

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\n 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.\n\n An AwAg Evaluation Request is presented as a structured JSON document, and its main elements are:\n\n 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\n 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.\n\n 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.\n\n 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.\n\n Use 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 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.\n\n 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.\n"
},
{
  "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\n 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.\n\n An AwAg Evaluation Request is presented as a structured JSON document, and its main elements are:\n 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\n 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.\n\n 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.\n\n 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.\n\n 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.\n\n 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.\n\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\n 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.\n\n ",
},
{
  "role": "user",
  "content": {
    "persona": null,
```

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