|  |  |
| --- | --- |
| **First Supervisor** | Fabrício Góes |
| **School/Department** | Computing and Mathematical Sciences |
| **Email**  | lfwg1@leicester.ac.uk | **Telephone Ext** |  |

|  |  |
| --- | --- |
| **Second Supervisor** | Marco Volpe |
| **School/Department** | Computing and Mathematical Sciences |
| **Email**  | mv163@leicester.ac.uk | **Telephone Ext** |  |

|  |  |
| --- | --- |
| **Additional Supervisor** |  |

**Section 2 – *Project Information***

|  |  |
| --- | --- |
| **Project Title** | Large Language Models as Autonomous Creativity Evaluators |
| **Project Highlights:** | 1. | Large Language Models Emergent Behaviour |
| 2. | Style/Personality Induction on Large Language Models |
| 3. | Automatic Generation/Evaluation of Creative Artefacts |
| **Project Summary (300 words maximum)** |
| Open AI’s ChatGPT has been recently in the news surprising people with its ability to chat as a human being. More than that, it can summarize texts, explain programming code, and also do creative tasks such as writing song lyrics, poetry and stories. In Computational Creativity, creative artefacts, such as stories and poetry, are considered creative if they present novelty (e.g. new compared to the existing ones) and value (e.g. funny for jokes). There are two main strategies to assess the creativity of artefacts: evaluation metrics and human judges. Evaluation metrics, proposed by humans, are usually used by generative systems to evaluate novelty and value of potential creative artefact candidates. The best ones are ultimately evaluated by humans, since they are still the ultimate judges on creativity. Despite evidence that non-expert judges cannot appropriately evaluate the creativity of a human or machine, studies have relied on hiring non-expert volunteers on crowd sourcing platforms, such as Amazon Mechanical Turk, to evaluate/rate artefacts in the creative domain. These studies usually do not ask volunteers to explain the reasoning behind their scores, but accept their judgement as valid. In this proposal, we assume that machines, like humans, can judge creative artefacts using their own intrinsic evaluation metrics. This assumption is backed on recent advances of large language models (LLMs) such as ChatGPT/GPT-4/Claude 3.0 that enable emergent behaviours, such as the ability of rating creative artefacts that they were not explicitly trained to rate. On top of it, some very recent publications show that LLMs can be configured/prompted to assume different styles/personalities with zero or few-shot learning. This means that LLMs can rate the same artefact from different points of view. We argue that by deploying a LLM configured with different styles we can accurately rate the creativity of artefacts, without the need of human judges (https://le.ac.uk/people/fabricio-goes). The goal of this proposal is to study, investigate and develop a novel method to enable LLMs as autonomous creativity evaluators. This novel method minimizes human participation in the loop of assessing creative artefacts, thus accelerating the automatic generation of creative artefacts.  |