**Types of Tests**

All the models were initially tested on one task to see how they responded initially, this was done either by Requests (Inference API) or LangChain.

If the models didn’t show a proper response or I wasn’t able to interact with them, I excluded them from the full testing. To include a model in the final testing, they had to at least perform well in one task.

During the initial inspection of the model, I did not change any hyperparameter of the models for the test.

**Initial Tests**

**Summarization**

The models were given different sections of medical text to summarize. The goal is to see if the models were just choosing the most important sentence or creating and actual breakdown of the ideas.  
  
**Criteria:** For the model to fail, they had to be creating new text or adding out of context information. They also had summarize and not just repeat the information given.

**Type Models Tested**: Text Generation, Text2Text, Summarization

**Results**: Some models couldn’t pass this stage as they didn’t take that many tokens like the “Falcon” model (Text Generation).   
Summarization only models are the lightest and perform ok, but they tend to only short sentences.  
Text2Text models performed the best as they “reason” and conceptualise the information better. For this tasks as they grew in parameters the better they performed, particularly the models tested on the HuggingFace Space’s like RWKV.

**Text Generation (Questions)**

In this section I asked a series of general and medical related questions to see what the models respond. To include models on the next section of test (formal test) all models have to pass this test.

**Criteria:** For the model to fail, they had to be creating new text or adding out of context information. They have to at least recognise who’s Elton John, from the general questions asked.

**Type Models Tested**: Text Generation, Text2Text.

**Results**: Most models do ok with general questions, but not as good with medical questions. Some finetuned models create to much out of context information.

Text Generation models tend to perform better with general hyperparameters.

Text2Text tend to use the question as answer but perform better once hyperparameters are changed.

**Question Asking with Context**

In this section I gave the models context and asked them questions about it. This test have to be performed different for the models that are specialised for this particular task vs the generative models.

For the generative models I created a prompt for the question and context.

**Criteria:** At least get one question right

**Type Models Tested**: Text Generation, Text2Text, QA models.

**Results**: Like summarization, Text Generation models can’t take to much context and don’t really understand the instructions.

QA models, are much lighter than the other models, they can do ok with simple questions that are one or two word answers but struggle with longer answers that are not explicitly written on text.

Text2Text models perform the best even with the light versions.

**Zero Shot Classification**

This type of test is for the Zero-shot Classification models, which they take some text and a list of possible classes to choose from.

This models are light but when I tested them I found that independently of the classes I gave them, the results were based on the location of the list of the classes I gave

As and example, I will give them a text and the following labels:

Labels: ['AI in Health', 'Clinical Trail', 'Medical Recovery', 'Patient Study', 'Digital Twin']

They will answer the following percentages:

Answer: [0.21885918080806732, 0.21561461687088013, 0.20043905079364777, 0.18664546310901642, 0.17844170331954956]

If I change the order of the labels, the results will still be the same.   
  
I used different models and they all did the same, this could be due the InferenceAPI interaction, but I didn’t purse this any further.

**Second Test**

After looking at how the models work and to make testing more efficient, I created a testing pipeline that allows to evaluate multiple models at the time in the following tasks:

* Summarization
* Text Generation (Question Asking)
* QA with Context

The program created allow testing multiple model and has some troubleshooting implementations when the models are not responsive or to large to be able to interact with.

This results are saved on a .csv table under the name of Test\_Results, the initial interaction is saved under Initial\_Test.

For this test, all the models had the same hyperparameters:

        self.\_model\_kwargs = {

            "min\_length": 10,

            "max\_length": 60,

            "temperature": 1,

            "top\_p": 0.3,

            "early\_stopping": True,

            "length\_penalty": 1,

            "num\_beams": 5,

            "no\_repeat\_ngram\_size": 2,

            "do\_sample": True,  # False for summarisation

            "repetition\_penalty": 1.2,

            }

The following is a list of the main hyperparameters found on the models and their meaning:

**min\_length**: The minimum length of the generated text. The model will continue generating text until this length is reached.

**max\_length**: The maximum length of the generated text. The model will stop generating text once this length is reached.

**temperature**: Controls the randomness in the model's responses. A higher temperature leads to more varied outputs, while a lower temperature results in more predictable text.

**top\_p (Nucleus Sampling)**: The model considers only the top p% of the most likely next words for generating text. It helps in focusing the model's outputs.

**early\_stopping**: If set to True, the model stops generating text once the end-of-text token is reached.

**length\_penalty**: Adjusts the length of the generated text. Higher values encourage shorter outputs, and lower values favor longer ones.

**num\_beams (Beam Search)**: Number of beams in beam search. A higher number improves output quality but increases computational cost.

**no\_repeat\_ngram\_size**: Prevents the model from repeating the same n-grams, helping to avoid repetitive text.

**do\_sample**: Determines if sampling is used for response generation. True enables more random, diverse outputs.

**verbose**: Controls the level of detail in the output. When True, more detailed information is provided.

**repetition\_penalty**: Discourages the model from repeating the same words, enhancing output diversity**.**

After creating the PDF extraction for the abstracts, I created two new type of test for the models. This test is not included (as of now) on the testing pipeline, mainly as this test only works with the Text2Text models.

The Test can be found on the Test\_LangChain code, and it has the following two options:

**Classification with Labels (Text2Text Models)**

For this test I followed the same idea of providing context and possible labels in the prompt. As the Text Generation models fail with long inputs, I excluded them from this testing. In the prompt I included the abstract of a research paper as input and gave a list of 5 different labels.

As a reference label, I did a Topic Modelling on the abstracts and classified according to it

**Criteria:** Choose a label from the label options

**Type Models Tested**: Text2Text

**Results**: The models understand the instructions, I had to reduce the number of outputs allowed as they kept generating more text.

**Classification no Labels (Text2Text Models)**

Same idea as the previous but no labels included.

**Criteria:** Give a reasonable answer, don’t create non sense information

**Type Models Tested**: Text2Text

**Results**: The answers make sense and do change from abstract to abstract. This models require fine tunning and can be done if you take the prompt as example. The Alpaca Methodology follows the idea of creating prompt as a training for the models to understand how to answer in the future.

“Effective was of Prompt Fine Tunning”

<https://www.philschmid.de/instruction-tune-llama-2>  
<https://github.com/tatsu-lab/stanford_alpaca#data-release>

**Appendix – Questions for the Tests  
  
Summarization:**

Question 1.-  
Diabetes Intervention Study. Multi-intervention trial in newly diagnosed NIDDM. OBJECTIVE: In a randomized 5-yr multi-intervention trial, we tested the efficacy of intensified health education (IHE) in improving metabolic control and reducing the level of coronary risk factors and incidence of ischemic heart disease (IHD). RESEARCH DESIGN AND METHODS: Within the intervention group, the benefit of clofibric acid was evaluated in a double-blind study. One thousand one hundred thirty-nine newly diagnosed middle-aged (30- to 55-yr-old) patients with non-insulin-dependent diabetes mellitus (NIDDM) entered the study. They were classified as diet controlled after a 6-wk screening phase with conventional dietary treatment. During the follow-up, the control group (n = 378) was cared for at different diabetes outpatient clinics with a standardized surveillance. The intervention group (n = 761) had a structured IHE that included dietary advice, antismoking and antialcohol education, and ways to enhance physical activity. RESULTS: Randomly, 379 of the IHE patients received 1.6 g clofibric acid/day, and the others received placebo. IHE resulted in improved glucose control (adjusted fasting blood glucose) levels after 5 yr (control subjects 9.27 mM, IHE group 8.71 mM, and IHE plus clofibric acid group 8.60 mM, P less than 0.01). The better glycemic control was achieved with fewer antidiabetic drugs. After 5 yr, antidiabetic drugs were prescribed to 47% of the control subjects, 28% of the IHE group, and 34% of the IHE plus clofibric acid group (cutoff limit for drug application was postprandial blood glucose of greater than or equal to 13.87 mM). The ratio of polyunsaturated to saturated fatty acids (0.26 vs. 0.40, P less than 0.01) and physical activity (174 vs. 327 scores, P less than 0.01) were increased, and blood pressure, tobacco, and alcohol consumption were significantly reduced by IHE. However, IHE had no effect on calorie intake, percentage of fat in the diet (45%), and body weight. The most important finding was the significant increase of blood cholesterol in all three groups (+0.47, +0.36, and +0.34 mM, respectively). Clofibric acid only prevented the increase of triglyceride levels (+0.56, +0.24, and +0.05 mM, respectively). The incidence rate per 1000 for myocardial infarction was 30.3 for control subjects, 53.6 for the IHE group, and 55.6 for the IHE plus clofibric acid group. The corresponding rates for IHD incidence were 90.9, 97.8, and 98.8, respectively. Men suffered more frequently from myocardial infarction, whereas women developed ECG criteria for IHD more frequently. Among the 35 cases of death, besides cardiovascular diseases, liver cirrhosis and neoplasia were the predominant causes. The death rate per 1000 in control subjects was 46.2, 30.6 in the IHE group, and 27 among patients with IHE plus clofibric acid. CONCLUSIONS: IHE was of substantial benefit for the control of glycemia, significantly diminished the need for antidiabetic drugs, and reduced a cluster of risk factors but had no effect on the control of blood lipids. This could be one major reason for the failure of IHE, effective lowering of blood pressure, and clofibric acid to prevent cardiovascular complications. Clofibric acid was only effective in reducing triglycerides.

Question 2.-  
Oxygen transport measurements to evaluate tissue perfusion and titrate therapy: dobutamine and dopamine effects. BACKGROUND: Increased cardiac index, oxygen delivery (DO2), and oxygen consumption (VO2) patterns were shown to characterize the physiologic status of surviving high-risk surgical patients, and indicate increased metabolic needs; relatively normal DO2 and VO2 values were found to characterize the sequential pattern of nonsurvivors who developed an early oxygen debt followed by lethal organ failure. The cardiac index, DO2, and VO2 values empirically determined from survivors' patterns were shown to improve outcome in prospective randomized trials. The present study considers these criteria to evaluate the tissue perfusion status as well as the effects of therapy on tissue perfusion and oxygenation. OBJECTIVE: To summarize new information on the temporal patterns of DO2, VO2, and oxygen debt on outcome and the effects of fluids and inotropes on these patterns in a wide range of clinical, temporal, and physiologic conditions. DESIGN: Descriptive analysis based on data gathered prospectively using a specified protocol. PATIENTS: High-risk patients with accidental or elective surgical trauma, and patients with or without sepsis or septic shock and organ failure. SETTING: University-run county hospital with a large trauma service. INTERVENTIONS: Fluids, dobutamine, and dopamine at various times and at various doses throughout critical illness of postoperative, posttraumatic, septic, and hypovolemic patients with and without lethal and nonlethal organ failure. MEASUREMENTS AND MAIN RESULTS: The pattern of DO2 plotted against the corresponding VO2 values in 437 consecutive critically ill surgical patients showed a wide variability and poor correlation probably because complex clinical conditions may obscure the supply-dependent and supply-independent VO2 relationships observed in normal dogs bled or given bacterial infusions. However, the use of specific therapy by well-defined protocols was shown to provide objective evidence of efficacy. Significant increases in DO2 and VO2 were previously shown after whole blood, packed red cells, and colloid administration, but not after crystalloid administration. Dobutamine administration in 715 circumstances in postoperative, traumatic, septic patients and patients with adult respiratory distress syndrome, renal failure, and multiple organ failure significantly improved DO2 and VO2. Dopamine under comparable conditions produced less improvement in DO2 and VO2 than that of dobutamine; most of the VO2 changes were not significant. CONCLUSIONS: The monitored patterns of cardiac index, DO2, and VO2 may be used to evaluate the adequacy of tissue perfusion as well as the relative effectiveness of alternative therapies. Second, these physiologic criteria may be used to titrate therapy in order to achieve optimal outcome. Third, after colloids optimally expand the plasma volume, dobutamine may be used to enhance flow and the distribution of flow in order to improve tissue oxygenation. Vasodilators may be used when hypertensive episodes occur or there is an inadequate response to inotropic agents. Vasopressors are used as a last resort, usually in the terminal or preterminal state.

Question 3.-  
Excision of limbal dermoids. We reviewed the clinical files of 10 patients who had undergone excision of unilateral epibulbar limbal dermoids. Preoperatively, all of the affected eyes had worse visual acuity (P less than .02) and more astigmatism (P less than .01) than the contralateral eyes. Postoperatively, every patient was cosmetically improved. Of the eight patients for whom both preoperative and postoperative visual acuity measurements had been obtained, in six it had changed minimally (less than or equal to 1 line), and in two it had improved (less than or equal to 2 lines). Surgical complications included persistent epithelial defects (40%) and peripheral corneal vascularization and opacity (70%). These complications do not outweigh the cosmetic and visual benefits of dermoid excision in selected patients.

**Text Generation:**

Question 1.- "What are ICD-10 codes used for?",

Question 2.- "Who is Elton John?",

Question 3.- "What’s clinical NLP?",

Question 4.- "what are the Symptoms of Malaria?",

Question 5.- "What is the ICD10 code of Malaria?"

**Question Asking With Context:**

**Context 1.-** My name is Clara and I live in Berkeley.  
**Question 1.-** What’s my name?

**Context 2, 3, 4.-** Excision of limbal dermoids. We reviewed the clinical files of 10 patients who had undergone excision of unilateral epibulbar limbal dermoids. Preoperatively, all of the affected eyes had worse visual acuity (P less than .02) and more astigmatism (P less than .01) than the contralateral eyes. Postoperatively, every patient was cosmetically improved. Of the eight patients for whom both preoperative and postoperative visual acuity measurements had been obtained, in six it had changed minimally (less than or equal to 1 line), and in two it had improved (less than or equal to 2 lines). Surgical complications included persistent epithelial defects (40%) and peripheral corneal vascularization and opacity (70%). These complications do not outweigh the cosmetic and visual benefits of dermoid excision in selected patients.

**Question 2.-** How many patient files? **Question 3.-** Give me the insights? **Question 4.-** What is the problem?

**Classification (No Label & Label)**

**Prompt No Label**From the following text, can you classify the type of text to a relevant topic in Medicine.

Text: {question}

Answer:

**Prompt Label**From the following text, can you classify the type of text with one of the following labels:

Labels: [ 'Digital Twin', 'Clinical Trail', 'Patient Study', 'AI in Health', 'Medical Recovery']

Text: {question}

Answer:

**Abstracts  
1.-** TECHNICAL ABSTRACT Background: Industry 4.0 presents itself as an ecosystem; a collection of elements endowed with Cyber-Physical Systems and Augmented Reality/Virtual Reality devices, which are connected through the Internet of Things, and uploaded to Cloud Platforms for analysis, knowledge extraction, and diagnostics through Cognitive Computing based on a large amount of data.

The concept is centered around data: managing data, analyzing data, and controlling data.

Many factors influence this interconnected working environment, and for that reason planning and implementing the digital transformation implies many challenges.

Industry 4.0 and Ergonomics are being integrated using a variety of tools and approaches, thus supporting the development of an Ergonomics 4.0 concept.

Purpose: This paper reviews studies focusing on the determinants of Ergonomics 4.0, identifying the main elements and their interrelationships with a focus on Digital Human Modeling and Cyber-Physical Systems.

We consider approaches such as Operator 4.0 and ‘Modeling and Simulation for Digital Twin Creation’, which aim to accelerate the decisionmaking and adaptation processes.

We identify the leading technologies, operations, and worker-related aspects through a qualitative meta-analysis, to establish elements and interrelationships of Ergonomics 4.0 determinants.

Methods: Literature was selected from articles recent journal publications, and a qualitative evaluation was performed using semantic meta-analysis.

The findings were then used to develop a theoretical taxonomy of determinants of Ergonomics 4.0 in Industry 4.0 based on various classifiers, which were structured and interlinked.

Results: The five areas categorized include: Industry 4.0 technology, Human-Cyber-Physical Systems, Operator 4.0, Human-Robot collaboration, Digital Twin and Digital Human Modeling.

The proposed conceptual framework for Ergonomics 4.0 describes processes, technology, information, and structures, which occur in Industry 4.0 as Operator 4.0, Human-Robot collaboration, Digital Twin, Digital Human Modeling and eventually define Ergonomics 4.0.

The concepts of Digital Twin and Digital Human Modeling are analyzed in detail, as they form the core of Ergonomics 4.0.

Conclusions: We propose a conceptual framework for Ergonomics 4.0 as a Cyber-Physical System and discuss aspects of Digital Human Modeling and Digital Twin that are essential to understanding the role of Ergonomics and the integration of Ergonomics into Industry 4.0.

**2.-**   
ABSTRACT Big data and AI have enabled digital simulation for prediction of future health states or behaviors of specific individuals, populations or humans in general. “Digital simulacra” use multimodal datasets to develop computational models that are virtual representations of people or groups, generating predictions of how systems evolve and react to interventions over time.

These include digital twins and virtual patients for in silico clinical trials, both of which seek to transform research and health care by speeding innovation and bridging the epistemic gap between population-based research findings and their application to the individual.

Nevertheless, digital simulacra mark a major milestone on a trajectory to embrace the epistemic culture of data science and a potential abandonment of medical epistemological concepts of causality and representation.

In doing so, “data first” approaches potentially shift moral attention from actual patients and principles, such as equity, to simulated patients and patient data.

**3.-**

Abstract The Digital Twins of product and production face the challenge that once the development process is closed, they does not reflect the real status of production where events as equipment failures, poor quality or missing compound parts happens continuously.

For assuring production resilience a holistic methodology, combining a top-down with a bottom-up 3D scanning approach for capturing real-time production parameters and embedding them in Digital Twins is developed.

The paper presents the methodology and a motivation scenario for further validation in an In today's business environment, the trend towards more product variety and customization is unbroken.

Due to this development, the need of innovative set-up of an automated measurement cell, where state-of-the-at technologies as autonomous AGV, mobile 3D laser scanning and agile and reconfigurable production systems emerged to cope with various products and product families.

To design and optimize production automated processes are integrated.

**4.-**

Abstract All people are unique and so are their diseases.

Our genomes, disease histories, behavior, and lifestyles are all different; therefore it is not too surprising that people often respond differently when administered the same drugs.

Cancer, in particular, is a complex and heterogeneous disease, originating in patients with different genomes, in cells with the different epigenomes, formed and evolving on the basis of random processes, with the response to therapy not only depending on the individual cancer cell but also on many features of the patient.

Selection of an optimal therapy will therefore require a deep molecular analysis comprising both the patient and their tumor (e.g., comprehensive molecular tumor analysis [CMTA]), and much better personalized prediction of response to possible therapies. As the knowledge base on cancer, cellular transduction and molecular interactions widens, so does our ability to generate computational models with the capacity to accurately represent the complex networks and cross-talk determining cancer progression and drug response the main advantages of mechanistic models are the integration of data from Currently, we are at an inflection point in which advances in technology, decreases in the costs of sequencing and other molecular analyses, and increases in computing advances are converging, forming the foundation to build a data-driven approach to personalized oncology.

In this article we discuss the deep molecular characterization of individual tumors and patients as the basis of not only current precision oncology but also of computational models (‘digital twins’), the foundation for a truly personalized therapy selection of the future.

**5.-**

Digital Twin (DT) is a 'living' entity that offers potential with monitoring and improving functionality of interconnected complex engineering systems (CESs).

However, lack of approaches for adaptively connecting the existing brownfield systems and their data limits the use of DTs.

This paper develops a new DT design framework that uses ontologies to enable co-evolution with the CES by capturing data in terms of variety, velocity, and volume across the asset life-cycle.

The framework has been tested successfully on a helicopter gearbox demonstrator and a mobile robotic system across their life cycles, illustrating DT adaptiveness without the data architecture needing to be modified.