## Term 2 Week 5 Progress

Evaluation:

* When doing research into how design ideas are evaluated, found this paper from 2015 offering a well-researched methodology for evaluating non-industry specific product ideas: <https://www.designsociety.org/publication/37913/A+MODEL+OF+IDEA+EVALUATION+AND+SELECTION+FOR+PRODUCT+INNOVATION>
  + A white rectangular box with black text

    Description automatically generated
  + A screenshot of a computer

    Description automatically generated
  + A white rectangular box with black text

    Description automatically generated
  + A close-up of a table

    Description automatically generated
  + A close-up of a table

    Description automatically generated
* This seems to me like the optimal way to have someone evaluate product ideas. I could create my own survey but it would not be likely to improve on this
  + Is it worth doing this anyway for the sake of the project being my own work?
* I also realised when doing this that this evaluation method is actually an optimal way for my software to choose the best product idea out of all the previous ones it generates (this was previously done with the AnalyseFeasibility and AnalyseMarketability actions)
  + I’ve rewritten that section of the program to get GPT to score the products using the above methodology instead.
* Having my software use the same evaluation methodology as I have the real-life evaluator use opens up a lot more room for interesting analysis
  + I’ve added the ability to bypass the product generation part of the program and instead allow a human to enter their own product ideas
  + This means I can now easily analyse the program in two sections

Current Evaluation Plan as-is:

* What will be analysed:
  + Idea generation section/Overall performance:
    - 1) Run MetaGPT three times and take what it decides is the best idea. Give that idea as well as the design student generated ideas to the evaluator and see how it performs across all categories.
  + Evaluation only performance:
    - 1) Enter student ideas into MetaGPT evaluation stage. Compare the MetaGPT evaluation scores the real evaluators score and also a ChatGPT webchat score (I expect they will need to be mean-centre normalised)
    - 2) For one of the MetaGPT runs, as well as the best idea take the middle idea and the worst idea. Have the real evaluator score these ideas also and compare to MetaGPT scores. Instead of only focusing on identifying good ideas, this will see if they evaluate bad ideas similarly
* Participants:
  + Design students: My flatmate has said he should be able to get at least five design students in a room for the coming Friday (he can be flexible with the dates)
  + Evaluator: Unfortunately, all the Design staff he spoke to shut it down instantly pretty much because they didn’t have the time. As of right now, that just leaves you as the only person who has said they might be able to do it and has an expertise
    - Assuming we will talk about this on Friday
* Time cost:
  + Students: 15 minutes to come up with a product idea and fill in the template
  + Evaluator: This is where it could be problematic, the methodology as described and assuming 5 design students would mean evaluating 10 products ideas. Generously assuming it will only take 5 minutes per product idea, that’s still 50 minutes of the evaluators time
    - Would be made a lot more reasonable with two evaluators.
* Test run:
  + I’m going to do a run through with Ash after the meeting on Friday using the design brief I created
  + I’ll act as the evaluator (not an expert but the brief is about outdoor kit and I do work retail in that industry)
* Why I think this method of evaluating is interesting:
  + Coming up with new product ideas is something humans are already very good at, and is a very creative task which from what I’ve experienced when working on this project, LLMs struggle on
  + However, the hard part is often figuring out what is actually a good product idea. Here research has found most companies to not have a definitive process, and it is often left up to the manager of the team of the person who had the idea. This ends up being problematic, because the manager is only an expert in the subject of the team they manage, so the idea ends up failing due to an outside factor they do not expect.
  + Here I think LLMs could have a use case, they are able to look at the idea with a much less biased scope, and I hypothesise this could lead to a better evaluation
    - The problem with this is making an evaluation requires quite a large amount of reasoning capabilities, something that if it exists at all (seen arguments back and forth) only exists in a very limited capacity for GPT-4
    - However, MetaGPT essentially implements Chain of Thoughts, a method Google developers released a paper on that is meant to improving reasoning capabilities in LLMs
    - It would be interesting to see if MetaGPT case can outperform the web based GPT. If it does, then it could have a use as a framework.
* Finalised task brief/forms:
  + Had a hold up with this due to a personal matter meaning I couldn’t work for the first few days this week
  + Assuming your onboard with using the evaluation methodology above, I’m going to write up the task brief for the students, convert the methodology into a form (for the evaluator) and get that sent to you for Monday.
  + Ideally would be able to get feedback and adjust in time for getting design students together on Friday. However, there is the problem with who is going to be the evaluator and how much time they are willing to spend. Might have to delay till the week after. Really don’t want it to be any later than that.