**Introduction**

Result of comparisons of DV-2 and DV-3 on the best prompts.

**Content**

| Prompt | Da Vinci 2 | Da Vinci 3 |
| --- | --- | --- |
| Baseline (best prompt) | 4 | 1 |
| Baseline + Mechanisms | 2 | 4 |
| Baseline + Environment | 2 | 3 |
| Baseline + Contexts | 5 | 1 |

* In Conclusion, the performance of which model of Da Vinci does better is highly subjective to the prompt.
* Thus, more samples will need to be generated with Da Vinci 3, and then compared against Da Vinci 2.
* If the Da Vinci 2 model does better in each of the prompts, then we already know that the best prompt is the baseline.
* If Da Vinci 3 does better in any prompt, then we would have to do an overall comparison with the same prompts, but with the better-performing language model.