-wide"], "daysOfWeek": "MONDAY-SUNDAY", "volumes": "18:1:12:0:15:2:18:1:20:2:22:0"]
personal-village	General chat among residents of the fictional village of Ambridge in the county of Dorsetshire, UK, as seen by our protagonist, Usha Franks. The chat is a mostly light-hearted exchange between residents of Ambridge, covering topics including comings & goings around Ambridge, parking problems, postal service & parcel delivery (problems), recommendations for good tradespeople, social events, local crime and plenty of village gossip. Residents of Ambridge are mostly retired villagers who have moved there for a quiet life and individual arrangements. A few of the messages should be of a nature that merits a prompt response, such as a request for assistance or information. Nearby villages include Penny Hassett, Loxley Barrett, Darrington, Hollerton, Edgeley, Waterley Cross and Lakey Green. The main local city is the cathedral city of Felpersham. Some of these locations should occasionally be mentioned in messages. In addition to the supplied list of people, you should make up identities as needed - the chat forum would contain maybe 50-30 members (although not all of these post messages often). Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spelling.	/awag sim daily add ["category":"usha-village","channels":["village"], "daysOfWeek": "MONDAY-SUNDAY", "volumes": "18:1:12:0:15:2:18:1:22:0"]
school-comms	Messages originating from the school Felpersham Grammar in Felpersham, Dorsetshire, UK, where our protagonist, Usha Franks sends her children. All messages are announcements and notices from school leaders or the school office, directed at parents with children at the school. In addition to the supplied list of people, you should make up identities as needed. Message content should be unique and not duplicated. The tone of the messages should be relatively formal British English with British spelling.	/awag sim daily add ["category":"usha-school-comms","channels":["school-comms"], "daysOfWeek": "MONDAY-FRIDAY", "volumes": "18:1:12:0:15:1:19:0"]
school-parents-chat	Messenger chat between parents with children attending Felpersham Grammar School in Felpersham, Dorsetshire, UK, where our protagonist, Usha Franks sends her children. This is a chat group independent of the school where parents can discuss school and other topics. This could include compliments or more often complaints about the school as well as sharing helpful information and coordinating things. Sometimes the parents group meets up in Felpersham cafes. There are some occasional jokes and off topic discussion. Some parents have a less positive outlook than others and tend to complain, while others are more positive - generate a mix of positive and negative comments. Only parents are in this chat - neither Usha nor anyone else's children are on it. Message content should be unique and not duplicated. The tone of the messages should be informal British English with British spelling.	/awag sim daily add ["category":"usha-school-parents-chat","channels":["school-parents-chat"], "daysOfWeek": "MONDAY-FRIDAY", "volumes": "18:2:10:0:13:2:18:1:22:0"]

**Figure S10.21: Usha – Data (Synthetic)**

# **Supplement S11**

## **Study Detailed Results**

# **Study Detailed Results**

## **Contents**

- S11.1 – Statistics
  - S11.1.1 – Classification Agreement
  - S11.1.2 – Evaluation Feedback
  - S11.1.3 – Evaluation Agreement (Modes 1 & 2)
  - S11.1.4 – Evaluation Agreement (Mode 3)
  - S11.1.5 – Evaluation Difference
  - S11.1.6 – Evaluation Ratings
  - S11.1.7 – Evaluation Results
- S11.2 – Selection of Evaluation Items
  - S11.2.1 – Susan
  - S11.2.2 – Adam
  - S11.2.3 – Phoebe
  - S11.2.4 – Kenton
  - S11.2.5 – Usha
- S11.3 – Evaluation Token Usage

## S11 Study Detailed Results

---

### S11.1 Statistics

These statistics are derived from programmatically generated study instance-specific stats packs located at [doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /study/data/stats/generated]. These have been combined into consolidated cross-instance Microsoft Excel spreadsheets, from which the tables and figures here have been taken. These are located at:

[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /study/data/stats/combined].

#### S11.1.1 Classification Agreement

Table S11.1 shows the percentage agreement between UD-ML classifications and manual classifications performed by the study participant for each study instance. Rows are broken down by UD-ML model, with an ALL row containing data across all models. Note that not all UD-ML models were included in each study – each row shows aggregated data for a model for all studies that contained that model. The second column indicates the number of studies that each model appeared in.

##### Percent of Classification Manual Agrees

This table shows the percentage agreement between UD-ML classifications and manual classifications performed by the study participant. This is a measure of how correctly the UD-ML models classify items as it compares the ML decision with the human decision. This data only includes items for which the participant entered classification values during the training phase of the study. It is consolidated across all personas - some models feature in more than one persona study, as indicated in the '# column. The ALL column is calculated as a mean of the individual persona values.

Model	#	Susan			Adam			Phoebe			Kenton			Usha			ALL			
		Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	
ALL	4,025	3,648	90.6%	12,021	11,344	94.4%	3876	3535	91.2%	964	908	94.2%	5,157	4,726	91.6%	26,043	24,161	92.8%	All	
urgency	5	503	499	99.2%	1,277	1,259	98.6%	646	609	94.3%	135	133	98.5%	573	540	94.2%	3,134	3,040	97.0%	+4.2 ur
work-logistics	5	503	482	95.8%	1,277	1,250	97.9%	646	622	96.3%	135	131	97.0%	573	545	95.1%	3,134	3,030	96.7%	+3.9 w-l
work-pers	5	503	423	84.1%	1,277	1,210	94.8%	646	586	90.7%	135	121	89.6%	573	511	89.2%	3,134	2,851	91.0%	-1.8 w-p
interested	4	504	361	71.6%	1,277	1,186	92.9%				136	122	89.7%	573	462	80.6%	2,490	2,131	85.6%	-7.2 in
work-relevant	4	503	462	91.8%	1,277	1,134	88.8%	646	576	89.2%				573	527	92.0%	2,999	2,699	90.0%	-2.8 w-r
company-law	1													573	539	94.1%	573	539	94.1%	+1.3 co-l
cycling	1				1,277	1,180	92.4%								1,277	1,180	92.4%	-0.4 cy		
cycling-logistics	1				529	492	93.0%								529	492	93.0%	+0.2 cy-l		
football	1										141	127	90.1%				141	127	90.1%	-2.7 fo
friend-group	1							646	575	89.0%							646	575	89.0%	f-g
golf	1										141	139	98.6%				141	139	98.6%	+5.8 go
golf-logistics	1										141	135	95.7%				141	135	95.7%	+3.0 g-l
pers-urgency	1				1,276	1,208	94.7%										1,276	1,208	94.7%	+1.9 p-u
personal-interested	1							646	567	87.8%							646	567	87.8%	-5.0 p-i
riding	1													573	526	91.8%	573	526	91.8%	-1.0 ri
riding-arrangements	1										573	558	97.4%				573	558	97.4%	+4.6 r-a
school-importance	1													573	518	90.4%	573	518	90.4%	-2.4 s-i
tech	1				1,277	1,225	95.9%										1,277	1,225	95.9%	+3.2 tc
tennis	1	503	445	88.5%													503	445	88.5%	-4.3 tn
tennis-arrangements	1	503	490	97.4%													503	490	97.4%	+4.6 t-a
tennis-organising	1	503	486	96.6%													503	486	96.6%	+3.8 t-o
work-urgency	1				1,277	1,200	94.0%										1,277	1,200	94.0%	+1.2 w-u
count_items					Number of items in the dataset 'phase2' having a manual classification record (the numerator in the percentage calculation)															
count_agree					Number of items in the dataset 'phase2' having a manual classification record that agrees with the ML classification (the denominator in the percentage calculation)															
percentage_agree					Percentage of records in dataset 'phase2' having a manual classification, which agrees with the ML classification															

Table S11.1: Percent of Manual Classification Agreement with UD-ML

## S11 Study Detailed Results

---

Table S11.2 and Figure S11.1 show how the classification agreement percentage changed over time during the training phase of each study, which occurred over a period of up to 7 days for each study.

### Percent of Classification Manual Agrees (Time Series)

This table shows the percentage agreement between UD-ML classifications and manual classifications performed by the study participant. This is a measure of how correctly the UD-ML models classify items as it compares the ML decision with the human decision. This consolidated time series data shows how the measure changed over time during the study. This data only includes items for which the participant-entered classification values during the training phase of the study.

Day	Susan			Adam			Phoebe			Kenton			Usha			ALL		
	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree	Items	Agree	% Agree
Day 1	472	417	88.3%	1,359	1,244	91.5%	366	338	92.3%	204	189	92.6%	342	307	89.8%	2,743	2,495	91.0%
Day 2	600	539	89.8%	1,547	1,428	92.3%	420	363	86.4%	266	253	95.1%	522	459	87.9%	3,355	3,042	90.7%
Day 3	776	692	89.2%	2,205	2,089	94.7%	426	387	90.8%	98	90	91.8%	747	662	88.6%	4,252	3,920	92.2%
Day 4	72	68	94.4%	1,830	1,736	94.9%	132	118	89.4%	98	91	92.9%	1,080	995	92.1%	3,212	3,008	93.6%
Day 5	672	608	90.5%	2,680	2,532	94.5%	852	773	90.7%	258	247	95.7%	675	615	91.1%	5,137	4,775	93.0%
Day 6	1,024	947	92.5%	920	890	96.7%	1,392	1,285	92.3%	37	36	97.3%	693	651	93.9%	4,066	3,809	93.7%
Day 7	409	377	92.2%	1,440	1,385	96.2%	288	271	94.1%	3	2	66.7%	900	851	94.6%	3,040	2,886	94.9%

Table S11.2: Percent of Manual Classification Agreement with UD-ML time Series

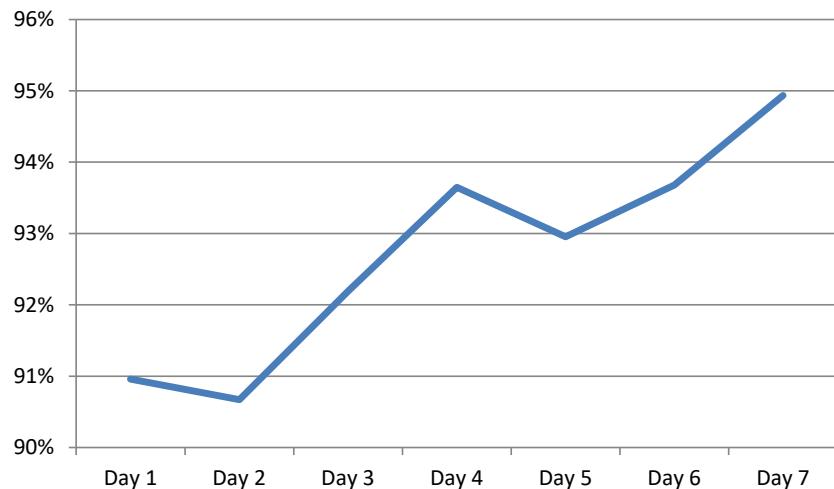


Figure S11.1: Percent of Manual Classification Agreement with UD-ML Time Series

## S11 Study Detailed Results

---

Table S11.3 and Figure S11.2 show the classification agreement data as a Cohen's Kappa.

### Cohen's Kappa For Manual Classification Agreement

This table shows Cohen's Kappa for agreement between UD-ML classifications and manual classifications performed by the study participant. This is a measure of how correctly the UD-ML models classify items as it compares the ML decision with the human decision. This data only includes items for which the participant-entered classification values during the training phase of the study. It is consolidated across all personas - some models feature in more than one persona study, as indicated in the '#' column. The ALL column is calculated as a mean of the individual persona values. The ALL row was calculated separately for each persona; it is not a mean of the individual model values.

Model	#	Susan	Adam	Phoebe	Kenton	Usha	ALL	
ALL		0.756	0.881	0.827	0.871	0.818	0.83	ALL
urgency	5	0.663	0.797	0.596	0.881	0.607	0.71	-0.12 ur
work-logistics	5	0.379	0.856	0.687	0.784	0.475	0.64	-0.19 w-l
work-pers	5	0.639	0.865	0.792	0.708	0.736	0.75	-0.08 w-p
interested	4	0.561	0.838		0.772	0.687	0.71	-0.12 in
work-relevant	4	0.000	0.462	0.68		0.766	0.48	-0.35 w-r
company-law	1					0.606	0.61	-0.22 co-l
cycling	1		0.790				0.79	-0.04 cy
cycling-logistics	1		0.555				0.56	-0.28 cy-l
football	1				0.791		0.79	-0.04 fo
friend-group	1			0.743			0.74	-0.09 f-g
golf	1				0.867		0.87	+0.04 go
golf-logistics	1				0.549		0.55	-0.28 g-l
pers-urgency	1		0.748				0.75	-0.08 p-u
personal-interested	1			0.696			0.70	-0.13 p-i
riding	1					0.725	0.73	-0.11 ri
riding-arrangements	1					0.638	0.64	-0.19 r-a
school-importance	1					0.427	0.43	-0.40 s-i
tech	1		0.765				0.77	-0.07 tc
tennis	1	0.475					0.48	-0.36 tn
tennis-arrangements	1	0.506					0.51	-0.32 t-a
tennis-organising	1	0.306					0.31	-0.52 t-o
work-urgency	1		0.642				0.64	-0.19 w-u

Note that the ALL column represents a MEAN of the Cohen's Kappa values for each of the Persona columns

Table S11.3: Cohen's Kappa for Manual Classification Agreement with UD-ML

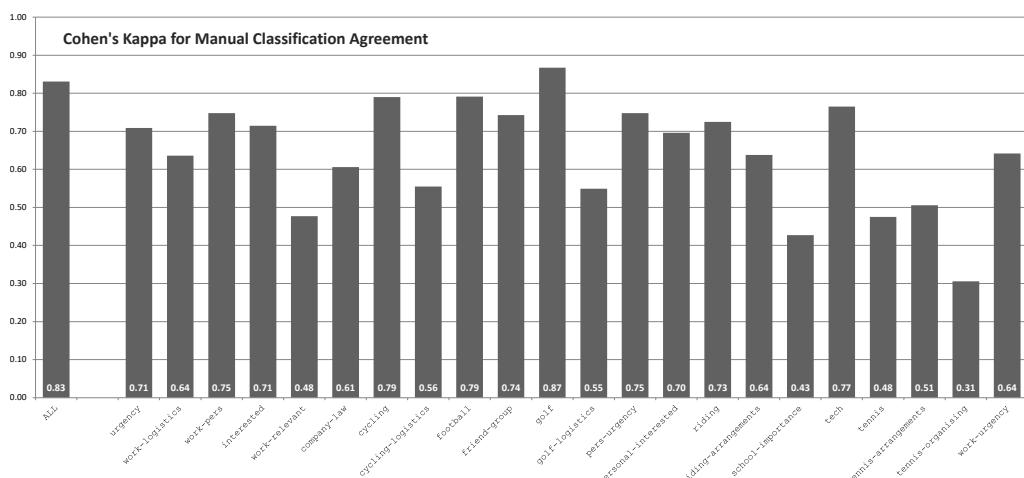


Figure S11.2: Cohen's Kappa for Manual Classification Agreement with UD-ML

## S11 Study Detailed Results

---

### S11.1.2 Evaluation Feedback

Table S11.4 details the Pearson's correlation coefficient ( $r$ ) between synthetic evaluations and participant-entered feedback, broken down by Persona, UD-ML Model and Evaluation Tag for Mode 1 and Mode 2 evaluations. Mean values of  $r$  have been calculated by applying a Fisher Z transformation before calculating a mean value and converting back to  $r$ . Only combinations of data having a significant correlation ( $p \leq 0.06$ ) value are included in the mean.

#### Pearson Correlation for Evaluation Feedback

This table contains values for each combination of Persona, UD-ML Model and Output Tag evaluated, showing the correlation between synthetic evaluations and participant-entered feedback. The higher the correlation, the more accurate the synthetic evaluation was. Mean values are calculated for rows and columns using only those  $r$  values having a significant correlation ( $p < 0.06$ ); other values are not included in means. Means are derived via a Fisher Z-Transformation. Only Mode 1 & 2 evaluations have feedback.

Persona	Model	vanilla-mode1-01		vanilla-mode1-02		vanilla-mode2-01		base-mode2-01		vanilla4-mode2-01		vanilla4o-mode2-01		Mean r
		r	p	r	p	r	p	r	p	r	p	r	p	
Susan	ALL	0.11	0.051	0.08	0.047	0.02	0.806	0.28	0.000	0.54	0.000	0.73	0.000	<b>0.38</b>
Susan	interested	0.44	0.004	0.07	0.545	0.25	0.161	0.38	0.044	0.54	0.005	0.63	0.000	<b>0.50</b>
Susan	tennis	0.38	0.011	0.16	0.149	-0.26	0.156	0.12	0.519	0.33	0.113	0.73	0.000	<b>0.58</b>
Susan	tennis-arrangements	0.20	0.374	-0.20	0.058	0.27	0.140	0.10	0.596	0.52	0.007	0.49	0.000	<b>0.29</b>
Susan	tennis-organising	-0.21	0.170	-0.13	0.226	-0.13	0.483	0.24	0.228	0.48	0.012	0.66	0.000	<b>0.58</b>
Susan	urgency	0.05	0.730	0.24	0.022	-0.07	0.711	0.00	0.000	0.41	0.037	0.85	0.000	<b>0.57</b>
Susan	work-logistics	0.14	0.371	0.30	0.005	0.02	0.933	-0.07	0.734	0.40	0.043	0.97	0.000	<b>0.74</b>
Susan	work-pers	0.16	0.312	-0.01	0.956	0.19	0.311	0.10	0.605	0.60	0.001	0.74	0.000	<b>0.68</b>
Susan	work-relevant	-0.17	0.266	0.24	0.025	-0.33	0.065	0.56	0.002	0.78	0.000	0.70	0.000	<b>0.45</b>
Adam	ALL	0.36	0.000	0.20	0.001	0.08	0.137	0.35	0.000	0.65	0.000	0.66	0.000	<b>0.46</b>
Adam	cycling	0.16	0.395	0.61	0.001	0.12	0.487	0.18	0.302	0.64	0.000	0.44	0.043	<b>0.57</b>
Adam	cycling-logistics	0.40	0.145	0.56	0.020	0.13	0.583	-0.13	0.568	0.49	0.011	0.60	0.009	<b>0.55</b>
Adam	interested	0.57	0.001	0.54	0.002	0.36	0.031	0.31	0.068	0.61	0.000	0.84	0.000	<b>0.57</b>
Adam	pers-urgency	0.35	0.058	0.13	0.497	0.07	0.666	0.31	0.070	0.57	0.000	0.61	0.003	<b>0.47</b>
Adam	tech	-0.11	0.560	0.19	0.321	-0.05	0.750	0.71	0.000	0.96	0.000	0.69	0.000	<b>0.84</b>
Adam	urgency	0.42	0.020	0.00	0.988	-0.07	0.688	0.53	0.001	0.35	0.026	0.49	0.022	<b>0.45</b>
Adam	work-logistics	0.81	0.000	0.25	0.204	-0.10	0.583	0.56	0.000	0.71	0.000	0.51	0.014	<b>0.66</b>
Adam	work-pers	-0.07	0.722	0.38	0.048	-0.12	0.494	0.43	0.015	0.68	0.000	0.64	0.002	<b>0.55</b>
Adam	work-relevant	0.40	0.028	-0.09	0.636	0.07	0.695	0.52	0.001	0.71	0.000	0.92	0.000	<b>0.70</b>
Adam	work-urgency	0.13	0.491	-0.18	0.341	-0.07	0.696	0.25	0.136	0.85	0.000	0.94	0.000	<b>0.90</b>
Phoebe	ALL									0.79	0.000	0.88	0.000	<b>0.84</b>
Phoebe	friend-group									0.66	0.002	0.81	0.000	<b>0.75</b>
Phoebe	personal-interested									0.88	0.000	0.61	0.001	<b>0.78</b>
Phoebe	urgency									0.53	0.016	0.69	0.000	<b>0.62</b>
Phoebe	work-logistics									0.80	0.000	0.90	0.000	<b>0.86</b>
Phoebe	work-pers									0.85	0.000	0.90	0.000	<b>0.88</b>
Phoebe	work-relevant									0.83	0.000	0.96	0.000	<b>0.91</b>
Kenton	ALL	0.24	0.000	0.20	0.000	0.06	0.278	0.34	0.000	0.84	0.000	0.73	0.000	<b>0.53</b>
Kenton	football	0.41	0.001	0.29	0.001	-0.06	0.676	0.28	0.067	1.00	0.000	0.74	0.000	<b>0.59</b>
Kenton	golf	-0.06	0.621	-0.07	0.429	0.02	0.872	0.51	0.000	0.00	0.000	0.28	0.014	<b>0.40</b>
Kenton	golf-logistics	0.00	0.975	0.21	0.021	0.04	0.774	0.47	0.002	0.00	0.000	0.21	0.074	<b>0.34</b>
Kenton	interested	0.30	0.016	0.25	0.005	0.10	0.523	0.14	0.379	0.78	0.000	0.87	0.000	<b>0.62</b>
Kenton	urgency	0.50	0.000	0.37	0.000	-0.11	0.468	0.60	0.000	0.72	0.000	0.69	0.000	<b>0.59</b>
Kenton	work-logistics	0.38	0.003	0.11	0.233	-0.02	0.908	0.01	0.963	0.67	0.000	0.86	0.000	<b>0.68</b>
Kenton	work-pers	0.14	0.267	0.26	0.003	0.01	0.970	0.19	0.288	0.94	0.000	0.72	0.000	<b>0.75</b>
Usha	ALL									0.70	0.000	0.82	0.000	<b>0.77</b>
Usha	company-law									0.40	0.008	0.87	0.000	<b>0.71</b>
Usha	interested									0.79	0.000	0.74	0.000	<b>0.76</b>
Usha	riding									0.91	0.000	0.79	0.000	<b>0.86</b>
Usha	riding-arrangements									0.83	0.000	0.83	0.000	<b>0.83</b>
Usha	school-importance									0.30	0.054	0.81	0.000	<b>0.61</b>
Usha	urgency									0.79	0.000	0.75	0.000	<b>0.77</b>
Usha	work-logistics									0.90	0.000	0.79	0.000	<b>0.85</b>
Usha	work-pers									0.60	0.000	0.92	0.000	<b>0.81</b>
Usha	work-relevant									0.83	0.000	0.75	0.000	<b>0.79</b>
Model label:		V-M1-A		V-M1-B		V-M2		B-M2		V4-M2		V4o-M2		
Percentage of significant rows:		31%		36%		4%		33%		93%		98%		
Mean r (significant only, excl. ALL):		<b>0.45</b>		<b>0.31</b>		<b>0.01</b>		<b>0.49</b>		<b>0.68</b>		<b>0.74</b>		
Mean r (significant only, ALL only):		<b>0.24</b>		<b>0.16</b>		n/a		<b>0.32</b>		<b>0.70</b>		<b>0.76</b>		

Table S11.4:  $r$  for Evaluation Feedback by Persona, UD-ML Model and Evaluation Tag

## S11 Study Detailed Results

---

Figure S11.3 shows the mean correlation ( $r$ ) between Modes 1 and 2 synthetic evaluation and participant feedback for all personas and UD-ML models by evaluation tag; this data is sourced from Table S11.4.

Figure S11.4 shows the percentages of rows in Table S11.4 for each evaluation tag having a significant correlation ( $p \leq 0.06$ ). Only these row values are used to calculate the mean data that is illustrated in Figures S11.5.

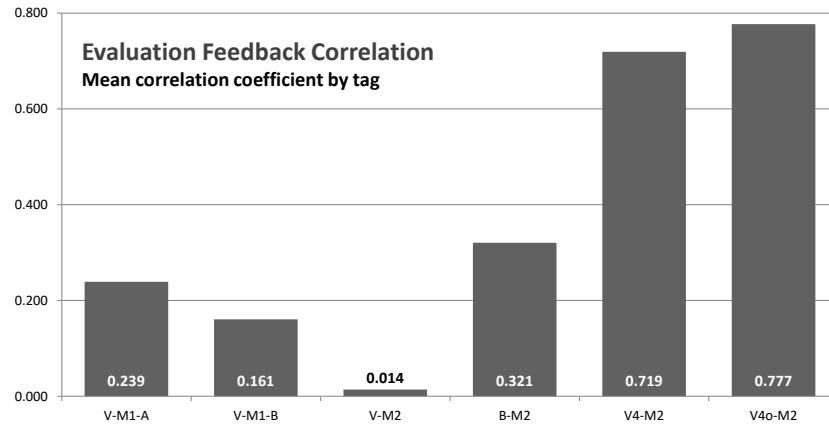


Figure S11.3: Mean Synthetic vs Participant Evaluation  $r$  by Evaluation Tag

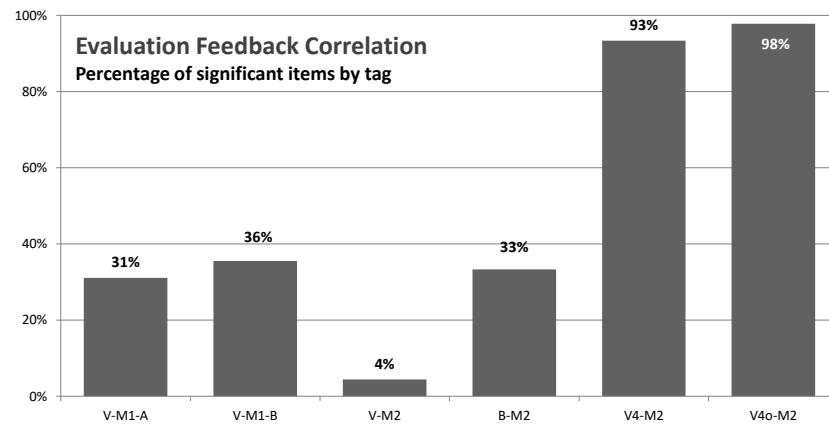


Figure S11.4: Percentage of Items Having Significant Correlation ( $r$ ) by Evaluation Tag

## S11 Study Detailed Results

---

Figures S11.5a to S11.5e show the mean correlation ( $r$ ) between Modes 1 and 2 synthetic evaluation and participant feedback by UD-ML Model for each persona; this data is sourced from Table S11.4.

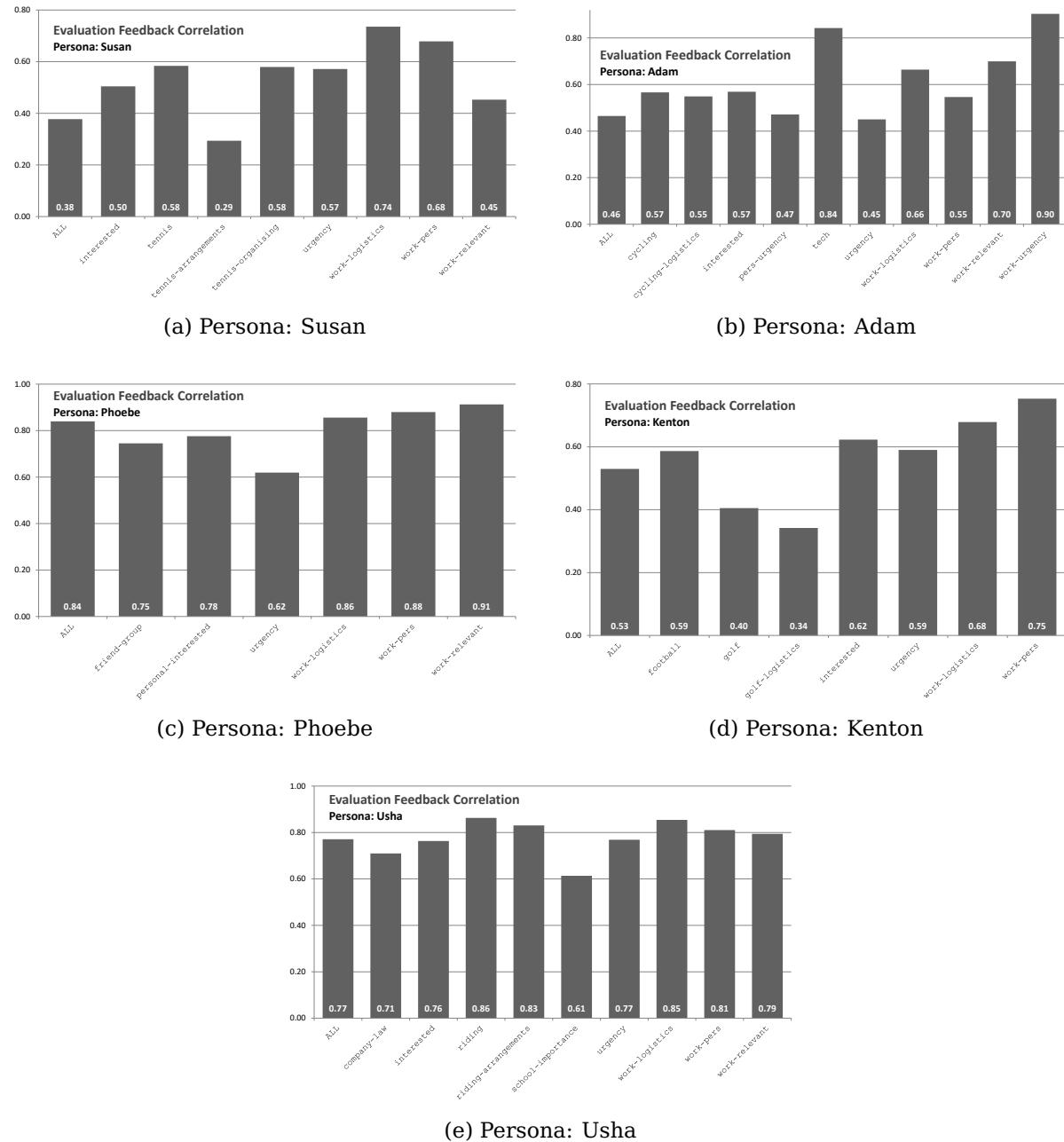


Figure S11.5: Synthetic vs Participant Evaluation  $r$  by UD-ML Model

## S11 Study Detailed Results

---

### S11.1.3 Evaluation Agreement (Modes 1 & 2)

Table S11.5 details the point-biserial correlation coefficient ( $r_{pb}$ ) between Mode 1 and Mode 2 synthetic evaluations and classification actions performed by the study participant during the training phase. In this case,  $r_{pb}$  is used because Mode 1 and Mode 2 evaluations have a categorical value (1-5 on the Likert scale) while classification actions are dichotomous (agree/disagree). The data is broken down by Persona, UD-ML Model and Evaluation Tag. Mean values of  $r_{pb}$  have been calculated by applying a Fisher Z transformation before calculating a mean value and converting back to  $r_{pb}$ . Only combinations of data having a significant correlation ( $p \leq 0.06$ ) value are included in the mean.

#### Point Biserial Correlation for Evaluation Agreement

This table contains values for each combination of Persona, UD-ML Model and Output Tag evaluated, showing the Point Biserial correlation between synthetic evaluations and classification actions entered by the participant. The higher the correlation, the more accurate the synthetic evaluation was. Mean values are calculated for rows and columns using only those r values having a significant correlation ( $p < 0.06$ ); other values are not included in means. Means are derived via a Fisher Z-Transformation. Only Mode 1 & 2 evaluations have feedback, and this data only includes items for which the participant entered classification values during the training phase of the study.

Persona	Model	vanilla-mode1-01		vanilla-mode1-02		vanilla-mode2-01		base-mode2-01		vanilla4-mode2-01		vanilla4o-mode2-01		Mean r	
		rbp	p	rbp	p	rbp	p	rbp	p	rbp	p	rbp	p		
Susan	ALL	0.12	0.000	0.03	0.282	0.00	0.989	0.11	0.002	0.36	0.000	0.50	0.000	<b>0.28</b>	ALL
Susan	interested	0.12	0.128	0.02	0.796	0.03	0.718	-0.01	0.937	0.32	0.001	0.38	0.000	<b>0.35</b>	+0.07 in
Susan	tennis	0.10	0.192	0.15	0.064	-0.15	0.048	0.07	0.472	0.46	0.000	0.48	0.000	<b>0.25</b>	-0.03 tn
Susan	tennis-arrangements	0.09	0.321	-0.10	0.179	-0.01	0.878	0.21	0.041	0.38	0.000	0.38	0.000	<b>0.32</b>	+0.04 t-a
Susan	tennis-organising	0.02	0.809	-0.02	0.798	-0.03	0.725	-0.03	0.774	-0.05	0.609	0.72	0.000	<b>0.72</b>	+0.44 t-o
Susan	urgency	-0.06	0.449	-0.04	0.593	-0.07	0.379	0.00	0.000	0.00	0.000	0.17	0.045	<b>0.17</b>	-0.11 ur
Susan	work-logistics	0.21	0.007	0.21	0.011	-0.03	0.729	-0.04	0.682	0.14	0.170	0.63	0.000	<b>0.37</b>	+0.09 w-l
Susan	work-pers	0.14	0.075	0.00	0.964	0.12	0.112	-0.23	0.030	0.27	0.008	0.38	0.000	<b>0.15</b>	-0.13 w-p
Susan	work-relevant	-0.02	0.830	0.14	0.080	-0.10	0.189	0.38	0.000	0.38	0.000	0.54	0.000	<b>0.43</b>	+0.15 w-r
Adam	ALL	0.14	0.000	0.04	0.018	0.00	0.949	0.12	0.000	0.24	0.000	0.41	0.000	<b>0.19</b>	ALL
Adam	cycling	-0.05	0.422	0.06	0.220	0.11	0.088	0.35	0.000	0.47	0.000	0.47	0.000	<b>0.43</b>	+0.24 cy
Adam	cycling-logistics	0.27	0.008	0.24	0.001	0.14	0.169	-0.14	0.350	0.32	0.001	0.45	0.000	<b>0.32</b>	+0.13 c-l
Adam	interested	0.27	0.000	0.14	0.003	0.17	0.010	0.11	0.265	0.18	0.005	0.29	0.000	<b>0.21</b>	+0.02 in
Adam	pers-urgency	0.26	0.000	0.01	0.767	-0.06	0.380	0.15	0.127	0.09	0.181	0.41	0.000	<b>0.33</b>	+0.14 p-u
Adam	tech	0.05	0.468	0.01	0.788	-0.05	0.454	0.13	0.190	0.58	0.000	0.68	0.000	<b>0.63</b>	+0.44 tc
Adam	urgency	0.11	0.087	0.05	0.268	-0.01	0.935	0.32	0.001	-0.04	0.583	0.16	0.002	<b>0.24</b>	+0.05 ur
Adam	work-logistics	0.18	0.006	-0.05	0.260	-0.09	0.200	-0.02	0.818	0.37	0.000	0.49	0.000	<b>0.35</b>	+0.16 w-l
Adam	work-pers	0.03	0.682	0.05	0.298	-0.03	0.676	-0.08	0.433	0.21	0.001	0.44	0.000	<b>0.33</b>	+0.14 w-p
Adam	work-relevant	0.20	0.002	0.10	0.026	-0.01	0.916	0.21	0.031	0.16	0.014	0.36	0.000	<b>0.21</b>	+0.02 w-r
Adam	work-urgency	0.10	0.118	-0.04	0.418	-0.16	0.018	0.10	0.330	0.21	0.001	0.39	0.000	<b>0.15</b>	-0.04 w-u
Phoebe	ALL									0.33	0.000	0.45	0.000	<b>0.39</b>	ALL
Phoebe	friend-group									0.29	0.000	0.40	0.000	<b>0.34</b>	-0.05 f-g
Phoebe	personal-interested									0.25	0.001	0.32	0.000	<b>0.28</b>	-0.11 p-i
Phoebe	urgency									0.25	0.001	0.47	0.000	<b>0.36</b>	-0.03 ur
Phoebe	work-logistics									0.25	0.001	0.50	0.000	<b>0.38</b>	-0.01 w-l
Phoebe	work-pers									0.44	0.000	0.46	0.000	<b>0.45</b>	+0.06 w-p
Phoebe	work-relevant									0.44	0.000	0.56	0.000	<b>0.50</b>	+0.11 w-r
Kenton	ALL	-0.03	0.653	-0.01	0.919	-0.10	0.120	0.14	0.002	0.46	0.000	0.58	0.000	<b>0.41</b>	ALL
Kenton	football	-0.13	0.473	0.21	0.254	-0.21	0.242	0.33	0.003	0.63	0.000	0.86	0.000	<b>0.66</b>	+0.25 fo
Kenton	golf	0.27	0.129	-0.02	0.932	0.13	0.461	0.16	0.176	0.70	0.000	0.83	0.000	<b>0.78</b>	+0.37 go
Kenton	golf-logistics	0.03	0.889	0.29	0.114	0.02	0.901	0.42	0.000	1.00	0.000	0.89	0.000	<b>0.74</b>	+0.34 g-l
Kenton	interested	-0.05	0.770	-0.10	0.585	-0.11	0.544	-0.07	0.573	0.22	0.156	0.32	0.004	<b>0.32</b>	-0.09 in
Kenton	urgency	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.57	0.000	0.36	0.001	<b>0.47</b>	+0.06 ur
Kenton	work-logistics	0.51	0.003	-0.13	0.500	-0.22	0.227	-0.04	0.709	0.39	0.009	0.46	0.000	<b>0.46</b>	+0.05 w-l
Kenton	work-pers	-0.36	0.047	-0.15	0.430	-0.41	0.022	-0.04	0.726	0.44	0.003	0.53	0.000	<b>0.06</b>	-0.34 w-p
Usha	ALL									0.40	0.000	0.50	0.000	<b>0.45</b>	ALL
Usha	company-law									0.26	0.022	0.63	0.000	<b>0.47</b>	+0.01 c-l
Usha	interested									0.34	0.002	0.42	0.000	<b>0.38</b>	-0.07 in
Usha	riding									0.53	0.000	0.62	0.000	<b>0.57</b>	+0.12 ri
Usha	riding-arrangements									-0.04	0.755	-0.04	0.706	<b>0.00</b>	-0.45 r-a
Usha	school-importance									0.44	0.000	0.46	0.000	<b>0.45</b>	-0.00 s-i
Usha	urgency									0.06	0.631	0.56	0.000	<b>0.56</b>	+0.10 ur
Usha	work-logistics									0.27	0.017	0.45	0.000	<b>0.37</b>	-0.09 w-l
Usha	work-pers									0.52	0.000	0.60	0.000	<b>0.56</b>	+0.10 w-p
Usha	work-relevant									0.35	0.002	0.22	0.061	<b>0.28</b>	-0.17 w-r
<b>Model label:</b>		V-M1-A		V-M1-B		V-M2		B-M2		V4-M2		V4o-M2			
<b>Percentage of significant rows:</b>		24%		13%		9%		24%		82%		98%			
<b>Mean rpb (significant only, excl. ALL):</b>		<b>0.19</b>		<b>0.17</b>		<b>-0.14</b>		<b>0.25</b>		<b>0.40</b>		<b>0.48</b>			
<b>Mean rpb (significant only, ALL only):</b>		<b>0.13</b>		<b>0.04</b>		n/a		<b>0.12</b>		<b>0.36</b>		<b>0.49</b>			

Table S11.5:  $r_{pb}$  for Evaluation Agreement by Persona, UD-ML Model and Evaluation Tag

## S11 Study Detailed Results

---

Figure S11.6 shows the mean correlation ( $r_{pb}$ ) between Modes 1 and 2 synthetic evaluation and participant classification actions for all personas and UD-ML models by evaluation tag; this data is sourced from Table S11.5.

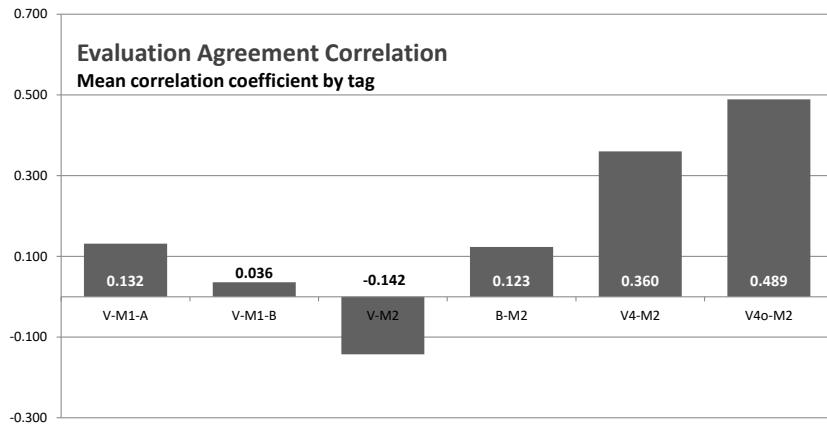


Figure S11.6: Mean Synthetic Evaluation vs Participant Classification  $r_{pb}$  by Eval Tag

Figure S11.7 shows the percentages of rows in Table S11.5 for each evaluation tag having a significant correlation ( $p \leq 0.06$ ). Only these row values are used to calculate the mean data that is illustrated in Figures S11.8.

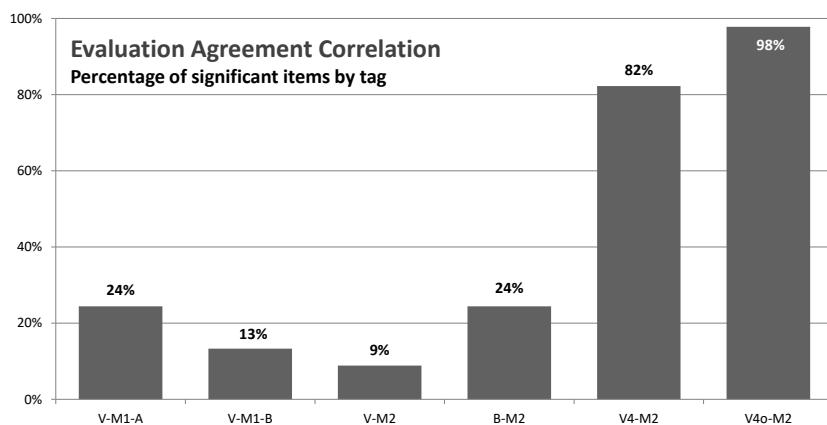


Figure S11.7: Percentage of Items Having Significant Correlation ( $r_{pb}$ ) by Evaluation Tag

## S11 Study Detailed Results

---

Figures S11.8a to S11.8e show the mean correlation ( $r_{pb}$ ) between Modes 1 and 2 synthetic evaluation and training classifications by UD-ML Model for each persona; this data is sourced from Table S11.5.

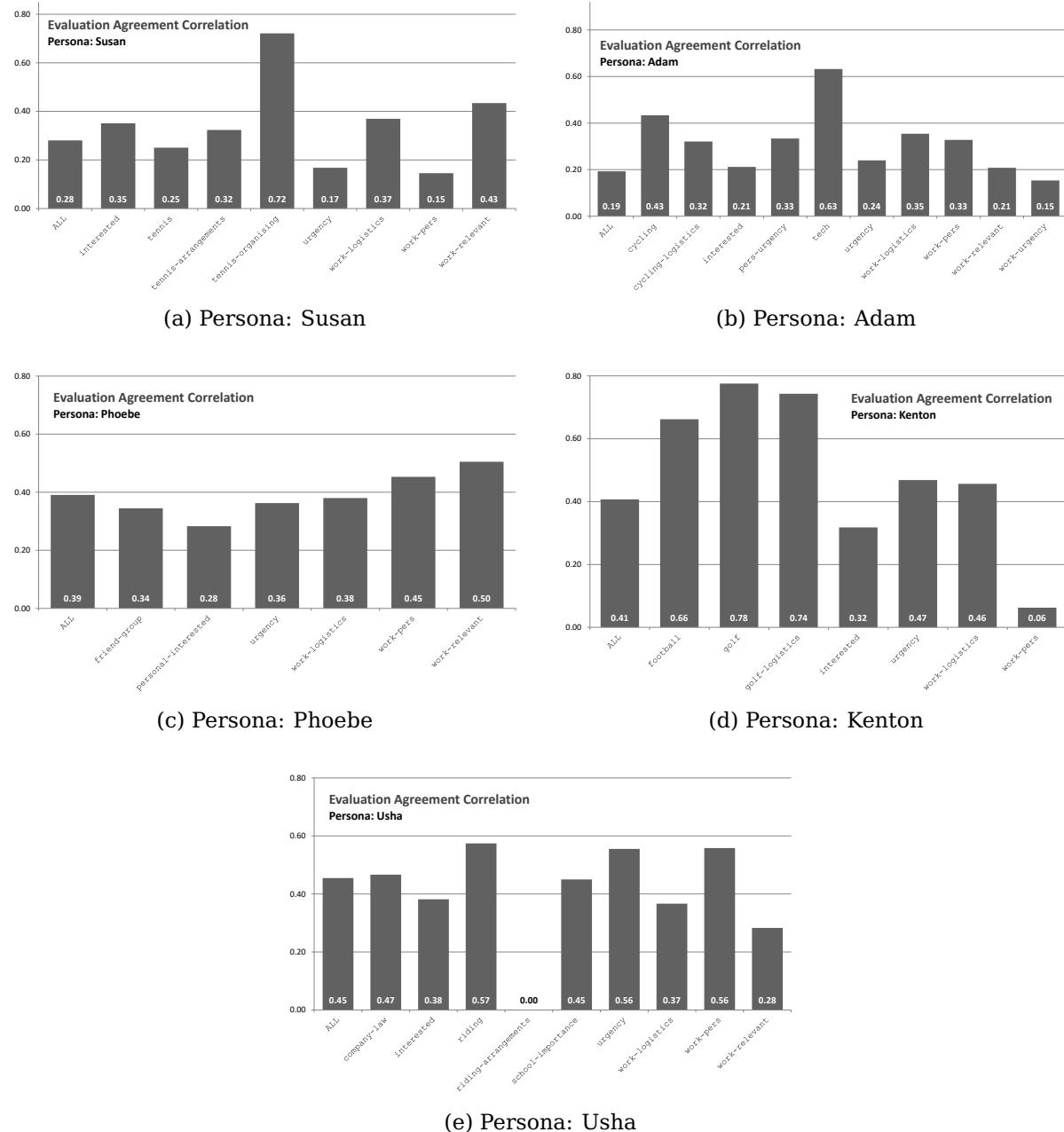


Figure S11.8: Synthetic vs Participant Evaluation  $r_{pb}$  by UD-ML Model

## S11 Study Detailed Results

---

### S11.1.4 Evaluation Agreement (Mode 3)

Table S11.6 details the phi coefficient ( $r_\phi$ ) between Mode 3 synthetic evaluations and classification actions performed by the study participant during the training phase. In this case,  $r_\phi$  is used because both Mode 3 evaluations and training classification actions are dichotomous (agree/disagree). The data is broken down by Persona, UD-ML Model and Evaluation Tag. Mean values of  $r_\phi$  have been calculated by applying a Fisher Z transformation before calculating a mean value and converting back to  $r_\phi$ . Only combinations of data having a significant correlation ( $p \leq 0.06$ ) value are included in the mean.

Figure S11.9 shows the mean correlation ( $r_\phi$ ) between Mode 3 synthetic evaluation and participant classification actions for all personas and UD-ML models by evaluation tag; this data is sourced from Table S11.6.

Figure S11.10 shows the percentages of rows in Table S11.6 for each evaluation tag having a significant correlation ( $p \leq 0.06$ ). Only these row values are used to calculate the mean data that is illustrated in Figures S11.11.

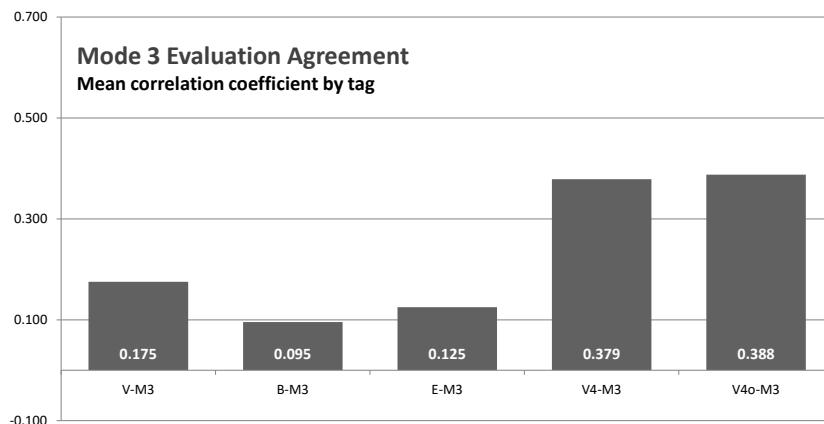


Figure S11.9: Mean Synthetic Evaluation vs Participant Classification  $r_\phi$  by Eval Tag

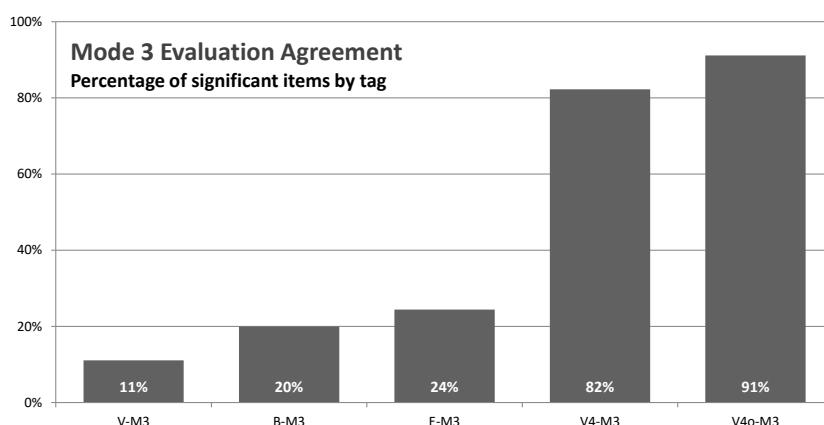


Figure S11.10: Percentage of Items Having Significant Correlation ( $r_\phi$ ) by Evaluation Tag

## S11 Study Detailed Results

---

### Phi Coefficient For Mode3 Agreement

This table contains values for each combination of Persona, UD-ML Model and Output Tag evaluated, showing the phi coefficient between Mode 3 synthetic evaluations and classification actions entered by the participant. The higher the coefficient, the more accurate the synthetic evaluation was. Mean values are calculated for rows and columns using only those phi values having a significant correlation ( $p < 0.06$ ); other values are not included in means. Means are derived via a Fisher Z-Transformation. Only Mode 3 evaluations have dichotomous values, and this data only includes items for which the participant entered classification values during the training phase of the study.

Persona	Model	vanilla-mode3-01		base-mode3-01		ext-mode3-01		vanilla4-mode3-01		vanilla4o-mode3-01		Mean r
		phi	p	phi	p	phi	p	phi	p	phi	p	
Susan	ALL	0.01	0.583	0.10	0.000	0.16	0.000	0.44	0.000	0.44	0.000	<b>0.29</b>
Susan	interested	0.04	0.495	0.07	0.282	0.03	0.680	0.46	0.000	0.33	0.000	<b>0.40</b>
Susan	tennis	0.02	0.762	0.00	1.000	0.14	0.044	0.47	0.000	0.65	0.000	<b>0.44</b>
Susan	tennis-arrangements	0.37	0.000	0.04	0.537	0.09	0.152	0.24	0.004	0.45	0.000	<b>0.36</b>
Susan	tennis-organising	0.01	0.829	0.00	1.000	0.00	0.950	0.12	0.149	0.20	0.003	<b>0.11</b>
Susan	urgency	0.11	0.056	0.19	0.005	0.00	1.000	0.13	0.111	0.00	1.000	<b>0.15</b>
Susan	work-logistics	0.15	0.013	0.10	0.137	0.11	0.102	0.38	0.000	0.33	0.000	<b>0.29</b>
Susan	work-pers	0.03	0.670	0.10	0.158	0.07	0.331	0.34	0.000	0.30	0.000	<b>0.24</b>
Susan	work-relevant	0.12	0.045	0.14	0.036	0.18	0.009	0.61	0.000	0.41	0.000	<b>0.34</b>
Adam	ALL	0.01	0.635	0.08	0.000	0.09	0.000	0.26	0.000	0.30	0.000	<b>0.18</b>
Adam	cycling	0.00	1.000	0.07	0.211	0.15	0.016	0.20	0.001	0.45	0.000	<b>0.27</b>
Adam	cycling-logistics	0.00	1.000	0.21	0.008	0.07	0.418	0.46	0.000	0.54	0.000	<b>0.41</b>
Adam	interested	0.04	0.430	0.00	1.000	0.04	0.550	0.23	0.000	0.14	0.001	<b>0.18</b>
Adam	pers-urgency	0.01	0.914	0.04	0.424	0.09	0.144	0.08	0.202	0.31	0.000	<b>0.31</b>
Adam	tech	0.00	1.000	0.12	0.021	0.00	1.000	0.52	0.000	0.47	0.000	<b>0.38</b>
Adam	urgency	0.04	0.444	0.00	1.000	0.11	0.062	0.04	0.542	0.19	0.000	<b>0.15</b>
Adam	work-logistics	0.00	1.000	0.01	0.889	0.00	1.000	0.26	0.000	0.44	0.000	<b>0.35</b>
Adam	work-pers	0.07	0.194	0.07	0.216	0.04	0.624	0.30	0.000	0.23	0.000	<b>0.27</b>
Adam	work-relevant	0.12	0.018	0.21	0.000	0.11	0.053	0.14	0.019	0.28	0.000	<b>0.18</b>
Adam	work-urgency	0.01	0.815	0.05	0.325	0.06	0.311	0.24	0.000	0.29	0.000	<b>0.26</b>
Phoebe	ALL							0.35	0.000	0.35	0.000	<b>0.35</b>
Phoebe	friend-group							0.29	0.000	0.37	0.000	<b>0.33</b>
Phoebe	personal-interested							0.24	0.000	0.21	0.001	<b>0.22</b>
Phoebe	urgency							0.22	0.001	0.41	0.000	<b>0.32</b>
Phoebe	work-logistics							0.28	0.000	0.39	0.000	<b>0.33</b>
Phoebe	work-pers							0.37	0.000	0.22	0.001	<b>0.30</b>
Phoebe	work-relevant							0.59	0.000	0.55	0.000	<b>0.57</b>
Kenton	ALL	0.07	0.084	0.11	0.002	0.13	0.001	0.47	0.000	0.41	0.000	<b>0.29</b>
Kenton	football	0.02	0.864	0.00	1.000	0.05	0.609	0.71	0.000	0.75	0.000	<b>0.73</b>
Kenton	golf	0.00	1.000	0.02	0.864	0.23	0.017	0.27	0.014	0.28	0.005	<b>0.26</b>
Kenton	golf-logistics	0.02	0.828	0.26	0.006	0.44	0.000	0.74	0.000	0.71	0.000	<b>0.57</b>
Kenton	interested	0.13	0.250	0.02	0.812	0.18	0.076	0.25	0.028	0.15	0.124	<b>0.25</b>
Kenton	urgency	0.00	1.000	0.00	1.000	0.00	1.000	0.13	0.278	0.10	0.318	<b>0.10</b>
Kenton	work-logistics	0.00	1.000	0.00	1.000	0.02	0.854	0.18	0.122	0.31	0.002	<b>0.00</b>
Kenton	work-pers	0.05	0.671	0.00	1.000	0.05	0.590	0.41	0.000	0.28	0.005	<b>0.35</b>
Usha	ALL							0.36	0.000	0.43	0.000	<b>0.40</b>
Usha	company-law							0.41	0.000	0.41	0.000	<b>0.41</b>
Usha	interested							0.30	0.006	0.35	0.001	<b>0.32</b>
Usha	riding							0.25	0.020	0.46	0.000	<b>0.36</b>
Usha	riding-arrangements							0.08	0.429	0.34	0.001	<b>0.34</b>
Usha	school-importance							0.20	0.065	0.32	0.002	<b>0.32</b>
Usha	urgency							0.10	0.368	0.15	0.151	<b>0.10</b>
Usha	work-logistics							0.20	0.061	0.13	0.228	<b>0.20</b>
Usha	work-pers							0.35	0.001	0.48	0.000	<b>0.42</b>
Usha	work-relevant							0.27	0.011	0.31	0.003	<b>0.29</b>
<b>Model label:</b>		V-M3		B-M3		E-M3		V4-M3		V4o-M3		
<b>Percentage of significant rows:</b>		11%		20%		24%		82%		91%		
<b>Mean phi (significant only, excl. ALL):</b>		<b>0.17</b>		<b>0.19</b>		<b>0.16</b>		<b>0.35</b>		<b>0.37</b>		
<b>Mean phi (significant only, ALL only):</b>		n/a		<b>0.10</b>		<b>0.13</b>		<b>0.38</b>		<b>0.39</b>		

Table S11.6:  $r_\phi$  for Mode 3 Evaluation Agreement by Persona, UD-ML Model and Evaluation Tag

## S11 Study Detailed Results

---

Figures S11.11a to S11.11e show the mean coefficient ( $r_\phi$ ) between Mode 3 synthetic evaluation and participant training classifications by UD-ML Model for each persona; this data is sourced from Table S11.6.

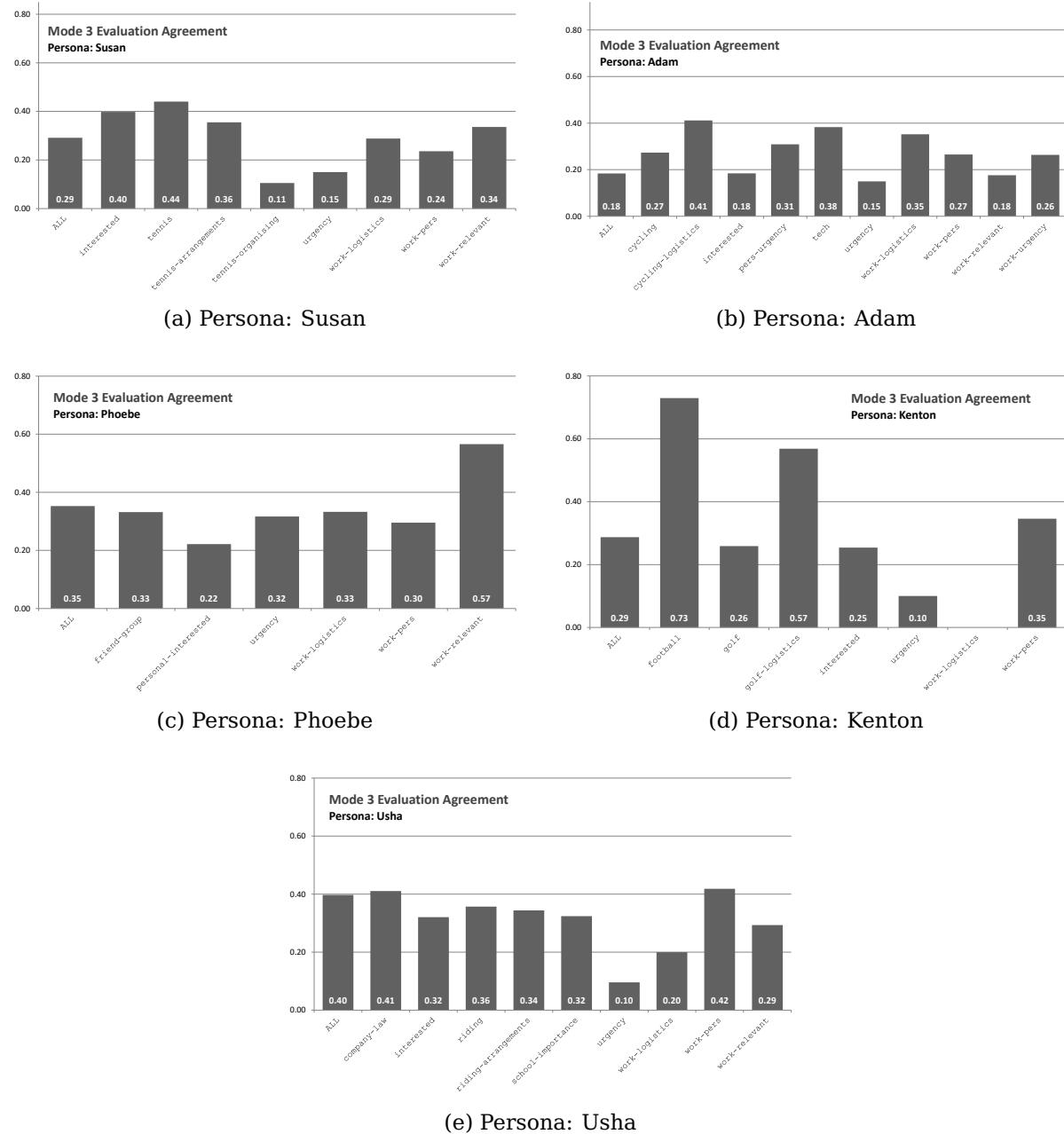


Figure S11.11: Mode 3 Synthetic Evaluation vs Participant Classification  $r_\phi$  by UD-ML Model

## S11 Study Detailed Results

---

### S11.1.5 Evaluation Difference

Table S11.7 shows aggregated data on the differences between numerical Likert values for synthetic evaluations and the corresponding participant feedback.

Figures S11.12 and S11.13 show this data broken down by evaluation output tag and UD-ML model respectively.

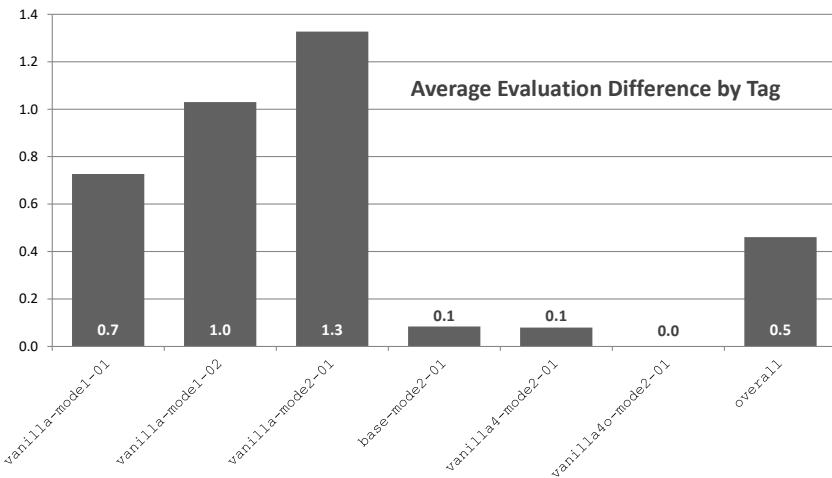


Figure S11.12: Average Evaluation Difference by Evaluation Tag

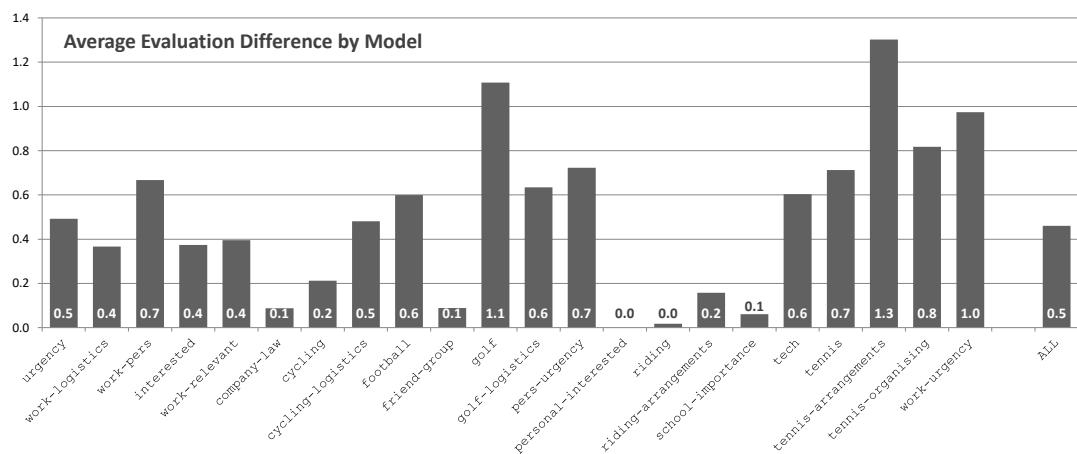


Figure S11.13: Average Evaluation Difference by UD-ML Model

## S11 Study Detailed Results

### Average and Spread for Evaluation Difference

This table shows the aggregated numerical differences between synthetic evaluations and the corresponding participant feedback actions, based on Likert value range of 1-5. A different close to zero indicates near perfect agreement on average, while a larger number represents a greater disagreement. The maximum value is 5, representing the difference between Likert values 1 & 5 or 5 & 1.

Model	vanilla-mode1-01				vanilla-mode1-02				vanilla-mode2-01				vanilla4-mode2-01				vanilla4-mode2-01				
	#	average	stddev	count	average	stddev	count	average	stddev	count	average	stddev	count	average	stddev	count	average	stddev	count	average	stddev
urgency	5	0.7	1.4	114	1.1	1.5	234	0.9	1.7	106	0.2	0.8	98	0.1	0.8	152	0.1	0.5	272	0.5	+0.0
work-logistics	5	0.5	1.5	114	0.5	1.6	232	1.3	1.7	104	0.4	1.3	96	0.1	0.8	153	0.0	0.5	272	0.4	-0.1
work-pers	5	0.7	1.9	114	0.9	1.9	232	1.3	1.9	105	0.9	1.9	88	0.4	1.1	153	0.4	0.9	271	0.7	+0.2
interested	4	0.4	1.5	114	0.5	1.8	233	0.9	1.6	106	0.5	1.8	97	0.2	1.1	133	0.1	0.8	247	0.4	-0.1
work-relevant	4	0.4	2.0	73	0.6	1.9	117	1.3	2.1	69	0.1	1.4	64	0.2	0.8	130	0.2	0.7	197	0.4	-0.1
company-law	1																			0.1	-0.4
cycling	1	0.1	1.2	30	0.1	0.8	29	0.4	1.3	37	0.3	1.3	36	0.2	0.9	42	0.1	1.2	22	0.2	-0.2
cycling-logistics	1	0.1	0.5	15	0.2	1.1	17	1.4	1.2	20	0.5	1.5	22	0.0	1.1	26	0.4	1.3	18	0.5	+0.0
football	1	0.7	1.3	41	1.0	1.8	115	1.5	1.4	37	0.0	1.0	33	0.0	0.0	23	0.0	0.7	75	0.6	+0.1
friend-group	1																		0.1	0.5	25
golf	1	2.1	1.3	42	1.9	1.6	121	2.1	1.0	37	0.2	1.1	34	0.1	0.6	23	0.0	0.9	75	1.1	-0.6
golf-logistics	1	0.9	1.7	41	1.0	1.7	116	1.7	1.2	37	0.0	0.9	34	0.0	0.2	23	0.0	0.9	75	0.6	+0.2
pers-urgency	1	0.5	1.2	30	1.3	1.9	29	1.5	1.6	37	0.4	1.1	36	0.1	1.0	41	0.2	0.7	22	0.7	+0.3
personal-interested	1																		0.0	0.4	25
riding	1																		0.1	0.5	43
riding-arrangements	1																		0.0	0.7	58
school-importance	1																		0.2	0.7	43
tech	1	0.2	1.1	30	1.2	1.8	29	1.7	1.6	37	0.0	0.6	36	0.1	0.3	41	0.1	0.2	22	0.6	-0.1
tennis	1	1.6	1.5	43	1.4	1.9	88	1.3	2.1	32	0.4	1.8	30	0.0	1.6	25	0.1	0.6	92	0.7	+0.3
tennis-arrangements	1	1.8	1.3	21	2.6	1.7	94	2.3	1.4	32	0.9	2.1	30	0.0	1.5	26	0.1	0.6	92	1.3	+0.8
tennis-organising	1	1.7	1.8	43	1.7	2.1	88	1.4	1.8	32	0.1	1.0	28	0.5	1.4	26	0.1	0.8	92	0.8	+0.4
work-urgency	1	0.8	1.3	30	2.0	1.7	30	2.2	1.6	37	0.3	1.1	37	0.2	0.6	41	0.1	0.4	22	1.0	+0.5
All	0.7	1.6	895	1.0	1.8	1.804	1.3	1.8	865	0.1	1.5	799	0.1	1.0	1,270	0.0	0.7	2,148	0.5	1.4	

Model label:

Table S11.7: Average and Spread for Evaluation Difference

## S11 Study Detailed Results

---

### S11.1.6 Evaluation Ratings

Table S11.8 shows overall ratings values for Synthetic Evaluations, identified by Output Tag<sup>1</sup>. These values are calculated by taking the Evaluation Feedback [S11.1.2], Mode 1/2 Evaluation Agreement [S11.1.3] and Mode 3 Evaluation Agreement [S11.1.4] as applicable to the mode. In the case of Modes 1 and 2, where we have two out of these available, the overall rating is a mean of the values. In the case of Mode 3, we use the single value that is available.

Note that these comparisons are not all like-for-like – while all of these source statistics measure correlation between synthetic evaluation results and some yardstick (either participant feedback or comparison with phase 2 classification actions), they are calculated in different ways ( $r$ ,  $r_{pb}$  and  $r_\phi$  respectively). However, they provide a relative comparison of the strengths of each tag.

This data is also summarised in Figure S11.14.

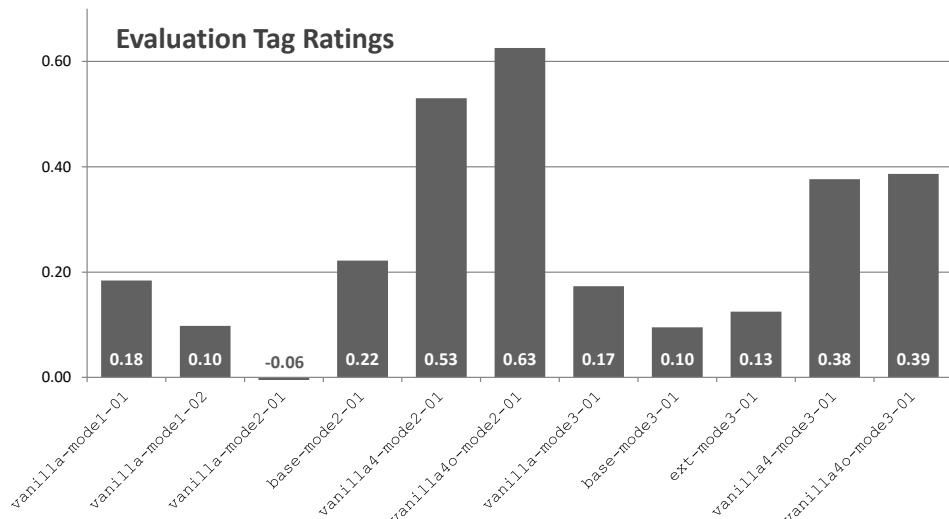


Figure S11.14: Synthetic Evaluation Tag Ratings

<sup>1</sup>Which represents a unique combination of OpenAI model, evaluation mode and iteration for the synthetic evaluation

**S11 Study Detailed Results****Evaluation Tag Ratings**

Rating values for models, taken from the mean of the 'ALL' correlation number across all personas using different techniques as applicable to the evaluation type. Where an ALL mean is not available, a mean of the individual items is used.

<b>Tag</b>	<b>PFEF</b>	<b>PFEA</b>	<b>PCFMA</b>	<b>Overall</b>
vanilla-mode1-01	0.24	0.13		0.18
vanilla-mode1-02	0.16	0.04		0.10
vanilla-mode2-01	0.01	-0.14		-0.06
base-mode2-01	0.32	0.12		0.22
vanilla4-mode2-01	0.70	0.36		0.53
vanilla4o-mode2-01	0.76	0.49		0.63
vanilla-mode3-01			0.17	0.17
base-mode3-01			0.10	0.10
ext-mode3-01			0.13	0.13
vanilla4-mode3-01			0.38	0.38
vanilla4o-mode3-01			0.39	0.39

PFEF: Pearson Correlation for Evaluation Feedback

PFEA: Point Biserial Correlation for Evaluation Agreement

PCFMA: Phi Coefficient for Mode3 Evaluation Agreement

Table S11.8: Synthetic Evaluation Tag Ratings

## S11 Study Detailed Results

---

### S11.1.7 Evaluation Results

Table S11.9 shows detailed results of Synthetic Evaluations of the performance of each UD-ML model. Rows represent the UD-ML models under evaluation, and columns represent the Output Tags of the evaluations. The evaluations are output as two metrics:

- **Evaluation Percentage**
  - For Mode 1&2 evaluations, this is defined as the percentage of evaluated items awarded Likert values Agree & Strongly Agree)
  - For Mode 3 evaluations, this is defined as the percentage of evaluated items that have an Agree result
- **Evaluation Score**
  - For Mode 1&2 evaluations, a score is assigned to items based on the Evaluation Likert value:
    - \* Strongly Disagree: 0
    - \* Disagree: 0.25
    - \* Neutral: 0.5
    - \* Agree: 0.75
    - \* Strongly Agree: 1
  - For Mode 3 evaluations, an award of Agree is assigned a score of 1 and Disagree is assigned 0
  - A mean is then calculated for these scores

Table S11.10 shows the Evaluation scores from Table S11.9 consolidated for UD-ML model via a simple mean and compared to the Manual Classification Agreement values taken from Tables S11.1 and S11.3. We have also added two metrics to measure similarity between the values, *Proximity Ratio* and *Squared Error x100 (SE<sub>100</sub>)*, defined as:

$$\text{Proximity Ratio} = \frac{\min(A, B)}{\max(A, B)}$$

$$\text{SE}_{100} = 100 \times (A - B)^2$$

Where *A* is the value for Percentage of Manual Classification Agreement [S11.1], and *B* is the Evaluation Score for the V4o-M2 column (corresponding to the vanilla4o-mode2-01 tag). This tag is selected as the prime Synthetic Evaluation as it has the highest Evaluation Rating as defined in Section S11.1.6.

Table S11.11 shows a subset of this information - containing only the agreement percentage, evaluation score and similarity metrics for clarity. This is shown in graphical form in Figure S11.15

## S11 Study Detailed Results

Evaluation Results															
Persona	Model	All Models												Best Model	
		vanilla-model-01			vanilla-model-02			vanilla-model-03			vanilla-model-04			agreement score	
		agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	agree %	score	Best Model	
Susan	All	59.9%	0.70	49.2%	0.57	35.6%	0.49	79.7%	0.33	76.1%	0.79	85.8%	0.87	82.2%	0.82
Susan	interested	39.9%	0.55	45.8%	0.55	32.1%	0.48	53.1%	0.59	48.2%	0.57	62.7%	0.67	64.6%	0.66
Susan	tennis	55.5%	0.67	35.5%	0.45	91.0%	0.32	81.7%	0.84	92.0%	0.93	33.4%	0.33	44.5%	0.45
Susan	tennis-arrangements	12.0%	0.49	1.8%	0.22	3.8%	0.23	58.6%	0.23	81.6%	0.45	96.5%	0.41	33.5%	0.33
Susan	teams-organising	45.3%	0.62	31.0%	0.41	35.1%	0.48	91.0%	0.32	82.7%	0.83	51.1%	0.51	84.0%	0.83
Susan	work-logistics	81.9%	0.74	51.0%	0.55	95.8%	0.36	87.1%	0.89	98.8%	0.96	74.0%	0.67	82.6%	0.67
Susan	work-relevant	68.0%	0.74	48.5%	0.59	96.7%	0.36	87.2%	0.89	92.0%	0.93	53.2%	0.53	70.0%	0.70
Susan	cycling-logistics	88.4%	0.74	66.6%	0.71	60.3%	0.47	66.3%	0.59	55.3%	0.59	54.4%	0.56	37.1%	0.46
Susan	work-relevant	72.8%	0.55	49.7%	0.52	83.6%	0.54	79.5%	0.62	43.7%	0.44	75.3%	0.75	69.3%	0.69
Adam	All	73.7%	0.79	57.0%	0.65	32.5%	0.51	84.6%	0.36	83.2%	0.85	89.4%	0.89	54.0%	0.54
Adam	cycling-logistics	94.8%	0.70	90.7%	0.32	70.6%	0.82	91.2%	0.31	92.8%	0.93	92.7%	0.74	70.6%	0.71
Adam	interested	91.8%	0.33	90.8%	0.31	48.5%	0.57	82.6%	0.33	86.7%	0.98	84.0%	0.76	77.2%	0.77
Adam	peps-urgency	68.6%	0.74	65.2%	0.71	39.5%	0.53	68.6%	0.66	63.3%	0.69	74.0%	0.76	40.6%	0.43
Adam	tech	84.7%	0.38	62.0%	0.58	38.1%	0.59	94.2%	0.34	90.2%	0.90	49.1%	0.76	76.6%	0.77
Adam	urgency	64.6%	0.74	41.9%	0.55	24.7%	0.47	95.1%	0.95	90.2%	0.92	56.4%	0.64	81.9%	0.76
Adam	work-logistics	88.1%	0.30	68.3%	0.75	33.8%	0.52	95.1%	0.95	91.5%	0.95	63.8%	0.64	66.8%	0.76
Adam	work-relevant	59.8%	0.59	56.0%	0.53	25.6%	0.38	62.6%	0.56	64.1%	0.71	82.1%	0.82	33.1%	0.33
Adam	work-urgency	68.6%	0.76	46.2%	0.57	26.0%	0.45	87.6%	0.86	81.4%	0.84	86.5%	0.87	48.9%	0.75
Adam	work-relevant	63.6%	0.71	28.3%	0.44	7.4%	0.33	85.7%	0.97	88.7%	0.90	92.0%	0.92	29.1%	0.29
Phoebe	All	76.8%	0.80	84.3%	0.84	76.8%	0.80	82.2%	0.84	84.8%	0.87	81.5%	0.82	81.9%	0.82
Phoebe	friend-group	57.6%	0.67	49.7%	0.50	29.3%	0.53	79.7%	0.62	84.3%	0.86	60.3%	0.60	69.8%	0.70
Phoebe	personal-interested	73.7%	0.79	59.2%	0.65	43.7%	0.61	92.6%	0.93	88.4%	0.89	60.5%	0.60	60.8%	0.60
Phoebe	urgency	20.9%	0.46	15.8%	0.36	31.7%	0.43	82.7%	0.69	83.6%	0.79	46.8%	0.74	79.4%	0.74
Phoebe	work-logistics	74.9%	0.31	61.7%	0.38	30.6%	0.56	95.7%	0.96	96.9%	0.97	97.4%	0.97	96.2%	0.96
Phoebe	work-pepers	54.8%	0.62	49.2%	0.59	16.3%	0.44	53.2%	0.58	55.5%	0.66	62.5%	0.68	54.2%	0.62
Phoebe	work-relevant	84.6%	0.63	45.1%	0.57	31.3%	0.59	92.7%	0.34	79.6%	0.83	94.3%	0.91	86.5%	0.91
Phoebe	work-urgency	46.8%	0.56	71.3%	0.76	45.3%	0.56	98.3%	0.38	87.0%	0.98	82.5%	0.86	91.6%	0.91
Kenton	All	85.7%	0.87	87.0%	0.87	89.4%	0.89	88.4%	0.97	82.3%	0.86	84.4%	0.84	87.4%	0.87
Kenton	football	93.6%	0.94	93.6%	0.94	92.6%	0.93	92.6%	0.97	97.8%	0.98	90.9%	0.92	92.6%	0.92
Kenton	golf	70.4%	0.46	72.7%	0.74	72.4%	0.72	97.4%	0.98	97.4%	0.98	77.9%	0.78	97.4%	0.78
Kenton	golf-logistics	74.9%	0.31	61.7%	0.38	92.4%	0.92	96.5%	0.91	97.1%	0.91	96.2%	0.91	97.4%	0.91
Kenton	interested	54.8%	0.62	49.2%	0.59	16.3%	0.44	53.2%	0.58	55.5%	0.66	62.5%	0.68	51.8%	0.58
Kenton	urgency	54.4%	0.63	45.1%	0.57	31.3%	0.59	92.7%	0.34	79.6%	0.83	94.3%	0.91	84.5%	0.91
Kenton	work-logistics	84.6%	0.65	71.3%	0.76	45.3%	0.56	98.3%	0.38	87.0%	0.98	82.5%	0.86	93.5%	0.91
Kenton	work-pepers	46.8%	0.56	47.3%	0.55	26.4%	0.40	46.7%	0.53	66.5%	0.71	28.5%	0.28	37.2%	0.38
Kenton	work-relevant	94.4%	0.95	93.5%	0.94	94.4%	0.95	93.5%	0.94	94.4%	0.95	94.4%	0.95	95.0%	0.95
Kenton	work-urgency	88.4%	0.90	95.5%	0.93	98.5%	0.98	98.5%	0.98	98.5%	0.98	98.5%	0.98	98.5%	0.98
Kenton	work-logistics	88.8%	0.90	90.5%	0.89	90.5%	0.90	90.5%	0.90	90.5%	0.90	90.5%	0.90	90.5%	0.90
Kenton	work-relevant	90.4%	0.91	93.5%	0.93	93.5%	0.94	93.5%	0.94	93.5%	0.94	93.5%	0.94	93.5%	0.94
Mean	All	65.0%	0.73	53.4%	0.62	32.9%	0.51	81.5%	0.84	81.2%	0.83	87.0%	0.87	52.2%	0.52
Mean	V-M1-A	V-M1-B	V-M2	B-M2	V-M3	V-M4	V-M5	V-M6	V-M7	V-M8	V-M9	V-M10	V-M11	All	

Table S11.9: Evaluation Results Detail

## S11 Study Detailed Results

### Classification Manual Agreement vs Evaluation Score

This table shows the Classification Manual Agreement value of each UP-ML model alongside the Evaluation Score awarded by each of the Synthetic Evaluations. The figure in square brackets below the evaluation tag label is the rating for that tag, a composite coefficient that represents the quality of that evaluation tag; the closer this rating is to 1, the higher the quality. The final columns are a proximity ratio and squared error value, comparing the closeness of the best evaluation (V40-M2) with the classification agreement.

<b>Model</b>	<b>Manual</b>		<b>V-M1-A</b>	<b>V-M1-B</b>	<b>V-M2</b>	<b>B-M2</b>	<b>V4-M2</b>	<b>V40-M2</b>	<b>V-M3</b>	<b>B-M3</b>	<b>E-M3</b>	<b>V4-M3</b>	<b>V40-M3</b>	<b>Proximity Ratio</b>	<b>Squared Error</b>
	#	kappa	% agree	[0..18]	[0..10]	[-0..06]	[0..22]	[0..53]	[0..63]	[0..17]	[0..10]	[0..13]	[0..38]	[0..39]	
ALL	0.83	92.8%	0.72	0.61	0.51	0.84	0.83	0.87	0.52	0.63	0.68	0.83	0.83	ALL	0.94
urgency	0.71	97.0%	0.72	0.62	0.57	0.95	0.88	0.93	0.73	0.85	0.81	0.89	0.94	u-r	0.96
work-logistics	0.64	96.7%	0.88	0.76	0.59	0.96	0.88	0.90	0.64	0.73	0.77	0.87	0.86	w-l	0.93
work-peers	0.75	91.0%	0.66	0.63	0.42	0.62	0.70	0.73	0.32	0.39	0.34	0.71	0.63	w-p	0.81
interested	0.71	85.6%	0.88	0.76	0.55	0.92	0.84	0.92	0.44	0.75	0.69	0.82	0.87	i-n	0.93
work-relevant	0.48	90.0%	0.82	0.67	0.50	0.87	0.85	0.89	0.47	0.65	0.72	0.84	0.86	w-r	0.99
company-law	0.61	94.1%	0.00	0.00	0.00	0.94	0.97	0.00	0.00	0.92	0.93	0.93	0.97	c-o_1	0.09
cycling	0.79	92.4%	0.95	0.92	0.82	0.91	0.93	0.93	0.74	0.71	0.75	0.93	0.91	c-y	0.99
cycling-logistics	0.56	93.0%	0.93	0.91	0.67	0.93	0.88	0.89	0.84	0.77	0.85	0.78	0.89	c-y-l	0.96
football	0.79	90.1%	0.79	0.65	0.61	0.86	0.93	0.89	0.60	0.61	0.83	0.91	0.89	f-o	0.01
friend-group	0.74	89.0%	0.00	0.00	0.00	0.84	0.87	0.00	0.00	0.82	0.84	0.89	0.98	f-g	0.04
golf	0.87	98.6%	0.46	0.38	0.43	0.85	0.96	0.98	0.47	0.79	0.90	0.96	0.97	g-o	0.99
golf-logistics	0.55	95.7%	0.81	0.68	0.56	0.96	0.97	0.98	0.74	0.91	0.91	0.96	0.97	g-l	0.98
pers-urgency	0.75	94.7%	0.69	0.55	0.44	0.93	0.87	0.90	0.49	0.77	0.81	0.89	0.87	p-u	0.95
personal-interested	0.70	87.8%	0.00	0.00	0.00	0.69	0.79	0.00	0.00	0.77	0.82	0.84	0.98	p-i	0.77
riding	0.73	91.8%	0.00	0.00	0.00	0.92	0.91	0.00	0.00	0.90	0.91	0.99	0.01	r-i	0.99
riding-arrangements	0.64	97.4%	0.00	0.00	0.00	0.89	0.96	0.00	0.00	0.88	0.95	0.95	0.99	r-a	0.02
school-importance	0.43	90.4%	0.00	0.00	0.00	0.95	0.94	0.00	0.00	0.97	0.95	s-i	0.96	0.13	
tech	0.77	95.9%	0.88	0.68	0.59	0.94	0.90	0.94	0.74	0.83	0.89	0.88	0.94	t-c	0.98
tennis	0.48	88.5%	0.67	0.45	0.47	0.92	0.84	0.93	0.33	0.35	0.33	0.84	0.88	t-n	0.95
tennis-arrangements	0.51	97.4%	0.49	0.22	0.23	0.65	0.85	0.96	0.04	0.33	0.50	0.84	0.93	t-a	0.02
tennis-organising	0.31	96.6%	0.62	0.41	0.48	0.92	0.83	0.96	0.38	0.51	0.51	0.83	0.91	t-o	0.99
work-urgency	0.64	94.0%	0.71	0.44	0.33	0.87	0.90	0.92	0.29	0.66	0.86	0.91	0.90	w-u	0.00
<b>Proximity Ratio</b>															0.04
<b>Squared Error</b>															0.98

Calculated as  $\min(A,B)/\max(A,B)$ , where A is manual classification percentage agreement and B is V40-M2 evaluation score

Calculated as  $100 \times (A-B)^2 / 2$ , where A is manual classification percentage agreement and B is V40-M2 evaluation score - note that this is multiplied by 100 for readability

Table S11.10: Classification Agreement vs Evaluation Score for All Evaluations

## S11 Study Detailed Results

---

### Classification Manual Agreement vs Evaluation Score

This table shows the Classification Manual Agreement value of each UD-ML model alongside the V4o-M2 Synthetic Evaluation Score. A proximity ratio and squared error are calculated to show the closeness of the synthetic to manual classifications.

Model	#	Manual Agreement	Evaluation Score	Proximity Ratio	Squared Error
ALL		0.93	0.87	0.94	0.33
urgency	5	0.97	0.93	0.96	0.16
work-logistics	5	0.97	0.90	0.93	0.42
work-pers	5	0.91	0.73	0.81	3.09
interested	4	0.86	0.92	0.93	0.41
work-relevant	4	0.90	0.89	0.99	0.01
company-law	1	0.94	0.97	0.97	0.09
cycling	1	0.92	0.93	0.99	0.00
cycling-logistics	1	0.93	0.89	0.96	0.16
football	1	0.90	0.89	0.99	0.01
friend-group	1	0.89	0.87	0.98	0.04
golf	1	0.99	0.98	0.99	0.00
golf-logistics	1	0.96	0.98	0.98	0.05
pers-urgency	1	0.95	0.90	0.95	0.22
personal-interested	1	0.88	0.79	0.90	0.77
riding	1	0.92	0.91	0.99	0.01
riding-arrangements	1	0.97	0.96	0.99	0.02
school-importance	1	0.90	0.94	0.96	0.13
tech	1	0.96	0.94	0.98	0.04
tennis	1	0.88	0.93	0.95	0.21
tennis-arrangements	1	0.97	0.96	0.99	0.02
tennis-organising	1	0.97	0.96	0.99	0.00
work-urgency	1	0.94	0.92	0.98	0.04

Table S11.11: Classification Agreement vs Evaluation Score for Prime Evaluation

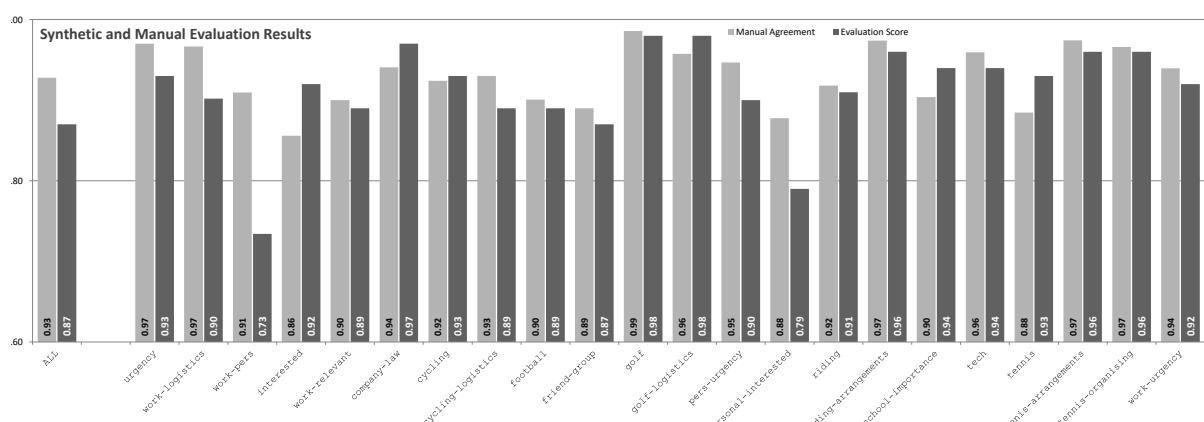


Figure S11.15: Classification Agreement vs Evaluation Score for Prime Evaluation

## S11.2 Selection of Evaluation Items

This section contains a selection of individual synthetic evaluation item summaries, organised by Persona. Each item is formatted to show the type of item, channel that it was sent to, output tag, unique Item UUID and a summary of the item text<sup>2</sup>). Each entry shows what the classification (UD-ML model) was, the classification selected by UD-ML, the synthetic evaluation, and the participant-entered value. It also includes the evaluation text generated by the synthetic evaluator.

### S11.2.1 Susan

#### S11.2.1.1 Susan Example 1 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>family</i>
phase2, phase2-vanilla-mode1-01 008c9eda-86bc-44f8-9a47-a36138577bbe	If anyone's looking for fresh eggs, our hens have been particularly productive! Let me know if you w...
Classification: <b>interested</b>	<i>As Susan, I strongly disagree with the classification 'not' for this item. The message is related to my hobby of tennis, and I am interested in it from a personal perspective. The correct classification should have been 'personal'.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Agree</b>	

#### S11.2.1.3 Susan Example 2 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>tennis-chat</i>
phase2-vanilla-mode1-01 02f38c70-c331-48bb-8b09-3b25942ae1c8	<b>Definitely, Mason. It's the camaraderie off the court that makes us stronger on it.</b>
Classification: <b>tennis</b>	<i>I completely agree with the classification chosen by the system as the item appears to be related to the user's hobby/interest of tennis. It is a casual conversation related to tennis and seems to be specifically about the user's interest in tennis.</i>
Categorised as: <b>tennis</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

<sup>2</sup>Which is also a hyperlink to the item – in the case of Slack messages, this link will no longer be accessible, and we cannot guarantee that any other link type will remain valid

## S11 Study Detailed Results

---

### S11.2.1.5 Susan Example 3 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>work-team-general</i>
phase2-vanilla-mode1-02 07f09430-3e9a-4a50-acef-1e66e448f288	<b>Sophie, can you put me down as a maybe for Thursday? Waiting to hear back about a dentist appointmen...</b>
Classification: <b>work-logistics</b>	<i>The system correctly marked this as 'not' for work-logistics. The message is not related to work logistics, but rather about a personal dental appointment. My likert value is 5 as I completely agree with the system's classification.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Disagree</b>	

### S11.2.1.7 Susan Example 4 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>susan-neil</i>
phase2-vanilla-mode1-02 165eec01-81fe-405e-9717-72df9a89b290	<b>Good idea. Let's sit down this weekend and brainstorm some ideas.</b>
Classification: <b>work-relevant</b>	<i>This message is not work-related and doesn't seem relevant to my work tasks at the university. The system's classification of 'not' is appropriate in this case.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.1.8 Susan Example 5 - vanilla - mode2

<b>SLACK MESSAGE</b>	<i>family</i>
phase2-vanilla-mode2-01 008c9eda-86bc-44f8-9a47-a36138577bbe	<b>If anyone's looking for fresh eggs, our hens have been particularly productive! Let me know if you w...</b>
Classification: <b>work-relevant</b>	<i>The system classified this as 'not' work-relevant, which is correct as the content is about fresh eggs. However, the context of this message may be misinterpreted, so I am neutral about this classification.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.1.9 Susan Example 6 - vanilla - mode2

**SLACK MESSAGE**

phase2-vanilla-mode2-01  
031e5f68-da66-44c0-b056-abb01dd6be14

*family*

@neil, can you check if we're out of bird seed? Noticed the feeder's looking a bit empty and the rob...

Classification: **work-logistics**

Categorised as: **not**

Synthetic Evaluation: **Agree**

Participant Evaluation: **Agree**

*This message does not seem to be related to work logistics. It's about checking on bird seed and the robins, which is more personal or social. I mostly disagree with the 'work logistics' classification and would choose 'not' instead. Therefore, I have given an evaluationLikert value of 4.*

### S11.2.1.10 Susan Example 7 - vanilla4 - mode2

**SLACK MESSAGE**

phase2-vanilla4-mode2-01  
0294b201-4cb2-4f2b-9182-9808cc83a9a4

*tennis-chat*

Pressure's on then! But we've got this, team. Let's show them what we're made of!  
:muscle::tennis:

Classification: **interested**

Categorised as: **personal**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*Given my big interest in tennis, classifying this under 'personal' because it's about my hobby is perfectly aligned. Therefore, I strongly agree with this classification.*

### S11.2.1.11 Susan Example 8 - vanilla4 - mode2

**SLACK MESSAGE**

phase2-vanilla4-mode2-01  
0294b201-4cb2-4f2b-9182-9808cc83a9a4

*tennis-chat*

Pressure's on then! But we've got this, team. Let's show them what we're made of!  
:muscle::tennis:

Classification: **tennis-arrangements**

Categorised as: **not**

Synthetic Evaluation: **Disagree**

Participant Evaluation: **Disagree**

*This message seems focused on team motivation and encouragement related to a tennis match, which leans more towards 'arrangements' rather than not being relevant at all. Therefore, I mostly disagree with the 'not' classification.*

#### S11.2.1.12 Susan Example 9 - vanilla4 - mode2

<b>RSS ITEM</b>	UK news / The Guardian
phase2-vanilla4-mode2-01 1d495126-cb4e-4a82-a7aa-2e438b504cc1	<b>Grocery price rises in Great Britain slow as cost of toilet rolls, butter and milk falls</b>
Classification: <b>work-relevant</b>	<i>I strongly agree with the classification 'not' for relevance to my work tasks. The content is focused on grocery market trends, which is unrelated to my administrative duties at the university.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

#### S11.2.1.13 Susan Example 10 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	work-team-general
phase2-vanilla4o-mode2-01 01df244a-719c-4c51-a086-4b2dd2b7407f	<b>Eleanor, thanks for the heads up! Will do. And Alisha, I thought it was just me! Glad to know I'm no...</b>
Classification: <b>interested</b>	<i>I mostly agree with the classification of 'not'. There isn't explicit content in the message that relates to any of my personal or work interests, but as a social and personable individual who values communication, Susan may have some minor interest in the social interaction.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Agree</b>	
Participant Evaluation: <b>Agree</b>	

#### S11.2.1.14 Susan Example 11 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	work-team-general
phase2-vanilla4o-mode2-01 01df244a-719c-4c51-a086-4b2dd2b7407f	<b>Eleanor, thanks for the heads up! Will do. And Alisha, I thought it was just me! Glad to know I'm no...</b>
Classification: <b>work-pers</b>	<i>I mostly disagree with this classification of 'work' as the message does not contain anything that seems related to work. It appears to be social and casual in tone, which would align better with 'personal' given the available classifications.</i>
Categorised as: <b>work</b>	
Synthetic Evaluation: <b>Disagree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.1.15 Susan Example 12 - vanilla4o - mode2

<b>RSS ITEM</b>	<i>BBC Sport</i>
phase2-vanilla4o-mode2-01	<b>Swiatek cruises into Madrid Open quarter-finals</b>
10516468-01a6-494d-a060-2f2fd1f45b31	
Classification: <b>tennis-arrangements</b>	<i>I strongly agree with the classification of 'not' for this item. While it pertains to tennis, it is a news update about a professional event and has no connection to arranging friendly matches or social plans.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.1.16 Susan Example 13 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>work-university-general</i>
phase2-vanilla4o-mode2-01	<b>Can't stress enough how helpful the new campus map app has been. Finding buildings and rooms has nev...</b>
16700bb7-ff82-4f68-8359-72b030514e7b	
Classification: <b>work-pers</b>	<i>The item is focused on a tool relevant to the campus and its students/staff, aligning it with the work context. I strongly agree with the classification as 'work'.</i>
Categorised as: <b>work</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.1.17 Susan Example 14 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>susan-neil</i>
phase2-vanilla4o-mode2-01	<b>Great! That's one less thing to worry about. Thanks for sorting it, love.</b>
1eb769b5-7896-4ff4-aa85-d1845267ae31	
Classification: <b>urgency</b>	<i>I mostly agree with this being classified as 'not urgent'. However, the response suggests something has been resolved, which might have been urgent previously. While it doesn't directly invoke urgency, there's a small chance that context might change interpretation. Therefore, I give a score of 4.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Agree</b>	
Participant Evaluation: <b>Agree</b>	

## S11.2.2 Adam

### S11.2.2.1 Adam Example 1 - vanilla4o - mode2

**SLACK MESSAGE** bwcc-general-chat  
phase2-vanilla4o-mode2-01 **Count me in too! I need to get some miles in.**  
03c33a30-f166-41a5-ac63-d52531cc1714

Classification: **cycling**

Categorised as: **not**

Synthetic Evaluation: **Strongly Disagree**

Participant Evaluation: **Strongly Disagree**

*The content directly mentions getting miles in, which is clearly related to cycling. As a keen cyclist, I find this relevant to my hobby. I strongly disagree with the classification of 'not'—it should have been classified as 'cycling'.*

### S11.2.2.2 Adam Example 2 - vanilla - mode1

**SLACK MESSAGE** bis-general  
phase2-vanilla-mode1-02 **That's great news, Henry! Congratulations to the team.**  
0b408103-02f4-450b-bc0e-f07da1e5b0ec

Classification: **work-pers**

Categorised as: **work**

Synthetic Evaluation: **Neutral**

Participant Evaluation: **Strongly Agree**

*The system selected 'work' for work-pers, indicating that the item relates to the work domain. However, based on the message content, it is difficult to ascertain if the item specifically pertains to work, personal life, or both. Therefore, a 'unsure' assessment is more fitting as the classification lacks clarity based on the available information.*

### S11.2.2.3 Adam Example 3 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>bwcc-zwift-chat</i>
phase2-vanilla-mode1-02	<b>Can't wait for tonight. Been practicing my sprint finish on Zwift all week.</b>
1479a470-9847-4b0a-878c-2bb59f5183b3	
Classification: <b>cycling-logistics</b>	<i>I strongly disagree with the classification 'not' for cycling-logistics. This item is not about logistics/arrangements in the cycle club, but rather about personal cycling practice on Zwift.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Strongly Disagree</b>	

### S11.2.2.4 Adam Example 4 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>bwcc-zwift-chat</i>
phase2-vanilla-mode1-02	<b>Can't wait for tonight. Been practicing my sprint finish on Zwift all week.</b>
1479a470-9847-4b0a-878c-2bb59f5183b3	
Classification: <b>tech</b>	<i>I strongly disagree with the classification 'not' for tech. This item is indeed related to technology as it mentions using an online platform for cycling practice.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.2.5 Adam Example 5 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>bis-general</i>
phase2-vanilla-mode1-02	<b>Good morning everyone, just a quick update. We've received positive feedback from Oceanic Research C...</b>
15179db6-8352-4e8a-a929-110ef240b374	
Classification: <b>pers-urgency</b>	<i>The system has incorrectly classified this message as 'not' urgent, while the positive feedback from Oceanic Research Corp. should likely be considered urgent. Additionally, the positive feedback affects work, making it both work-related and urgent.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

#### S11.2.2.6 Adam Example 6 - vanilla - mode1

**SLACK MESSAGE**

phase2-vanilla-mode1-02  
15179db6-8352-4e8a-a929-110ef240b374

*bis-general*

**Good morning everyone, just a quick update.  
We've received positive feedback from Oceanic  
Research C...**

Classification: **urgency**

Categorised as: **not**

Synthetic Evaluation: **Strongly Disagree**

Participant Evaluation: **Agree**

*The system has incorrectly classified this message as 'not' urgent. The positive feedback should likely be considered urgent due to its work-related nature. The urgency of the content is not accurately reflected in the classification.*

#### S11.2.2.7 Adam Example 7 - vanilla - mode1

**SLACK MESSAGE**

phase2-vanilla-mode1-02  
15a95c41-5884-412c-81e2-96bad8c31264

*ian-chat*

**Hey, do you fancy going to The Bull tonight?  
Fallon's got a new ale on.**

Classification: **work-urgency**

Categorised as: **not**

Synthetic Evaluation: **Neutral**

Participant Evaluation: **Strongly Agree**

*I would have expected this message to be classified as - 'not' for 'work-urgency' because it is not connected with work or of an urgent nature. - 'not' for 'tech' because it is not related to technology. - 'not' for 'cycling-logistics' because it is not related to logistic-s/arrangements in my cycle club. - 'urgent' for 'pers-urgency' because it is a personal message and is urgent. - 'not' for 'work-relevant' because it is not work-related and relevant to my work tasks. - 'not' for 'urgency' because it is not in relation to something that I think is urgent. - 'not' for 'work-logistics' because it is not related to work logistics. - 'not' for 'cycling' because it is not related to my hobby of cycling. - 'personal' for 'interested' because it is related to my personal life and I am interested. - 'personal' for 'work-pers' because it relates to my personal life.*

## S11 Study Detailed Results

---

### S11.2.2.8 Adam Example 8 - vanilla4o - mode2

<b>RSS ITEM</b>	<i>UK news / The Guardian</i>
phase2-vanilla4o-mode2-01	<b>Humza Yousaf vows to stay on as Scottish first minister</b>
005f0b39-1fcf-4a53-96e8-4dd7bfd7f3c0	
Classification: <b>cycling</b>	<i>I strongly agree with the classification of 'not' for cycling. This political news content from Scotland has no connection to my cycling hobby. An evaluationLikert value of 5 reflects my strong agreement.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.2.9 Adam Example 9 - vanilla - mode2

<b>SLACK MESSAGE</b>	<i>family-group-chat</i>
phase2-vanilla-mode2-01	<b>I can help with the laptop, Granny. Let's set up a time this weekend.</b>
8120c964-870c-4708-b837-6d77cc87bed6	
Classification: <b>tech</b>	<i>This message is offering to help with a laptop, so it is definitely related to technology.</i>
Categorised as: <b>not</b>	<i>The classification of 'not' is not appropriate, and the better classification should be 'technology'. I strongly disagree with the current classification.</i>
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Neutral</b>	

### S11.2.2.10 Adam Example 10 - vanilla - mode2

<b>SLACK MESSAGE</b>	<i>bwcc-general-chat</i>
phase2-vanilla-mode2-01	<b>Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan</b>
c16cb886-36fd-4f9c-89f8-d10c98ead4b6	
Classification: <b>interested</b>	<i>Although the message is of personal interest, it does not specifically mention whether it relates to work. Therefore, I neither agree nor disagree with the classification of 'personal' for interested.</i>
Categorised as: <b>personal</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Agree</b>	

## S11 Study Detailed Results

---

### S11.2.2.11 Adam Example 11 - vanilla - mode2

<b>SLACK MESSAGE</b>	<i>bwcc-general-chat</i>
phase2-vanilla-mode2-01	<b>Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan</b>
c16cb886-36fd-4f9c-89f8-d10c98ead4b6	
Classification: <b>work-urgency</b>	<i>The Sunday ride cancellation message is not work-related and there is no indication of urgency. Therefore, I neither agree nor disagree with the classification of 'not' for work-urgency.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.2.12 Adam Example 12 - vanilla4 - mode2

<b>SLACK MESSAGE</b>	<i>bwcc-general-chat</i>
phase2-vanilla4-mode2-01	<b>Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan</b>
c16cb886-36fd-4f9c-89f8-d10c98ead4b6	
Classification: <b>work-urgency</b>	<i>The notification about missing a cycling event is clearly neither urgent nor work-related. The system's classification of 'not urgent' is accurate, reflecting the appropriate urgency level for this particular item.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.2.13 Adam Example 13 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>bwcc-general-chat</i>
phase2-vanilla4o-mode2-01	<b>Just a heads up, I won't be able to make the Sunday ride. Have a good one! - Ethan</b>
c16cb886-36fd-4f9c-89f8-d10c98ead4b6	
Classification: <b>work-urgency</b>	<i>I strongly agree with the selected classification of 'not' for this item in the context of work urgency. The content relates to a personal cycling commitment and does not pertain to work matters, urgent or otherwise. This classification is correct, so I have rated it a 5.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

#### S11.2.2.14 Adam Example 14 - base - mode2

**SLACK MESSAGE**

phase2-base-mode2-01

00f42aa0-d1bb-4e62-83e8-f1cd2bbde41c

*bis-team-general*

**Just a reminder that we have a team meeting with the CEO tomorrow. Make sure you're prepared with an...**

Classification: **pers-urgency**Categorised as: **not**Synthetic Evaluation: **Strongly Agree**Participant Evaluation: **Strongly Agree**

*This item is a work task and is also moderately urgent (as it is in relation to a team meeting with the CEO tomorrow), so I mostly agree with the selected classification of 'not' as this is not personal in nature. I have given an evaluationLikert value of 5 to indicate strong agreement.*

#### S11.2.2.15 Adam Example 15 - base - mode2

**SLACK MESSAGE**

phase2-base-mode2-01

aa335f2b-212c-4f2a-9b67-d2d9dc45da1e

*bis-general*

**Thanks for the reminder, Liam. Looking forward to the updates.**

Classification: **interested**Categorised as: **work**Synthetic Evaluation: **Strongly Disagree**Participant Evaluation: **Strongly Agree**

*I strongly disagree with the selected classification of 'work' for this item - I'm not interested in this message. The interest level also seems neutral, and the content is not really about my work duties. Given this, I would have instead chosen 'not'. Because I strongly disagree I have awarded an evaluationLikert value of 1.*

#### S11.2.2.16 Adam Example 16 - vanilla4 - mode2

**SLACK MESSAGE**

phase2-vanilla4-mode2-01

223d77aa-6942-432f-9880-a46b83a6f2cf

*bis-team-manager*

**Got it, Adam. I'll review and get back to you if I need any clarification.**

Classification: **work-pers**Categorised as: **personal**Synthetic Evaluation: **Strongly Disagree**Participant Evaluation: **Strongly Disagree**

*The content of the message is clearly work-related, concerning a professional task. Classifying it as 'personal' is incorrect; it should be classified as 'work'.*

**S11 Study Detailed Results****S11.2.3 Phoebe****S11.2.3.1 Phoebe Example 1 - vanilla4 - mode2**

<b>RSS ITEM</b>	UK news / The Guardian
phase2-vanilla4-mode2-01 02d6faf8-22d9-4be8-b6df-13a42ed07ee8	<b>?All eyes are on Glasgow?: UK's first legal drug consumption room ready to open</b>
Classification: <b>friend-group</b>	The content on Glasgow's drug consumption room doesn't relate to any specific group of friends nor is it a typical topic to discuss with my university or work-related friends. I strongly agree with its classification as 'not' relevant to a friend group, scoring it a 5 on the likert scale.
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Agree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

**S11.2.3.2 Phoebe Example 2 - vanilla4 - mode2**

<b>RSS ITEM</b>	UK news / The Guardian
phase2-vanilla4-mode2-01 02d6faf8-22d9-4be8-b6df-13a42ed07ee8	<b>?All eyes are on Glasgow?: UK's first legal drug consumption room ready to open</b>
Classification: <b>personal-interested</b>	Although the topic of the drug consumption room is compelling, it doesn't align directly with my immediate interests, especially considering my busy work life. Hence, I'm neutral about the classification of 'not interested' for personal relevance, giving it a likert score of 3.
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Neutral</b>	

**S11.2.3.3 Phoebe Example 3 - vanilla4o - mode2**

<b>SLACK MESSAGE</b>	work-fasttrack
phase2-vanilla4o-mode2-01 014cfebb-7ede-40ba-8097-556f4e14a19f	<b>Does anyone know if we're allowed to swap rotations if both parties agree? Asking for a friend...</b>
Classification: <b>work-logistics</b>	The question is about the logistics of swapping rotations, which makes it very much related to work logistics. I mostly disagree with the classification of 'not' since the item's content directly involves logistical planning in a work context.
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Disagree</b>	
Participant Evaluation: <b>Agree</b>	

#### S11.2.3.4 Phoebe Example 4 - vanilla4o - mode2

**SLACK MESSAGE** work-fasttrack  
phase2-vanilla4o-mode2-01 **Phoebe, I heard you're organising a Fast Track team-building event. Let me know if you need help wit...**  
046fa6af-cafe-41ff-95ca-79c76ea28f9a

Classification: **work-logistics**

Categorised as: **not**

Synthetic Evaluation: **Disagree**

Participant Evaluation: **Disagree**

*I mostly disagree with the classification of 'not' under 'work-logistics'. Organising a team-building event involves planning and logistics, which makes it relevant under this category. I have given a Likert value of 2 indicating disagreement.*

#### S11.2.3.5 Phoebe Example 5 - vanilla4o - mode2

**SLACK MESSAGE** work-team  
phase2-vanilla4o-mode2-01 **Megan, do we have a deadline for submitting the Beacon Industries compliance report?**  
088c4a1b-050c-4e84-bc65-19a78e588676

Classification: **urgency**

Categorised as: **not**

Synthetic Evaluation: **Agree**

Participant Evaluation: **Agree**

*I mostly agree with the classification of 'not' urgent for this item. While knowing the deadline could be an urgent matter depending on the timeframe, the content does not explicitly express urgency, making 'not urgent' a mostly appropriate classification.*

#### S11.2.3.6 Phoebe Example 6 - vanilla4 - mode2

**SLACK MESSAGE** friends-work  
phase2-vanilla4-mode2-01 **Phoebe, your brownies are legendary. Can we get a repeat performance next week?**  
35dfde69-518a-42de-b784-94d97fd356bd

Classification: **friend-group**

Categorised as: **work**

Synthetic Evaluation: **Agree**

Participant Evaluation: **Agree**

*This message is likely from a work colleague, suggesting it relates to my 'work' friend group. I mostly agree with this classification, though it borders on a personal activity done in a work context.*

## S11.2.4 Kenton

### S11.2.4.1 Kenton Example 1 - vanilla - mode1

<b>RSS ITEM</b>	<i>Football / The Guardian</i>
phase2-vanilla-mode1-01	<b>Virgil van Dijk unsure if he will be part of post-Jürgen Klopp era at Liverpool</b>
0d1dc70e-4309-4f1f-bed1-e20092a0c883	
Classification: <b>golf</b>	<i>As a client manager for a global consumer products company, I am not interested in this football-related news from a golf perspective.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Strongly Disagree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.4.2 Kenton Example 2 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>work-team-manager</i>
phase2-vanilla-mode1-01	<b>Great. Keep up the good work, Kenton!</b>
0349e52c-4a95-404a-a744-c3adbcf3f6b9	
Classification: <b>football</b>	<i>Since the message is not related to my interest in football, the 'not' classification for football is appropriate. It's just a casual message of encouragement with no football-related content.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.4.3 Kenton Example 3 - vanilla - mode1

<b>SLACK MESSAGE</b>	<i>work-team-general</i>
phase2-vanilla-mode1-01	<b>I'm in as well. It's been a while since we all caught up.</b>
0802a886-d429-450a-b938-db3b6adcd45d	
Classification: <b>urgency</b>	<i>I believe this item has been correctly classified as 'not' urgent, as the message content doesn't indicate any urgent nature.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Strongly Agree</b>	

#### S11.2.4.4 Kenton Example 4 - vanilla - mode1

**SLACK MESSAGE**

phase2-vanilla-mode1-01  
0af3e7fa-c20b-4774-973c-9ed510264866

*work-team-general*

**Hey team, just a heads up that I'll be in a meeting with Frescos this afternoon. If anyone needs me,...**

Classification: **golf-logistics**

Categorised as: **not**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*I strongly agree with the classification of 'not' for the item's relevance to logistics in the context of my golfing arrangements. The content of the message does not suggest any logistical arrangements related to golf.*

#### S11.2.4.5 Kenton Example 5 - vanilla - mode1

**SLACK MESSAGE**

phase2-vanilla-mode1-01  
0af3e7fa-c20b-4774-973c-9ed510264866

*work-team-general*

**Hey team, just a heads up that I'll be in a meeting with Frescos this afternoon. If anyone needs me,...**

Classification: **work-pers**

Categorised as: **personal**

Synthetic Evaluation: **Agree**

Participant Evaluation: **Disagree**

*I mostly agree with the classification of 'personal' for work/personal relevance. The message indicates availability after 4pm, implying a personal appointment.*

#### S11.2.4.6 Kenton Example 6 - vanilla - mode1

**RSS ITEM**

phase2-vanilla-mode1-01  
0f37afda-aca8-494c-a244-6ab57bf18e51

*UK news / The Guardian*

**Post Office inquiry has heard 'chorus of cowards', lawyers say**

Classification: **golf**

Categorised as: **not**

Synthetic Evaluation: **Strongly Disagree**

Participant Evaluation: **Strongly Agree**

*This item is not related to my hobby or interest in golf, it is focused on the Post Office scandal inquiry.*

## S11 Study Detailed Results

---

### S11.2.4.7 Kenton Example 7 - vanilla - mode2

<b>RSS ITEM</b> phase2-vanilla-mode2-01 006f844f-556d-42c4-adee-d5d10f4e9b0c	<i>UK news / The Guardian</i> <b>'Unacceptable greenwashing': Scottish farmed salmon should not be labelled organic, say charities</b>
Classification: <b>football</b>	<i>The content is not related to the persona's hobby/interest of football, so 'not' seems suitable. I mostly agree with the selected classification of 'not' for this item, therefore I have given an evaluationLikert value of 2 to indicate agreement.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Disagree</b>	
Participant Evaluation: <b>Strongly Agree</b>	

### S11.2.4.8 Kenton Example 8 - vanilla - mode2

<b>SLACK MESSAGE</b> phase2-vanilla-mode2-01 0af3e7fa-c20b-4774-973c-9ed510264866	<i>work-team-general</i> <b>Hey team, just a heads up that I'll be in a meeting with Frescos this afternoon. If anyone needs me,...</b>
Classification: <b>football</b>	<i>The message does not relate directly to football, so 'not' is a reasonable classification. However, the mention of being available after 4pm might indirectly relate to potential football activities later in the day. Therefore, it's not entirely clear-cut, but overall, 'not' seems more appropriate. I have awarded an evaluationLikert value of 3 because the classification is not entirely clear-cut.</i>
Categorised as: <b>not</b>	
Synthetic Evaluation: <b>Neutral</b>	
Participant Evaluation: <b>Strongly Agree</b>	

#### S11.2.4.9 Kenton Example 9 - base - mode2

##### RSS ITEM

phase2-base-mode2-01

04494042-1ff7-4a92-8d0c-2eec9981c3ac

The back page of the internet.

**Iran 0 - [1] Japan - Hidemasa Morita 28'**

Classification: **urgency**

Categorised as: **not**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*I strongly agree with the selected classification of 'not' for this item - I think this item is about a football match so is unlikely to be urgent. I have given an evaluationLikert value of 5 to indicate strong agreement.*

#### S11.2.4.10 Kenton Example 10 - base - mode2

##### SLACK MESSAGE

phase2-base-mode2-01

204a8a9a-784f-4edb-8b7f-b7026f8f121b

work-company-general

**Agreed, Ethan. It's always good to stay informed about what our competitors are up to. Let's continu...**

Classification: **work-pers**

Categorised as: **personal**

Synthetic Evaluation: **Strongly Disagree**

Participant Evaluation: **Strongly Disagree**

*I do mostly agree with the selected classification of the content 'not' related to my work or personal life. There's not much about this that screams either 'work' or 'personal' specifically, but on balance is probably slightly closer personal, so I have awarded an evaluationLikert value of 1 to indicate slight disagreement.*

#### S11.2.4.11 Kenton Example 11 - vanilla4 - mode2

##### SLACK MESSAGE

phase2-vanilla4-mode2-01

00033b65-f1a9-42b7-89e0-fd99c9365773

golf

**Great! Let's meet at HLGC at 10am on Saturday. And Benjamin, leave your terrible jokes at home pleas...**

Classification: **urgency**

Categorised as: **not**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*Arranging a casual meeting for golf does not present any urgency. I strongly agree with the classification of 'not urgent'.*

## S11 Study Detailed Results

---

### S11.2.4.12 Kenton Example 12 - vanilla4 - mode2

<b>RSS ITEM</b>	UK news / The Guardian
phase2-vanilla4-mode2-01 006f844f-556d-42c4-adee-d5d10f4e9b0c	<b>'Unacceptable greenwashing': Scottish farmed salmon should not be labelled organic, say charities</b>

Classification: **work-pers**

Categorised as: **personal**

Synthetic Evaluation: **Neutral**

Participant Evaluation: **Agree**

*While the article could potentially bear relevance to my professional world through the lens of corporate responsibility and public affairs, the direct connection to my personal interest in current and impactful news makes me lean more towards the article being a personal interest. However, I'm not fully convinced, hence a neutral Likert score of 3.*

### S11.2.4.13 Kenton Example 13 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>golf</i>
phase2-vanilla4o-mode2-01 00033b65-f1a9-42b7-89e0-fd99c9365773	<b>Great! Let's meet at HLGC at 10am on Saturday. And Benjamin, leave your terrible jokes at home pleas...</b>

Classification: **football**

Categorised as: **not**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*I strongly agree with the classification of 'not' for football. The content clearly pertains to a golf arrangement, with no indication or relevance to Kenton's interest in football. The system has made the correct distinction here.*

### S11.2.4.14 Kenton Example 14 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>football</i>
phase2-vanilla4o-mode2-01 0775c96f-17ab-4cbe-8515-2323e33ea6d2	<b>I'm in! I've heard RRFC's new striker is a real talent. Can't wait to see him in action.</b>

Classification: **football**

Categorised as: **football**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*I strongly agree with the classification of 'football'. The content explicitly discusses RRFC's new striker and an eagerness to see him play, directly aligning with the interest of football. A Likert value of 5 accurately reflects this strong alignment.*

## S11 Study Detailed Results

---

### S11.2.4.15 Kenton Example 15 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>football</i>
phase2-vanilla4o-mode2-01	I'm in! I've heard RRFC's new striker is a real talent. Can't wait to see him in action.
0775c96f-17ab-4cbe-8515-2323e33ea6d2	
Classification:	<b>urgency</b>
Categorised as:	<b>not</b>
Synthetic Evaluation:	<b>Strongly Agree</b>
Participant Evaluation:	<b>Strongly Agree</b>
	<i>I strongly agree with the classification of 'not'. The content is casual and about football, with no indication of urgency in either a work or personal context. The evaluationLikert is 5 to indicate strong agreement.</i>

### S11.2.4.16 Kenton Example 16 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>family</i>
phase2-vanilla4o-mode2-01	I'll join you. I could do with a break from work.
110aea4a-d389-4bd6-bb2d-aa8dee1f8a5d	
Classification:	<b>interested</b>
Categorised as:	<b>not</b>
Synthetic Evaluation:	<b>Neutral</b>
Participant Evaluation:	<b>Neutral</b>
	<i>I am neutral about the classification of 'not' for interest. The content does not provide sufficient information to determine if the sender considers this of personal or work interest. While it implies a casual response, it remains unclear.</i>

### S11.2.4.17 Kenton Example 17 - vanilla4o - mode2

<b>SLACK MESSAGE</b>	<i>work-team-client</i>
phase2-vanilla4o-mode2-01	Sure, Mila. I'll send him an invite. Also, I think it would be a good idea to include some customer ...
1630efb8-312c-4802-84af-e748b51bdfdb	
Classification:	<b>work-logistics</b>
Categorised as:	<b>logistics</b>
Synthetic Evaluation:	<b>Strongly Agree</b>
Participant Evaluation:	<b>Strongly Agree</b>
	<i>I strongly agree with the selected classification of 'logistics' for this item. The discussion about sending invites and including testimonials in a presentation clearly relates to logistical planning in a work context. It is an accurate classification.</i>

### S11.2.5 Usha

#### S11.2.5.1 Usha Example 1 - vanilla4 - mode2

##### SLACK MESSAGE

phase2-vanilla4-mode2-01

19721721-27da-4257-b827-24315a7e9ba7

*work-partners*

**Has anyone else noticed that some of our older clients seem less engaged lately? Maybe we should thi...**

Classification: **company-law**

Categorised as: **not**

Synthetic Evaluation: **Neutral**

Participant Evaluation: **Agree**

*While the content is work-related, it specifically discusses client engagement strategies rather than company law. Therefore, it's not directly pertaining to UK company law, but considering it broadly involves my professional environment, I'm neutral on whether this should be classified under 'company-law' or 'not'.*

#### S11.2.5.2 Usha Example 2 - vanilla4 - mode2

##### SLACK MESSAGE

phase2-vanilla4-mode2-01

23f84034-2a13-481f-a3a6-22918c60eab2

*family-close*

**Dad, did you see the new cricket bat Ravi got? It's practically glowing!**

Classification: **work-pers**

Categorised as: **personal**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*The content pertaining to my child's cricket bat clearly falls under personal life, not work. The classification as 'personal' is therefore correct.*

#### S11.2.5.3 Usha Example 3 - vanilla4o - mode2

##### SLACK MESSAGE

phase2-vanilla4o-mode2-01

3be30bc0-dfaa-4118-884d-d430186050ad

*work-clients*

**Usha, I've noticed some discrepancies in the compliance report. Could we go over them before submitt...**

Classification: **work-pers**

Categorised as: **work**

Synthetic Evaluation: **Strongly Agree**

Participant Evaluation: **Strongly Agree**

*I strongly agree with the selected classification of 'work'. Given the subject matter is a compliance report, it is undoubtedly related to my professional life and not personal.*

## S11 Study Detailed Results

---

### S11.2.5.4 Usha Example 4 - vanilla4o - mode2

**SLACK MESSAGE** family-wide  
phase2-vanilla4o-mode2-01 Yes, Usha! It's been dreadful. I had to  
3c348d77-b5ac-4155-90ee-a5f88f20b0f7 detour through Lakey Green last week and still  
barely made i...

Classification: **interested**  
Categorised as: **personal**  
Synthetic Evaluation: **Agree**  
Participant Evaluation: **Strongly Agree**

*I mostly agree with the classification of 'personal' for this item. The content appears to be from a personal conversation discussing traffic delays seemingly relatable to Usha's interests outside work. I rated this a Likert value of 4 as it aligns reasonably well but with mild uncertainty on strong personal relevance.*

### S11.2.5.5 Usha Example 5 - vanilla4o - mode2

**SLACK MESSAGE** school-parents-chat  
phase2-vanilla4o-mode2-01 'More reliable than the Wi-Fi,' Lucy? True,  
3f58eaba-028b-432e-91fb-11cd10a6edaa but I don't fancy cleaning up after them!  
Let's stick to...

Classification: **school-importance**  
Categorised as: **school-other**  
Synthetic Evaluation: **Neutral**  
Participant Evaluation: **Strongly Agree**

*The classification of 'school-other' could be correct if this content pertains to some context involving children's school, but the message itself is quite ambiguous without additional details. As such, I neither strongly agree nor disagree, and assigning an evaluation score of 3 seems appropriate to reflect the uncertainty.*

### S11.2.5.6 Usha Example 6 - vanilla4o - mode2

**RSS ITEM** UK news / The Guardian  
phase2-vanilla4o-mode2-01 Rachel Reeves must handle bond sell-off with  
3fe10da9-42af-46d7-a388-8a514d9297dd care, but this is not a Truss-level event

Classification: **interested**  
Categorised as: **not**  
Synthetic Evaluation: **Neutral**  
Participant Evaluation: **Agree**

*I am neutral about the classification of 'not' for this item. While it is not directly personal or work-related, the broad economic implications are somewhat interesting, though not necessarily actionable from my perspective.*

## S11 Study Detailed Results

### S11.3 Evaluation Token Usage

Figure S11.12 summarises the OpenAI token usage by persona, evaluation mode and model during the study.

OpenAI pricing<sup>3</sup> used different prices for different models, with input (prompt) tokens having a different cost to output (completion) tokens. Pricing is listed in USD per million tokens.

Data refers to the OpenAI models: gpt-3.5-turbo-1106 and gpt-4-turbo. Where "(FT)" appears in the model name (i.e. for -base and -ext models), the model is a fine-tuned model based on gpt-3.5-turbo-1106. Token usage for the process of fine-tuning models is not listed.

Agent	Tag	Model	Mode	Token Cost/M		Items Processed	Tokens			Tokens/Item			Cost		
				P	C		Prompt	Completion	Total	Prompt	Completion	Total	1	1000	ALL
adam	phase2-vanilla-mode1-01	gpt-3.5-turbo	1	\$1	\$2	234	518,819	151,265	670,084	2,217	646	2,864	0.0035	\$3.51	\$0.82
adam	phase2-vanilla-mode1-02	gpt-3.5-turbo	1	\$1	\$2	517	1,953,236	354,259	2,307,495	3,778	685	4,463	0.0051	\$5.15	\$2.66
adam	phase2-vanilla-mode2-01	gpt-3.5-turbo	2	\$1	\$2	231	1,051,534	230,749	1,282,283	4,552	999	5,551	0.0065	\$6.55	\$1.51
adam	phase2-vanilla-mode3-01	gpt-3.5-turbo	3	\$1	\$2	390	1,663,872	343,622	2,007,494	4,266	881	5,147	0.0060	\$6.03	\$2.35
adam	phase2-base-mode2-01	gpt-3.5-turbo (FT)	2	\$3	\$6	109	507,668	101,545	609,213	4,658	932	5,589	0.0196	\$19.56	\$2.13
adam	phase2-base-mode3-01	gpt-3.5-turbo (FT)	3	\$3	\$6	383	1,655,263	342,743	1,998,006	4,322	895	5,217	0.0183	\$18.33	\$7.02
adam	phase2-ext-mode3-01	gpt-3.5-turbo (FT)	3	\$3	\$6	349	1,518,425	215,559	1,733,984	4,351	618	4,968	0.0168	\$16.76	\$5.85
adam	phase2-vanilla4-mode2-01	gpt-4-turbo	2	\$10	\$30	291	1,341,900	251,664	1,593,564	4,611	865	5,476	0.0721	\$72.06	\$20.97
adam	phase2-vanilla4-mode3-01	gpt-4-turbo	3	\$10	\$30	317	1,356,422	265,797	1,622,219	4,279	838	5,117	0.0679	\$67.94	\$21.54
adam	phase2-vanilla40-mode2-01	gpt-4o-2024-11-20	2	\$2.50	\$10	387	1,763,360	357,413	2,120,773	4,556	924	5,480	0.0206	\$20.63	\$7.98
adam	phase2-vanilla40-mode3-01	gpt-4o-2024-11-20	3	\$2.50	\$10	536	2,267,217	441,157	2,708,374	4,230	823	5,053	0.0188	\$18.81	\$10.08
kenton	phase2-vanilla-mode1-01	gpt-3.5-turbo	1	\$1	\$2	451	1,012,696	236,988	1,249,684	2,245	525	2,771	0.0033	\$3.30	\$1.49
kenton	phase2-vanilla-mode1-02	gpt-3.5-turbo	1	\$1	\$2	458	1,744,297	252,682	1,996,979	3,809	552	4,360	0.0049	\$4.91	\$2.25
kenton	phase2-vanilla-mode2-01	gpt-3.5-turbo	2	\$1	\$2	448	2,006,944	334,207	2,341,151	4,480	746	5,226	0.0060	\$5.97	\$2.68
kenton	phase2-vanilla-mode3-01	gpt-3.5-turbo	3	\$1	\$2	846	3,531,854	544,814	4,076,668	4,175	644	4,819	0.0055	\$5.46	\$4.62
kenton	phase2-base-mode2-01	gpt-3.5-turbo (FT)	2	\$3	\$6	304	1,320,699	215,205	1,535,904	4,344	708	5,052	0.0173	\$17.28	\$5.25
kenton	phase2-base-mode3-01	gpt-3.5-turbo (FT)	3	\$3	\$6	1009	4,177,058	658,709	4,835,767	4,140	653	4,793	0.0163	\$16.34	\$16.48
kenton	phase2-ext-mode3-01	gpt-3.5-turbo (FT)	3	\$3	\$6	959	3,976,276	513,531	4,489,807	4,146	535	4,682	0.0157	\$15.65	\$15.01
kenton	phase2-vanilla4-mode2-01	gpt-4-turbo	2	\$10	\$30	665	3,029,514	451,190	3,480,704	4,556	678	5,234	0.0659	\$65.91	\$43.83
kenton	phase2-vanilla4-mode3-01	gpt-4-turbo	3	\$10	\$30	1318	5,600,389	874,223	6,474,612	4,249	663	4,912	0.0624	\$62.39	\$82.23
kenton	phase2-vanilla40-mode2-01	gpt-4o-2024-11-20	2	\$2.50	\$10	849	3,762,689	616,776	4,379,465	4,432	726	5,158	0.0183	\$18.34	\$15.57
kenton	phase2-vanilla40-mode3-01	gpt-4o-2024-11-20	3	\$2.50	\$10	1162	4,831,972	732,661	5,564,633	4,158	631	4,789	0.0167	\$16.70	\$19.41
susan	phase2-vanilla-mode1-01	gpt-3.5-turbo	1	\$1	\$2	367	818,699	217,156	1,035,855	2,231	592	2,822	0.0034	\$3.41	\$1.25
susan	phase2-vanilla-mode1-02	gpt-3.5-turbo	1	\$1	\$2	465	1,763,503	311,639	2,075,142	3,792	670	4,463	0.0051	\$5.13	\$2.39
susan	phase2-vanilla-mode2-01	gpt-3.5-turbo	2	\$1	\$2	388	1,733,698	342,816	2,076,514	4,468	884	5,352	0.0062	\$6.24	\$2.42
susan	phase2-vanilla-mode3-01	gpt-3.5-turbo	3	\$1	\$2	731	3,024,532	561,877	3,586,409	4,138	769	4,906	0.0057	\$5.67	\$4.15
susan	phase2-base-mode2-01	gpt-3.5-turbo (FT)	2	\$3	\$6	179	787,194	146,701	933,895	4,398	820	5,217	0.0181	\$18.11	\$3.24
susan	phase2-base-mode3-01	gpt-3.5-turbo (FT)	3	\$3	\$6	496	2,057,125	389,809	2,446,934	4,147	786	4,933	0.0172	\$17.16	\$8.51
susan	phase2-ext-mode3-01	gpt-3.5-turbo (FT)	3	\$3	\$6	465	1,928,707	307,136	2,235,843	4,148	661	4,808	0.0164	\$16.41	\$7.63
susan	phase2-vanilla4-mode2-01	gpt-4-turbo	2	\$10	\$30	203	893,114	153,897	1,047,011	4,400	758	5,158	0.0667	\$66.74	\$13.55
susan	phase2-vanilla4-mode3-01	gpt-4-turbo	3	\$10	\$30	237	984,352	168,404	1,152,756	4,153	711	4,864	0.0629	\$62.85	\$14.90
susan	phase2-vanilla40-mode2-01	gpt-4o-2024-11-20	2	\$2.50	\$10	339	1,504,733	276,520	1,781,253	4,439	816	5,254	0.0193	\$19.25	\$6.53
susan	phase2-vanilla40-mode3-01	gpt-4o-2024-11-20	3	\$2.50	\$10	514	2,123,878	371,929	2,495,807	4,132	724	4,856	0.0176	\$17.57	\$9.03
phoebe	phase2-vanilla4-mode2-01	gpt-4-turbo	2	\$10	\$30	338	1,428,572	202,597	1,631,169	4,227	599	4,826	0.0602	\$60.25	\$20.36
phoebe	phase2-vanilla4-mode3-01	gpt-4-turbo	3	\$10	\$30	475	1,890,393	263,831	2,154,224	3,980	555	4,535	0.0565	\$56.46	\$26.82
phoebe	phase2-vanilla40-mode2-01	gpt-4o-2024-11-20	2	\$2.50	\$10	331	1,400,673	212,808	1,613,481	4,232	643	4,875	0.0170	\$17.01	\$5.63
phoebe	phase2-vanilla40-mode3-01	gpt-4o-2024-11-20	3	\$2.50	\$10	450	2,768,037	254,736	2,022,773	3,929	566	4,495	0.0155	\$15.48	\$6.97
usha	phase2-vanilla4-mode2-01	gpt-4-turbo	2	\$10	\$30	278	1,194,154	238,208	1,432,362	4,296	857	5,152	0.0687	\$68.66	\$19.09
usha	phase2-vanilla4-mode3-01	gpt-4-turbo	3	\$10	\$30	295	1,175,693	235,724	1,411,417	3,985	799	4,784	0.0638	\$63.83	\$18.83
usha	phase2-vanilla40-mode2-01	gpt-4o-2024-11-20	2	\$2.50	\$10	200	860,896	180,573	1,041,469	4,304	903	5,207	0.0198	\$19.79	\$3.96
usha	phase2-vanilla40-mode3-01	gpt-4o-2024-11-20	3	\$2.50	\$10	322	1,284,484	254,598	1,539,082	3,989	791	4,780	0.0179	\$17.88	\$5.76
														Total cost: \$365.33	
ALL	All Mode 1	ALL	1			1,052	2,350,214	605,409	2,955,623	2,234	575	2,810			
ALL	All Mode 2	ALL	2			5,540	24,587,342	4,312,869	28,900,211	4,438	778	5,217			
ALL	All Mode 3	ALL	3			11,254	46,815,949	7,740,860	54,556,809	4,160	688	4,848			

Table S11.12: OpenAI token Usage in Evaluations by Model and Mode

<sup>3</sup>OpenAI pricing data was current as of 27 March 2024:

<https://web.archive.org/web/20240327174934/https://openai.com/pricing>

## **Supplement S12**

**Source Code & Additional Materials**

# Source Code & Additional Materials

## Contents

- S12.1 – Source Code
  - S12.1.1 – AwAg Services
  - S12.1.2 – AwAg UI
  - S12.1.3 – Simple Persistent Object Store
- S12.2 – L<sup>A</sup>T<sub>E</sub>X Template

## S12.1 Source Code

This section contains details of source code for some software related to this research.

### S12.1.1 AwAg Services

The awagdata and awagml services are implemented in a combined Python Flask application<sup>1</sup>, **AwAg Services**. This is available at the following addresses:

- doi:10.21954/ou.rd.28303898
- <https://github.com/revisionist/ou-phd-code-awag-svc>

The application structure is based on Flask boilerplate from Idris Rampurawala<sup>2</sup>. Significant files are:

- **awag\_evaluation\_processor.py** – Used to run synthetic evaluations
- **awag\_stats\_engine.py** – Generates statistics from the stored research data; it contains the class StatsEngine, which is extended by the classes StatsEngineExtended and StatsEngineExtendedDataFrames – which generate statistics by classification and generate JSON/Excel stats packs respectively.
- **dataset\_manager.py** – Manages datasets and fine-tuning content on OpenAI
- **simulation\_manager.py** – Generates and retrieves synthetic content
- **views.py** – Implements awagml routes

---

<sup>1</sup><https://flask.palletsprojects.com/> [<https://perma.cc/XC37-Q3ZT>]

<sup>2</sup><https://github.com/idris-rampurawala/flask-boilerplate>, open-sourced under the MIT License

## S12 Source Code & Additional Materials

---

Exports of some Postman<sup>3</sup> collections for interacting with these services are located at:  
[doi:10.21954/ou.rd.28044944](https://doi.org/10.21954/ou.rd.28044944) [path: /study/postman]

### S12.1.2 AwAg UI

The Training and Evaluation user interfaces for the study are provided by the **AwAg UI** application (awagUi). This is written in Angular<sup>4</sup> and is available at the following addresses:

- [doi:10.21954/ou.rd.28303901](https://doi.org/10.21954/ou.rd.28303901)
- <https://github.com/revisionist/ou-phd-code-awag-ui>

This UI uses AwAg Services as the back end.

### S12.1.3 Simple Persistent Object Store

We built a simple JSON object store that was focussed on our specific needs. This is used in this project for fine-tuning related data. We have made it freely available under an Apache 2.0 license at:

- <https://github.com/revisionist/python-apps/tree/main/flask/sposs>

## S12.2 LATEX

A LATEX template based on the files used to generate the thesis and supplement documents is available on GitHub at:

- <https://github.com/revisionist/latex/blob/master/examples/phd-thesis/>

No warranties express or implied ☺

---

<sup>3</sup><https://www.postman.com/product/rest-client/> [<https://perma.cc/8RQL-WAK3>]

<sup>4</sup><https://angular.io/>