Report on sample size

```python
from matplotlib import pyplot as plt
import pandas as pd
import numpy as np
```

Models

Note that we will perform all of the below for both experts and non-experts separately (but using the same images) to estimate

RQ 1: Can AI mimic the styles of artists?

This model uses AI-real comparisons, and simply bundles all AI models together for analysis.

$$\mathbb{P}(A > B) = \text{sigmoid}\left(\beta_{AI} \cdot (x_A^{\text{is\_AI}} - x_B^{\text{is\_AI}})\right). \tag{1}$$

This model has 1 parameter to estimate.

We can also split out by image generation model, to see whether some AI models are more or less effective:

$$\mathbb{P}(A > B) = \text{sigmoid}\left(\beta \cdot (x_A^{\text{models}} - x_B^{\text{models}})\right) \tag{2}$$

This has M parameters to estimate, where M is the number of models. Note that this is taking into account dropping out the coefficient term for human-generated art.

We can also split out by artist, to see how AI performance varies across artists:

$$\mathbb{P}(A > B) = \text{sigmoid}\left(\beta \cdot (x_A^{\text{is\_AI, artists}} - x_B^{\text{is\_AI, artists}})\right) \tag{3}$$

This groups together all AI models, and compares their performance against human models for each artist. Since one artist is dropped, this model has $A-1$ parameters to estimate (where A is the number of artists).

Finally, we can split by both artist and model to estimate how AI performance varies across both artists and models:

$$\mathbb{P}(A > B) = \text{sigmoid}\left(\beta \cdot (x_A^{\text{models, artists}} - x_B^{\text{models, artists}})\right) \tag{4}$$

This has $(A-1) \cdot M$ parameters to estimate.

RQ 2: How does anachronistic subject matter influence model mimicry capability?

This uses AI-AI comparisons. Our primary analysis measures how AI models' relative performance varies with anachronistic subject matter. We have already performed the following analysis for RQ 1:

$$\mathbb{P}(A > B) = \text{sigmoid}\left(\beta \cdot (x_A^{\text{models}} - x_B^{\text{models}})\right) \tag{5}$$

Because no human-generated art is being included, we adjust this slightly, estimating $m-1$ parameters. This means that coefficients are with reference to one model. The choice is arbitrary; we can go with the best-performing model, for example.

We do this analysi across 3 different types of subject matter: in-sample subject matter (e.g., apples for Cezanne), out-of-sample subject matter (Eiffel tower for Cezanne), and anachronistic subject matter (laptop for Cezanne). This yields $3 \cdot (M-1)$ parameters, which may be compared to analyse how relative model performance varies across types of subject matter.

RQ 3: What aspects of style is AI better or worse at mimicing?

All of the models so far have been straightforward from the literature. For RQ3, however, we need to extend the standard models a little.

Consider the following:

$$\mathbb{P}(A > B) = \text{sigmoid}(\lambda_A - \lambda_B) \tag{6}$$

λ_A is the 'overall characteristicness' of image A. This is the Bradley-Terry model for pairwise comparison. Assuming that $\lambda_A = \beta \cdot x + \epsilon$ yields the models used in RQ1 and RQ2.

However, we break it down further for RQ3, assuming that λ_A is the result of an unobserved linear combination of S style aspects λ_A^i for $i = 1, ..., S$, weighted by ‘style aspect weighting vector’ γ:

$$\lambda_A = \sum_{i=1}^{S} \gamma^i \cdot \lambda_A^i \tag{7}$$

Further, we assume that:

$$\lambda_A^i = \theta^i \cdot x_A + \epsilon, \tag{8}$$

where x_A are some properties of image A, and θ^i are regression coefficients.

In plain English, the characteristicness of image A for a certain aspect of style i is determined by *something* along with random noise, which captures individual variation.

Survey participants choose a top reason out of S reasons for why they preferred image A over image B. We assume that the probability that they choose reason r_i is proportional to $\gamma^i \cdot (\lambda_A^i - \lambda_B^i)$. In plain English, it depends on a) how much the respondent cares about style aspect i, as well as b) the actual perceived difference between the two images along style aspect i. This yields the following model,

$$\mathbb{P}(r_i \mid A > B) = \frac{\exp(\gamma^i \cdot (\lambda_A^i - \lambda_B^i))}{\sum_{j=1}^{S} \exp(\gamma^j \cdot (\lambda_A^j - \lambda_B^j))}.$$

Now, we plug in (Equation 8) to give:

$$\mathbb{P}(r_i \mid A > B) = \frac{\exp(\gamma^i \cdot \theta^i \cdot (x_A - x_B))}{\sum_{j=1}^{S} \exp(\gamma^j \cdot \theta^j \cdot (x_A - x_B))}.$$

We woud like to estimate γ and θ. However, note from this model that they cannot be disentangled. For example, for any given values of $\theta and \gamma$ that we estimate, $\varepsilon \cdot \theta and \frac{1}{\varepsilon}\gamma$ provide an identical estimate. To eliminate this over-parameterisation, we set $\beta = \gamma \cdot \theta$, giving the following model:

$$\mathbb{P}(r_i \mid A > B) = \frac{\exp(\beta^i \cdot (x_A - x_B))}{\sum_{j=1}^{S} \exp(\beta^j \cdot (x_A - x_B))}.$$

This is a standard multinomial regression. For RQ3, we consider two models. Firstly, we group all AI models together:

$$\mathbb{P}(r_i \mid A > B) = \frac{\exp(\beta^i \cdot (x_A^{\text{is\_AI}} - x_B^{\text{is\_AI}}))}{\sum_{j=1}^{S} \exp(\beta^j \cdot (x_A^{\text{is\_AI}} - x_B^{\text{is\_AI}}))}. \tag{9}$$

This requires estimating S parameters via standard multinomial regression.

For the second model, we try to estimate how different AI models' performance varies across different aspects. This gives the following model:

$$\mathbb{P}(r_i \mid A > B) = \frac{\exp(\beta^i \cdot (x_A^{\text{models}} - x_B^{\text{models}}))}{\sum_{j=1}^{S} \exp(\beta^j \cdot (x_A^{\text{models}} - x_B^{\text{models}}))} \tag{10}$$

Again, we drop the term for human-generated images, so that each coefficient term measures the relative performance to human-generated images. This yields $S \cdot M$ parameters to estimate.

Summary of models

Research question	Model name	Equation	# of parameters
1	AI mimicry skill	Equation 1	1
1	Model mimicry skill	Equation 2	M
1	AI mimicry skill across artists	Equation 3	$A - 1$
1	Model mimicry skill across artists	Equation 4	$M \cdot (A - 1)$
2	Model mimicry skill across anachronistic content	Equation 5	$3 \cdot (M - 1)$
3	Reason for preference, AI vs human	Equation 9	S
3	Reason for preference, model comparison	Equation 10	$S \cdot M$

Rule of thumb analysis

There is existing literature for rules of thumb for how many participants are needed for discrete choice analyses. I draw on Assele, Meulders, and Vandebroek (2023) in particular, which gives the following equation (assuming 5% significance level and 90% desired statistical power),

$$N \geq 150 \cdot \frac{K}{S}, \text{ equivalently } N \cdot S \geq 150 \cdot K$$

where N is the number of participants, K is the number of parameters that must be estimated, and S is the number of choice tasks performed by each participant.

Applying this to our table of analyses above, consider first RQ1.

Model name	Equation	# of parameters	Required # of choices
AI mimicry skill	Equation 1	1	150
Model mimicry skill	Equation 2	M	$150M$
AI mimicry skill across artists	Equation 3	$A-1$	$150(A-1)$
Model mimicry skill across artists	Equation 4	$M \cdot (A-1)$	$150M(A-1)$

Each of the first 3 should be doable, respectively requiring 150, 450, and 450 choices if we assume 3 models and 4 artists being investigated. If we assume that each expert makes 24 choices, then we need 19 experts.

The most demanding case is the last one, and if we assume that we investigate 3 models and 4 artists, then we require $150 \cdot 3 \cdot 3 = 1350$ choices. Even assuming 30 choices per expert, this is 45 experts, which is beyond our 'safe bet' of 40. So maybe doable.

Considering RQ2, we have $3 \cdot (M-1)$ parameters, and therefore, assuming 3 models being tested, 900 choices required. If we have 30 choices per expert, this requires 30 experts. Note that there is **no overlap** with the RQ1 data, and so this would amount to 60 questions per expert - likely too many.

These are the studies for RQ3:

Model name	Equation	\# of parameters	Required # of choices
Reason for preference, AI vs human	Equation 9	S	$150 \cdot S$
Reason for preference, model comparison	Equation 10	$S \cdot M$	$150 \cdot S \cdot M$

Each of these can re-use data from RQ1, so long as users fill out the reason for each preference. If $S = 6$, then the required # of choices for the basic model is 900. Assuming 24 choices per expert, this requires 38 experts, which is realistic.

On the other hand, the second model, assuming $M = 3$, requires 2700 choices. This seems quite unfeasible, requiring 90 experts to answer 30 questions each, for example.

Choice set design

Let's assume the following:

- Only consider RQs 1 and 3 for now
- each user will do 24 choices at most
- 4 blocks of artists, so 6 questions per artist
- 3 AI models (so 2 per artist-AI combination)
- 2 subjects per artist.

This gives a nice set of 4 blocks (1 per artist). Each block has a factorial structure of 6 questions across 2 subject types and 3 AI models. Every user sees the same choice set.

```python
from src import simulate as sim

res = sim.generate_data_for_simulation()

print(res.keys())

res['outcomes']
```

```
dict_keys(['outcomes', 'models', 'artworks', 'artists', 'participants'])
```

	participant	real_artwork	ai_artwork
0	par~WVM2b8NK8xKYbqzU99rCzP	img~8KyHcPvaZ6SK4BqXeFuDNU	img~gSyD2htypcKegpc
1	par~WVM2b8NK8xKYbqzU99rCzP	img~8KyHcPvaZ6SK4BqXeFuDNU	img~HK8Rha4c37ogYz
2	par~WVM2b8NK8xKYbqzU99rCzP	img~8KyHcPvaZ6SK4BqXeFuDNU	img~oGTCVeoGptnoN
3	par~WVM2b8NK8xKYbqzU99rCzP	img~WqGHzM6qnkJFTpMzHqfgah	img~L3WWPJ7ifW5N
4	par~WVM2b8NK8xKYbqzU99rCzP	img~WqGHzM6qnkJFTpMzHqfgah	img~EUaxwMZnU8JEz
...	...	...	...
955	par~gJuVrwubNSNL7GiwKu9Gh5	img~6PHXmMBSKMbGv2T2DQCGPX	img~Lpe8HaERhrpM5
956	par~gJuVrwubNSNL7GiwKu9Gh5	img~6PHXmMBSKMbGv2T2DQCGPX	img~GYW5kLPkoPVB
957	par~gJuVrwubNSNL7GiwKu9Gh5	img~KofxTjnkqahgWXKUEwbFna	img~nhGNEpcqGoBfx
958	par~gJuVrwubNSNL7GiwKu9Gh5	img~KofxTjnkqahgWXKUEwbFna	img~WUqZWsGnsiN56
959	par~gJuVrwubNSNL7GiwKu9Gh5	img~KofxTjnkqahgWXKUEwbFna	img~N2B5gvbFDMMja

Calibration of settings

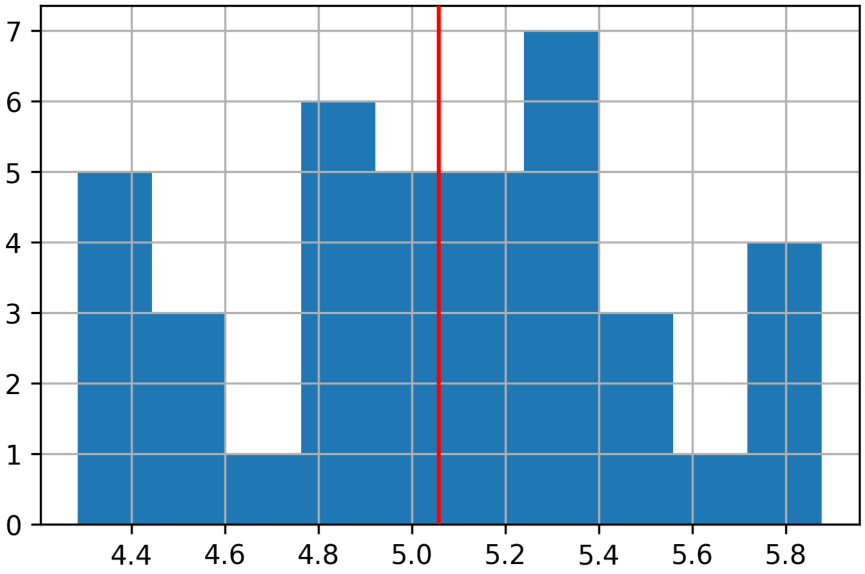
Let’s look at the participants in our simulation:

```
res['participants'].head()
```

	id	skill	style_aspect_weights
0	par~WVM2b8NK8xKYbqzU99rCzP	4.776783	[0.15125544813944436, 0.1825106648578071, 0.16...
1	par~NYGh6r4sLR9fS7JHgL3iR3	4.921638	[0.1826637568970099, 0.1595408483555649, 0.175...
2	par~cur2A3mArjXpCbuD5FZikB	4.481020	[0.17844955983854718, 0.16159939914742613, 0.1...
3	par~8rYw4bpu5FByJiG5j6E2xL	4.953546	[0.1778702756466676, 0.18193322942222329, 0.16...
4	par~htk3nuqibLjHWcVoBJSCqQ	4.896690	[0.1794390398877533, 0.16739439803892894, 0.15...

Note the ‘skill’ parameter, which should have a mean of roughly 5:

```
res['participants']['skill'].hist(zorder=1)
plt.axvline(res['participants']['skill'].mean(), color = 'red', zorder=10)
plt.show()
```



Why did I set this? Similar to when building the analytical models, the ‘skill’ of participants and the scale of images’ ‘characteristicness’ are coupled together. I want the characteristicness to roughly sit between 0 and 1 so I can model it with a beta distribution in the simulation.

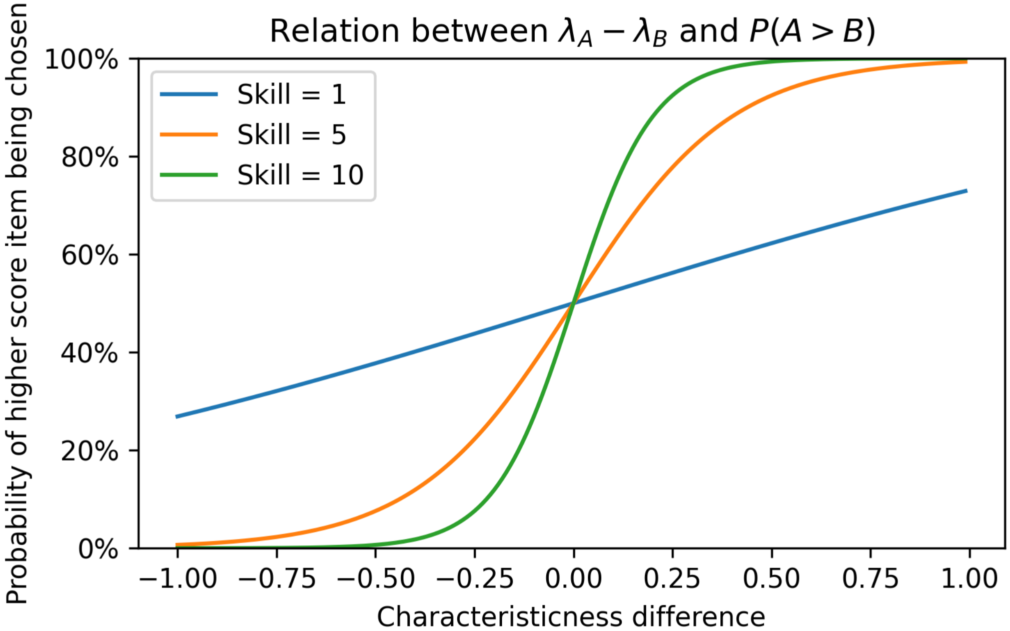
Hence, I choose a mean value for skill which produces sensible results when the difference in characteristicness between two images can range, at the extremes, between 0 and 1.

The following shows how difference between the characteristicness of two images relates to the probability of choosing each image, for differing participant skills.

```
from matplotlib import ticker

for skill in (1,5,10):
    x = np.arange(-1, 1, 0.01)
    y = 1 - sim.score_diff_to_probability(skill * x)
    plt.plot(x, y, label = f'Skill = {skill}')

plt.ylim((0,1))
plt.xlabel('Characteristicness difference')
plt.ylabel('Probability of higher score item being chosen')
plt.legend()
plt.gca().yaxis.set_major_formatter(ticker.PercentFormatter(1.0))
plt.title('Relation between $\\lambda_A - \\lambda_B$ and $P(A>B)$')
plt.tight_layout()
```



You can see that the skill of 5 gives a reasonable range, where a difference of 1 will basically guarantee that the better image is chosen. So we just fix the mean of users' skill to 5, and allow variance around that to model varying user skill.

But when choosing scenarios to investigate, we still need to know the consequence of differences in score. Assuming user skill of 5 (we won't vary this mean in analyses), the following characteristicness differences have the following effects:

```
probs_of_interest = [0.99, 0.9, 0.75, 0.66, 0.55]
x = np.arange(0, 1, 0.01)
y = 1 - sim.score_diff_to_probability(5 * x)
for p in probs_of_interest:
    idx = (np.abs(y - p)).argmin()
    print(p, x[idx])
```

```
0.99 0.92
0.9 0.44
0.75 0.22
0.66 0.13
0.55 0.04
```

Scenarios to investigate

```
res['artworks'].head()
```

	id	is_ai	artist	creator
0	img~8KyHcPvaZ6SK4BqXeFuDNU	False	art~eMxbEZezqGrbMytr58UmBi	art~eMxbEZezqGrbMy
1	img~WqGHzM6qnkJFTpMzHqfgah	False	art~eMxbEZezqGrbMytr58UmBi	art~eMxbEZezqGrbMy
2	img~2625TvUBTfiaHYpe26jkoZ	False	art~6fXGe69gXq3qajufHFJ99W	art~6fXGe69gXq3qajuf
3	img~Pg6BEB6kQy8SpzPBKYNDFC	False	art~6fXGe69gXq3qajufHFJ99W	art~6fXGe69gXq3qajuf
4	img~89jTdHWEJtnx2NrwzK9pwd	False	art~VXxggQhbYx6fQfq9um2UBG	art~VXxggQhbYx6fQf

Analyses

```
from src import data_prep
data = data_prep.process_simulation_data(res['outcomes'], res['artworks'])

print(data['model'].shape, res['outcomes'].shape)

# AI basic
```

```
(960, 3) (960, 11)
```

Assele, Samson Yaekob, Michel Meulders, and Martina Vandebroek. 2023. “Sample Size Selection for Discrete Choice Experiments Using Design Features.” *Journal of Choice Modelling* 49 (December): 100436. https://doi.org/10.1016/j.jocm.2023.100436.