

order number:

Thesis submitted to the
UNIVERSITY OF MOHAMED BOUDIAF – MSILA



FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

In partial fulfillment of the requirements for the degree of
Master in Artificial Intelligence

By
Abdat Aya

Title of the thesis

AI-Powered Skin Analysis and Recommendation System

Under the supervision of
Lamri Sayad

Composition of the jury

Rached Yagoubi	University of Msila	President
Lamri Sayad	University of Msila	Reporter
Abdelouahab Khettaf	University of Msila	Examiner

June, 2026

Dedications :

" This thesis is dedicated to the people who built me, supported me, and walked beside me every step of the way. To my Parents, Thank you for your endless love and sacrifices. You gave me the foundation to dream and the strength to reach this moment. Everything I achieve is a reflection of your hard work. To my Grandma and my late Grandpa, Grandma, your quiet prayers and supportive words kept me going. Grandpa, though you are not here to see this day, your memory stays with me. I hope I have made you proud. To my Friends, Hiba, Nourhane and Imane, Thank you for turning a challenging journey into the most beautiful years of my life. From the stressed, late-night study sessions to the endless laughter, I couldn't have crossed this finish line without you. To my Teachers, who gave me a passion that I will carry into my future."

Aya Abdat.

Acknowledgments

I am deeply indebted to my thesis supervisor Prof. Lamri Sayad, whose unlimited steadfast support and inspiration have made this project a great success. In a very special way, I thank him for every form of support he has rendered unto me to see that I succeed in this challenging study.

Contents

List of figures	8
List of tables	9
General Introduction.....	10
Chapter 1: Artificial Intelligence	13
1.1 Introduction	13
1.2 Artificial Intelligence.....	14
1.2.1 Definition and Brief History.....	14
1.2.2 Machine Learning.....	16
1.2.3 Deep Learning	18
1.2.4 Large Language Models and Multimodal Models	21
1.2.5 Artificial Intelligence in Healthcare	22
1.3 Recommendation Systems.....	24
1.3.1 Definition and Purpose	24
1.3.2 Main Types of Recommendation Systems	25
1.3.3 Recommendation Systems in Health and Skincare	27
1.3.4 Evaluation of Recommendation Systems	28
1.4 Conclusion.....	28
Chapter 2: Skincare and Related Work	30
2.1 Introduction	30
2.2 The Skincare Domain	30
2.2.1 The Global Skincare Industry.....	30
2.2.2 Human Skin: Structure and Function	32
2.2.3 The Fitzpatrick Skin Phototype Classification	34
2.2.4 Active Ingredients in Skincare	35
2.2.5 The Consumer Information Problem.....	37

2.2.6 How Skincare Guidance Is Currently Delivered	38
2.3 Related Work.....	39
2.3.1 Artificial Intelligence in Dermatology: Research Systems	39
2.3.2 Commercial Skincare AI Applications.....	42
2.4 Conclusion	43
Chapter 3: Implementation and Results	45
3.1 Introduction	45
3.2 Tools	46
3.2.1 Overview of the Technology Stack	46
3.2.2 Backend: Python and FastAPI.....	47
3.2.3 Frontend: React and Vite.....	47
3.2.4 The Multimodal Model: GPT-4o.....	47
3.3 Dataset	48
3.3.1 A Test Set, Not Training Data	48
3.3.2 Source of the Images	48
3.3.3 The Two Ground-Truth Attributes	49
3.3.4 The Concern Taxonomy	49
3.3.5 Composition of the Test Set	50
3.3.6 Evaluation Protocol	50
3.4 Implemented System	51
3.4.1 Architecture Overview	51
3.4.2 The Data Contract.....	52
3.4.3 Agent 1: The Image Quality Gate.....	53
3.4.4 Agent 2: The Vision Agent (Seed)	54
3.4.5 Agent 3: The Vision Refinement Agent (Overwrite)	54
3.4.6 Agent 4: The Profile Agent	55

3.4.7 Agent 5: The Recommendation Agent	55
3.4.8 Agent 6: The Safety Validation Agent	56
3.4.9 Agent 7: The Prose Layer Agent	57
3.4.10 The User-Facing Flow	57
3.4.11 An End-to-End Example	59
3.5 Results	61
3.5.1 Evaluation Metrics.....	61
3.5.2 Fitzpatrick Classification.....	61
3.5.3 Per-Concern Detection	62
3.5.4 Where the System Works and Where It Struggles	63
3.5.5 Behaviour of the Safety Layer.....	64
3.5.6 Comparison with Commercial Approaches.....	64
3.5.7 Threats to Validity	65
3.6 Conclusion	65
General Conclusion	67
Bibliography.....	70

List of figures

Fig. 1: Main milestones in the history of artificial intelligence	15
Fig. 2: The three main paradigms of machine learning	16
Fig. 3: An artificial neuron (a) and a simple feedforward neural network (b).....	19
Fig. 4: A typical Convolutional Neural Network (CNN) architecture for image classification	20
Fig. 5: High-level structure of a single Transformer layer.....	21
Fig. 6: How a multimodal large language model processes a text-and-image input	22
Fig. 7: General structure of a recommendation system	25
Fig. 8: Comparison between content-based filtering and collaborative filtering.....	25
Fig. 9: Cross-section of human skin, showing the three main layers and their main functions	33
Fig. 10: The Fitzpatrick scale of skin phototypes	34
Fig. 11: Common interactions between active skincare ingredients.....	37
Fig. 12: The five channels through which skincare guidance is currently delivered, compared on reliability, accessibility, and personalisation	38
Fig. 13: The seven-agent pipeline. Learned stages (GPT-4o) are shaded; all safety-bearing stages are deterministic.....	52
Fig. 14: The photo-capture screen, requesting a front view and optional cheek profiles	58
Fig. 15: The four questions of the skin questionnaire: skin type, reactivity, age band, and primary concern.....	59
Fig. 16: The progress screen, naming the stages of the pipeline as they execute	59
Fig. 17: The data contract as it grows through the pipeline for a single request	60
Fig. 18: Fitzpatrick accuracy: exact match against within-one-type.....	62
Fig. 19: F1 score for each of the eight concerns, ranked	63

List of tables

Table 1: The three commonly distinguished categories of artificial intelligence.....	15
Table 2: Common families of machine learning algorithms	17
Table 3: Selected studies on the application of deep learning in dermatology	23
Table 4: Comparison of the main families of recommendation systems	26
Table 5: Approaches to skincare recommendation in existing tools, compared with the system proposed in this thesis	27
Table 6: Selected industry estimates for the global skincare market.....	32
Table 7: Active ingredients commonly used in modern skincare products and their primary effects.....	35
Table 8: Selected studies in the academic literature on artificial intelligence in dermatology	39
Table 9: Categories of commercial AI-based skincare tools, compared with the system proposed in this thesis	43
Table 10: Principal components of the implemented system and their roles	46
Table 11: The eight-category concern taxonomy and the criterion for judging each present .	49
Table 12: Principal fields of the data contract that passes between the agents	53
Table 13: The nine fields carried by each product in the catalogue and their role in the logic	56
Table 14: Fitzpatrick classification accuracy on the test set.....	61
Table 15: Per-concern detection performance: precision, recall, and F1	62
Table 16: Structural comparison with a typical commercial skincare tool	65

General Introduction

Skin is the body's largest organ and, for most people, the most visible one. It is also the part of the body people are most often asked to take care of without ever being taught how. A person standing in a pharmacy aisle, or scrolling through an online store, is expected to choose between dozens of products that promise contradictory things, written in a vocabulary; actives, barrier function, comedogenic ratings, pH; that was never explained to them. The result is a familiar kind of quiet confusion. People buy products that do not suit their skin, combine ingredients that should not be combined, and abandon routines that might have worked if they had been built correctly in the first place. Very few of these people have done anything wrong. They simply never had access to the one thing that would have helped: a moment of informed, personalised guidance.

That guidance does exist. It is called a dermatologist. But a dermatology consultation is not something a person can reasonably obtain every time they need to buy a moisturiser. Appointments are expensive, waiting lists are long, and in many regions specialists are concentrated in a handful of cities. A doctor's time is a scarce and valuable resource, and it should be spent on diagnosis and treatment; on the problems that genuinely require medical expertise. The everyday question of "what should I be using, and in what order" is real, it affects a person's comfort and confidence and money, but it rarely rises to the level that justifies a clinical appointment. It falls into a gap: too important to leave to guesswork, not severe enough to take to a doctor.

This thesis is about that gap, and about whether it can be narrowed responsibly with the tools now available in artificial intelligence. The work presented here is a system that accepts a photograph of a person's face together with a short structured questionnaire, processes that input through a sequence of specialised software agents, and returns a personalised, bilingual skincare report. The report describes the person's apparent skin characteristics, flags anything that should be checked by a professional, and proposes a concrete routine built from a real, grounded product catalogue. The system is designed from the first line of code around a single principle, and it is worth stating that principle plainly before anything else: this tool assists, it does not diagnose, and it does not replace a doctor. Everything in the architecture exists to keep that promise true.

Stating the principle is easy. Enforcing it inside a working AI system is the difficult and interesting part, and it is where this thesis makes its contribution. Modern large language models, and GPT-4o in particular, are remarkably fluent. They can look at an image and produce a confident, well-written paragraph describing it. But fluency is not the same as reliability. A language model does not know when it is wrong, and it will phrase a guess in exactly the same assured tone it uses for a fact. In a system that touches a person's health, however indirectly, that property is dangerous. A model that smoothly recommends combining two irritating actives, or that fails to flag a lesion that a human should examine, has not made an obvious error; it has made a fluent one, and fluent errors are the hardest for a non-expert user to catch.

The central design decision of this work is therefore a division of labour, drawn deliberately and enforced structurally. The critical logic of the system; safety validation, the triage decision about whether a case should be referred to a professional, and the detection of conflicts between recommended products; is handled entirely by deterministic Python code. This logic is explicit, testable, and it behaves the same way every time it is run. The language model is given a narrower and more appropriate role: it handles language. It performs the visual analysis of the photograph and it writes the final report in natural, readable prose. It is, in effect, the system's eyes and its voice, but never its judgement. The judgement lives in code that a human wrote, can read, and can be held accountable for.

This separation is the answer this thesis offers to a question that the wider field is still working through: how does one build an AI application that is genuinely helpful in a sensitive domain without quietly inheriting the unreliability of the model at its core? The answer proposed here is not to make the model more trustworthy, that is not within the power of a developer, but to architect the system so that the model is only ever trusted with the things it is actually good at. Where correctness matters, code decides. Where expression matters, the model writes. The boundary between those two responsibilities is the real subject of this thesis.

Two further commitments shape the work and should be introduced here. The first concerns grounding. A skincare recommendation is only useful if it points to products a person can actually obtain, so the system does not invent products or describe idealised ingredients in the abstract. Its recommendations are drawn from a specific, real catalogue — that of Touché — which means every suggestion the system makes corresponds to something that exists and can be purchased. The second commitment concerns evaluation. A system of

this kind must be tested against the full range of human skin, not a narrow and convenient subset of it, because a tool that performs well only on lighter skin tones would reproduce one of the most documented failures in dermatological AI. For this reason the system is validated against the Fitzpatrick17k dataset, which spans the complete Fitzpatrick scale of skin types. The details of that validation, and an honest account of where the system performs well and where it does not, are given in Chapter 3.

It should be said clearly what this thesis does not claim. It does not claim to have built a diagnostic instrument, and it does not claim that an AI system can substitute for clinical judgement. The contribution is narrower and, the author would argue, more honest: a working demonstration that the everyday guidance gap can be addressed with a system that is useful, grounded, and bilingual, and that has its own limits built into its structure rather than appended as a disclaimer. The system knows what it is not allowed to decide, and it is built so that it cannot quietly decide those things anyway.

The remainder of this thesis is organised into three chapters. Chapter 1 establishes the technical ground the work stands on: artificial intelligence and how the field is usually divided, the branches of machine learning and deep learning, the large multimodal models the system relies on, and the family of recommendation systems to which it ultimately belongs. Chapter 2 turns from the technique to the domain: the structure of the skincare industry and the consumer information problem within it, the dermatological background needed to reason responsibly about skin, and a survey of the related academic and commercial work, which together frame the contribution of this thesis. Chapter 3 presents the system itself: the tools it is built on, the dataset used to evaluate it, a complete description of the seven-agent pipeline, agent by agent, including the reasoning behind the deterministic-versus-language division at each stage, and an honest account of the results, of where the system performs well and where it does not. A general conclusion then summarises the contribution, outlines directions for future work, and offers a closing reflection.

Chapter 1: Artificial Intelligence

1.1 Introduction

This thesis presents an AI-based system that helps people make better skincare decisions from a single photograph and a short questionnaire. Before describing what was built and how, it is necessary to establish the technical ground the work stands on. This chapter does that. It introduces the field of artificial intelligence, the branches of it that the system relies on, and the family of techniques, recommendation systems, that the system ultimately belongs to.

The chapter is organised as a movement from the general to the specific. Section 1.2 begins with artificial intelligence as a whole: what it is, how the field developed, and how it is usually divided into categories. From there it narrows to machine learning, the branch of AI concerned with systems that learn patterns from data, and then to deep learning, the sub-branch that has driven most of the visible progress of the last decade. The discussion then turns to large language models and multimodal models, the most recent and the most relevant developments for this thesis, since the system presented here uses a multimodal large language model to interpret photographs of the user's face. The section closes with a short look at how artificial intelligence is currently used in healthcare and, in particular, in dermatology.

Section 1.3 then turns to recommendation systems. A recommendation system, in the broad sense used in this thesis, is a system that takes some description of a user and returns a selection of items chosen from a catalogue. This is, structurally, what the proposed system does: it takes a description of a user's skin and returns a set of products. The section reviews the main families of recommendation systems, examines how they have been applied in health and skincare contexts, and discusses how they are typically evaluated.

Section 1.4 closes the chapter with a short summary and a transition to Chapter 2, which moves from the technical background established here to the domain background, the skincare industry and the existing landscape of related work. The aim of this chapter is not to be exhaustive. Each of the topics introduced here has a literature large enough to fill a thesis on its own. The aim is to give the reader the specific technical context needed to understand the design decisions made in Chapter 3, and no more than that.

1.2 Artificial Intelligence

1.2.1 Definition and Brief History

Artificial Intelligence (AI) is the branch of computer science concerned with building systems that perform tasks normally requiring human intelligence, such as understanding language, recognising images, solving problems, and learning from experience [1]. In other words, the goal of AI is to design machines that can carry out activities that humans usually do with their minds.

The idea of intelligent machines is older than the term itself. In 1950, the British mathematician Alan Turing published a paper titled "Computing Machinery and Intelligence" in the journal *Mind* [2]. In this paper, he asked the question "Can machines think?" and proposed a practical test, now known as the Turing test, in which a machine is considered intelligent if a human cannot distinguish its written answers from those of another human. The paper is widely regarded as one of the earliest foundations of artificial intelligence as a research field.

The term "artificial intelligence" itself was introduced six years later. In the summer of 1956, a small group of researchers, John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, organised a workshop at Dartmouth College in the United States, known today as the Dartmouth Summer Research Project on Artificial Intelligence [3]. This workshop is widely considered the official birth of AI as a research discipline, and the original proposal was based on the idea that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it [3].

Since 1956, the field has gone through several phases. The first decades, often called the era of symbolic AI, focused on systems that reasoned using explicit logical rules written by human experts. Expert systems, which encoded the knowledge of specialists in domains such as medicine and chemistry, were a typical product of this era. These systems were useful within their narrow domains but were difficult to extend and could not handle situations outside the rules they had been given.

A significant change occurred from the 1990s onward, as researchers began to rely less on hand-written rules and more on statistical methods that allowed systems to learn patterns directly from data. This shift gave rise to the modern field of machine learning, which is discussed in the next section. A further turning point came around 2012, when a new

generation of methods known as deep learning began to outperform previous approaches on tasks involving images, sound, and language. The most recent phase, beginning in the early 2020s, has been marked by the rise of large language models such as GPT, which can process both text and images and produce fluent natural-language responses [1]. The system presented in this thesis belongs to this most recent phase, and uses a multimodal large language model as one of its components. The main milestones of this development are summarised in Fig. 1.

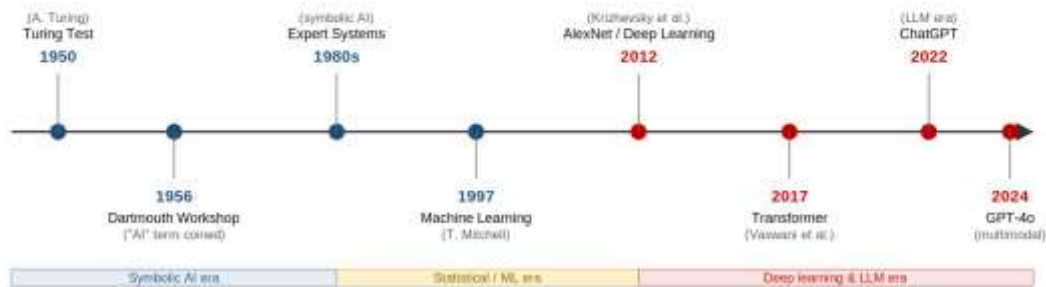


Fig. 1: Main milestones in the history of artificial intelligence.

A common way to organise the field of AI is to distinguish between narrow AI and general AI [1]. Narrow AI refers to systems designed to perform a single, well-defined task, such as recognising faces in photographs, translating between two languages, or recommending products to a user. All AI systems in practical use today, including the system presented in this thesis, are forms of narrow AI. General AI, sometimes called Artificial General Intelligence (AGI), refers to a hypothetical system that would match human intelligence across a wide range of tasks. General AI does not currently exist, and there is no agreement among researchers about when, or whether, it will be achieved. Table 1 summarises the three commonly distinguished categories of AI.

Table 1: The three commonly distinguished categories of artificial intelligence.

Type	Description	Examples
Narrow AI	Designed for one specific task; cannot transfer to unrelated tasks.	Image classification, language translation, product recommendation.
General AI (AGI)	Hypothetical system matching human intelligence across a wide range of tasks.	No real-world example exists yet.

Type	Description	Examples
Super AI	Hypothetical system surpassing human intelligence in essentially all domains.	Purely speculative; no existing system.

1.2.2 Machine Learning

Machine learning is the branch of artificial intelligence concerned with building systems that learn from data rather than being programmed with explicit rules. A widely cited definition by Mitchell describes a learning system as follows: a computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance at T , as measured by P , improves with experience E [4]. In simpler terms, a machine learning system is one that gets better at a task as it sees more examples.

This is a different approach from classical programming. In a classical program, a human writes the rules that decide the output from a given input. In a machine learning program, the human provides examples of inputs and the expected outputs, and the system itself adjusts its internal parameters so that, over many examples, its predictions become closer to the expected outputs. The rules are not written by hand; they are learned from data.

Machine learning is generally divided into three main categories, according to the kind of feedback the system receives during learning [1]. These three categories are illustrated in Fig. 2.

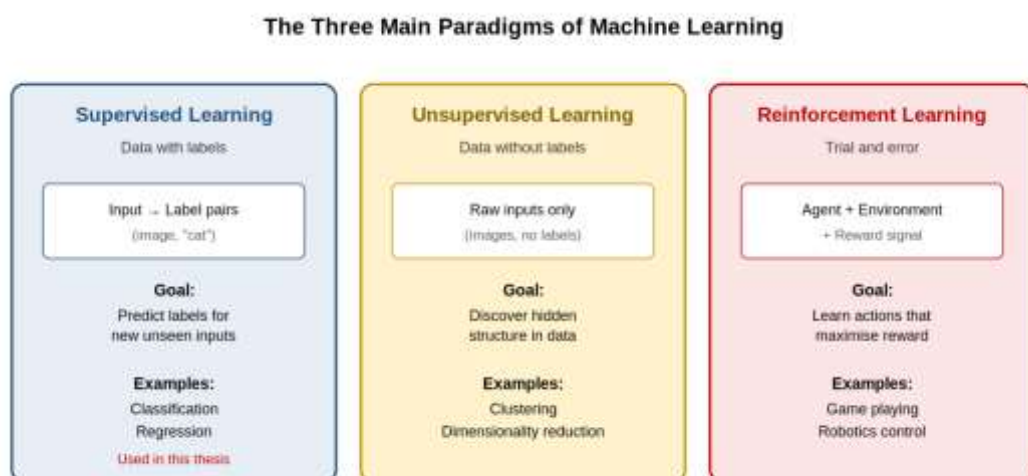


Fig. 2: The three main paradigms of machine learning.

Supervised learning is the most common category, and it is the one most relevant to this thesis. In supervised learning, the system is given a dataset in which each input example is paired with a correct output, often called a label. The system tries to learn the relationship between the inputs and the outputs, so that when a new, unseen input is given, it can predict the correct output. Image classification is a typical example: the system is shown many images, each labelled with the object it contains, and learns to predict the object class of new images. Supervised learning itself contains two main sub-types, classification, where the output is a category such as "cat" or "dog", and regression, where the output is a numerical value such as a price or a probability.

Unsupervised learning is used when the data does not come with labels. The goal in this case is to discover structure in the data on its own, for example by grouping similar examples together (clustering) or by reducing the number of variables needed to describe the data (dimensionality reduction). Unsupervised learning is useful for exploring data when the relevant labels are not known in advance.

Reinforcement learning is a third category in which a system, called an agent, learns by interacting with an environment. The agent takes actions, receives rewards or penalties, and gradually adjusts its behaviour so that, over time, the total reward it receives is as large as possible. Reinforcement learning has been used in domains such as game playing and robotics. It is not used in the system presented in this thesis.

Within these three categories, many specific algorithms have been developed. Table 2 summarises some of the most widely used families of machine learning algorithms, with their typical use cases and their main strengths and weaknesses. The choice of algorithm in any practical project depends on the kind of data available and the task to be solved.

Table 2: Common families of machine learning algorithms.

Family	Typical use	Strengths	Weaknesses
Decision Trees	Classification, regression	Easy to interpret; handles mixed data	Tends to overfit; unstable
Random Forests	Classification, regression	More stable than single tree; robust	Slower; less interpretable

Family	Typical use	Strengths	Weaknesses
Support Vector Machines	Classification	Effective on small datasets	Sensitive to parameter choice
k-Nearest Neighbours	Classification, regression	Simple; no explicit training step	Slow at prediction time
k-Means	Clustering (unsupervised)	Fast; simple to implement	Number of clusters must be chosen
Neural Networks	Many tasks	Very flexible; basis of deep learning	Needs much data and computation

Machine learning forms the foundation on which the next section, deep learning, is built. Deep learning is not a separate field from machine learning; it is a particular family of machine learning methods that has become especially important in the last decade.

1.2.3 Deep Learning

Deep learning is a sub-field of machine learning that uses artificial neural networks with many layers, hence the word "deep". In a landmark review article published in Nature in 2015, LeCun, Bengio, and Hinton described deep learning as a set of methods that allow computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction [5]. The key idea is that each layer of the network transforms the data into a slightly more abstract representation than the previous one, so that, by the end of the network, the data is described in terms that are useful for the task at hand.

The basic building block of a deep learning model is the artificial neuron, a simple mathematical unit that takes several numerical inputs, multiplies each by a weight, adds them together, and passes the result through a non-linear function called an activation function. A single neuron on its own can only solve very simple problems. The power of neural networks comes from connecting many neurons together in layers, so that the output of one layer becomes the input of the next. The weights of all the neurons are adjusted automatically during training using an algorithm called backpropagation, which calculates how each weight contributed to the error and updates it accordingly [5]. Fig. 3 illustrates a single artificial neuron and a small feedforward network.

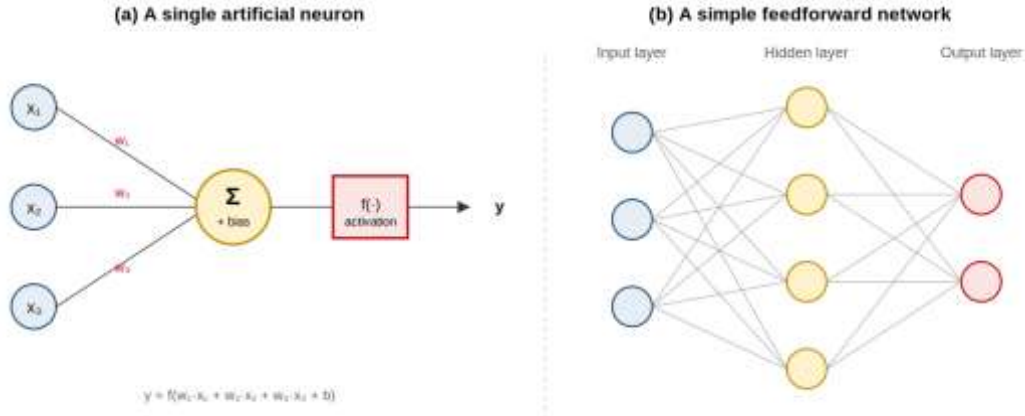


Fig. 3: An artificial neuron (a) and a simple feedforward neural network (b).

Although the ideas behind neural networks date back to the 1950s and 1960s, deep learning only became dominant around 2012. Three factors made this possible. The first was the availability of large labelled datasets, such as ImageNet, which contained millions of images organised into thousands of categories. The second was the development of more powerful hardware, particularly graphics processing units (GPUs), which allowed neural networks with many layers to be trained in a reasonable amount of time. The third was a series of methodological advances, including better activation functions and better regularisation techniques.

The turning point is widely associated with the work of Krizhevsky, Sutskever, and Hinton, who in 2012 trained a deep convolutional neural network, later called AlexNet, on the ImageNet dataset and achieved a top-5 error rate considerably lower than the previous state of the art [6]. From that point on, deep learning rapidly became the dominant approach in computer vision, and it later spread to other fields, including speech recognition, natural language processing, and medical imaging.

Several specific types of deep neural networks have been developed for different kinds of data. The three most important for the present discussion are convolutional neural networks, recurrent neural networks, and transformers.

Convolutional neural networks (CNNs) are designed for data that has a grid-like structure, such as images. Instead of connecting every neuron in one layer to every neuron in the next, a CNN uses a small filter that is applied across the entire image, detecting patterns such as edges, textures, and shapes. Deeper layers in the network detect more complex patterns built from the simpler patterns found in earlier layers. This is the architecture used by AlexNet and

by most modern image classification systems, including those used in dermatology [5], [6]. A typical CNN architecture is shown in Fig. 4.

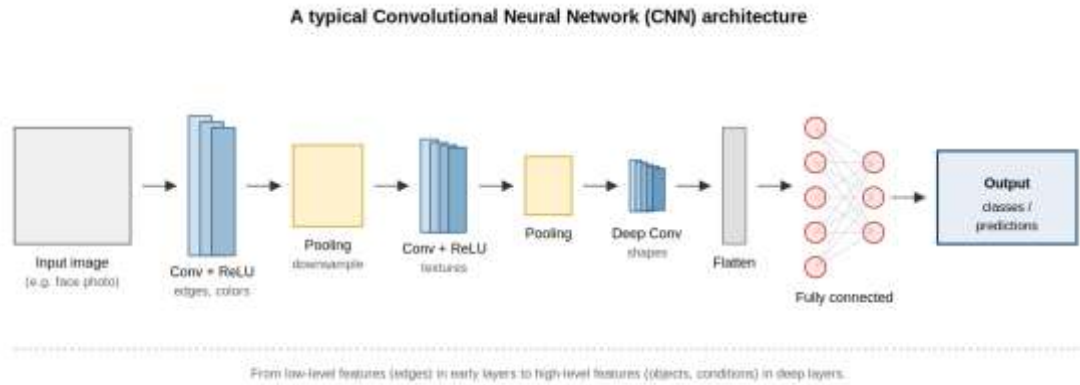


Fig. 4: A typical Convolutional Neural Network (CNN) architecture for image classification.

Recurrent neural networks (RNNs) are designed for sequential data, such as text or time series, in which the order of the elements matters. They contain connections that loop back, allowing the network to maintain a kind of internal memory of what it has seen so far. Variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were developed to address some of the difficulties of training basic RNNs. RNNs have largely been replaced by transformers in the most recent generation of language models, but they remain important in some sequence-processing applications.

Transformers are a more recent architecture, introduced by Vaswani et al. in 2017 in a paper titled "Attention Is All You Need"[7]. Instead of processing data step by step, as RNNs do, transformers use a mechanism called self-attention to consider all the elements of a sequence at the same time and decide which ones are most relevant to each other. This makes them much faster to train on modern hardware and much better at handling long-range relationships in the data. Transformers are the architecture behind all modern large language models, including the model used in the system presented in this thesis. The high-level structure of a transformer layer is shown in Fig. 5.

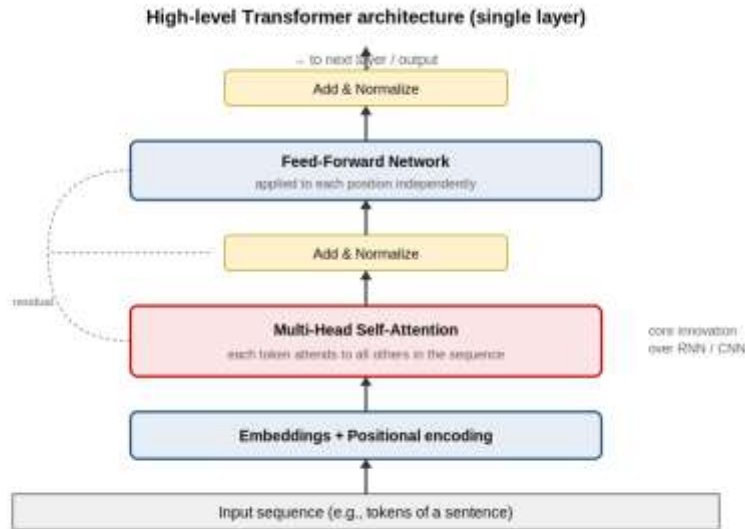


Fig. 5: High-level structure of a single Transformer layer.

1.2.4 Large Language Models and Multimodal Models

Large language models, often abbreviated as LLMs, are a particular kind of deep learning system. They are based on the transformer architecture described in the previous subsection, and they are trained on very large amounts of text drawn from the internet, books, and other sources. The training procedure is straightforward in principle: at each step, the model is shown a piece of text and asked to predict the next word, and its parameters are adjusted so that its prediction is closer to the actual next word. Because the training data covers a very wide range of topics and writing styles, an LLM trained in this way ends up able to produce fluent text on almost any subject and to follow instructions written in natural language.

The first widely known LLM was the GPT-2 model released by OpenAI in 2019, followed by GPT-3 in 2020 and GPT-4 in 2023. Each new generation of these models is larger than the previous one, both in the number of parameters (the internal numerical values that the model learns during training) and in the amount of training data. Other organisations, including Google, Meta, and Anthropic, have developed their own LLMs, and the field has grown rapidly in the last few years.

A multimodal model is a model that can process more than one type of input. Most early LLMs were text-only, meaning they could only read and produce text. More recent models can also accept images, audio, or video as input, alongside text. The model used in this thesis, GPT-4o, is one such multimodal model. It was released by OpenAI in May 2024 and is designed to accept text and images as input and produce text as output [8]. The letter "o" in its name stands for "omni", indicating that it is intended to work across multiple modalities. For

the purposes of the system presented in this thesis, the most important feature of GPT-4o is its ability to look at a photograph of a person's face and produce a written description of the visible skin characteristics. Fig. 6 shows, at a high level, how a multimodal model of this kind processes a text-and-image input.

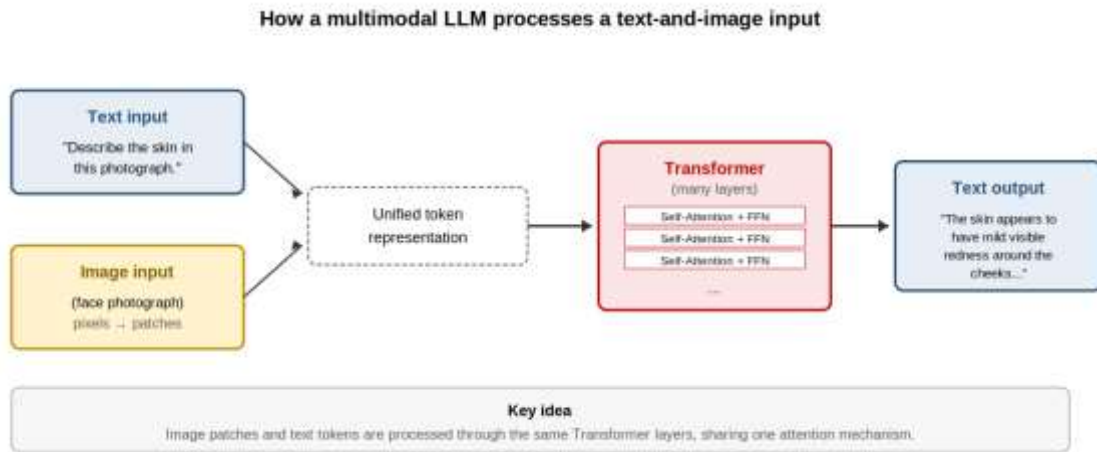


Fig. 6: How a multimodal large language model processes a text-and-image input.

Despite their impressive abilities, LLMs have a property that must be understood carefully before they are used in any application that affects people's health or safety. They generate text by predicting the most probable continuation of the input, and they do not have a reliable internal mechanism to distinguish between statements they have strong evidence for and statements they have invented. When a model produces a fluent but factually incorrect statement, the phenomenon is known as hallucination [9]. A model will state a guess in the same confident tone it uses for a fact, and a reader who is not an expert in the subject may have no way of telling the two apart. This is a documented and widely studied property of LLMs, not a defect of any particular model, and it is one of the main reasons why the system presented in this thesis does not rely on the language model alone for any decision that affects the user's safety. The detailed handling of this issue is described in Chapter 3.

1.2.5 Artificial Intelligence in Healthcare

Artificial intelligence has been applied to a wide range of problems in healthcare, including the analysis of medical images, the prediction of patient outcomes, the discovery of new drugs, and the support of clinical decision making. Dermatology, in particular, has been one of the most active areas of application, for a straightforward reason: a large part of dermatological assessment begins with looking at the skin, and a skin image is exactly the kind of data on which a deep learning system can be trained.

A widely cited study by Esteva et al., published in *Nature* in 2017, demonstrated that a deep convolutional neural network trained on 129,450 clinical images of skin lesions could classify those lesions at a level of accuracy comparable to board-certified dermatologists on a set of benchmark tasks [10]. This result was significant, and it helped to establish deep learning as a serious tool for dermatological image analysis. However, it also raised an issue that has since become well known in the field: the data on which such models are trained is not always representative of the full range of human skin.

Most publicly available dermatological image datasets contain a far greater proportion of lighter skin tones than darker ones. A model trained predominantly on light skin will tend to perform better on light skin and worse on dark skin, simply because it has seen less of the latter during training. This issue was studied in detail by Groh et al., who annotated 16,577 clinical images with Fitzpatrick skin type labels and showed that a deep neural network trained on this dataset was most accurate on skin types similar to those it had seen most often during training [11]. The Fitzpatrick scale is a six-point classification of skin phototype, originally developed to describe the way different skin reacts to sun exposure, and it is now commonly used in computer vision research to evaluate whether a model performs fairly across skin tones [11]. Table 3 summarises a few notable studies on the application of deep learning to dermatology.

Table 3: Selected studies on the application of deep learning in dermatology.

Study	Year	Task	Dataset	Result
Esteva et al.	2017	Skin lesion classification (benign vs malignant)	129,450 clinical images	Performance comparable to 21 board-certified dermatologists
Haenssle et al.	2018	Melanoma classification	ISIC dermoscopy images	CNN outperformed 58 dermatologists on test set

Study	Year	Task	Dataset	Result
Tschandl et al.	2019	Pigmented lesion diagnosis	HAM10000	CNN matched or exceeded average dermatologist
Groh et al.	2021	Skin condition classification across skin tones	Fitzpatrick17k (16,577 images)	Accuracy varied substantially across Fitzpatrick types

The relevance of this work to the present thesis is direct. A skincare guidance system that works well on some skin types and poorly on others would not only be unfair, but also unreliable in its core function. For this reason, the system presented in this thesis is evaluated on the Fitzpatrick17k dataset, and the evaluation is reported across the full range of skin types rather than as a single overall figure. The details of this evaluation, and the limitations the system still has, are given in Chapter 3.

1.3 Recommendation Systems

1.3.1 Definition and Purpose

A recommendation system, also known as a recommender system, is a software tool that suggests items to a user from a potentially large catalogue, based on what the system knows about the user, the items, or both [12]. Recommendation systems are widely used in everyday digital services: they suggest films on streaming platforms, products on online shops, and songs on music services, among many other examples. Their purpose is to help users find items that are relevant to them without having to search through the entire catalogue manually.

The relevance of this concept to the present thesis is direct. The system developed in this work takes a description of a user's skin, in the form of a photograph and a short questionnaire, and returns a set of skincare products selected from a real catalogue. It is, in other words, a recommendation system for skincare products, even though it also performs additional tasks such as image quality checking and safety validation. Understanding the

general principles of recommendation systems is therefore necessary background for understanding the design choices made in Chapter 3. The general structure of a recommendation system is shown in Fig. 7.

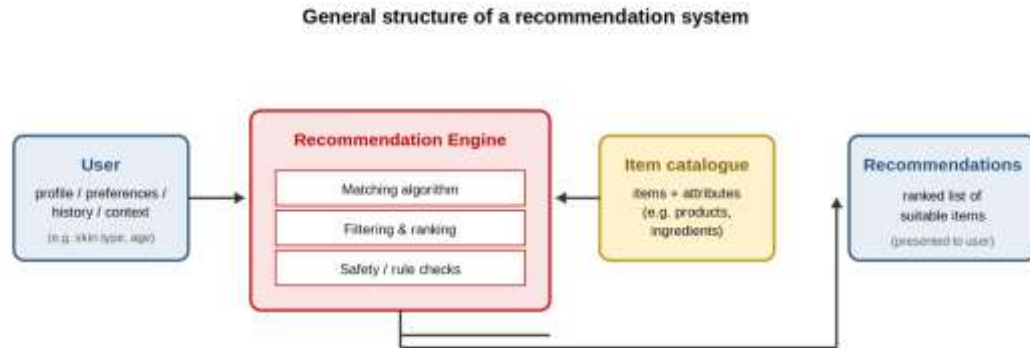


Fig. 7: General structure of a recommendation system.

1.3.2 Main Types of Recommendation Systems

Recommendation systems are usually classified into a small number of families, according to the kind of information they rely on to produce their recommendations [12]. The two most common families, content-based filtering and collaborative filtering, are illustrated side by side in Fig. 8.

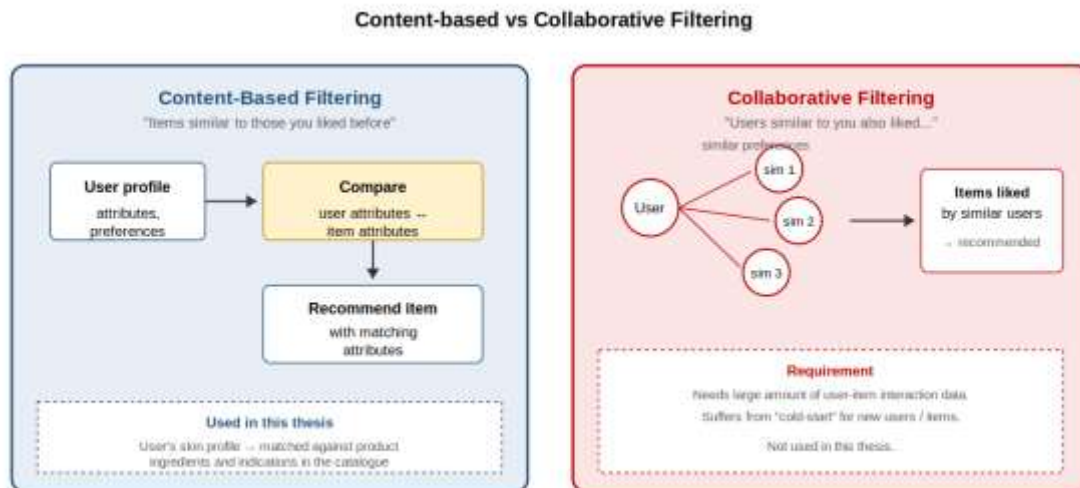


Fig. 8: Comparison between content-based filtering and collaborative filtering.

Content-based filtering recommends items to a user based on the features of the items themselves and the user's previous preferences for items with similar features [13]. For example, a content-based film recommender might note that a user has enjoyed science fiction

films directed by a particular director, and then recommend other science fiction films by the same director. Content-based filtering is particularly relevant to the system presented in this thesis, because the system matches the user's skin profile (their skin type, sensitivities, and concerns) against the attributes of products in the catalogue (ingredients, indications, contraindications) in order to decide which products to recommend.

Collaborative filtering takes a different approach. Instead of analysing the features of items, it identifies users who have shown similar preferences in the past and recommends items that those similar users have liked [12]. The classic phrase associated with collaborative filtering is "users who liked this also liked..." Collaborative filtering is the technique behind many large-scale commercial recommenders, but it requires a substantial amount of data about user interactions, which is not always available, particularly for new users or new items. This problem is known as the cold-start problem.

Hybrid systems combine content-based and collaborative methods to take advantage of the strengths of each and reduce their individual weaknesses [12]. Many real-world recommendation systems are hybrid in some form. A fourth family, knowledge-based or rule-based systems, relies on explicit rules or constraints, written or curated by domain experts, rather than on patterns learned from data [12]. A knowledge-based skincare recommender, for example, might encode a rule stating that products containing two strong active ingredients should not be recommended together to a user with a compromised skin barrier. Knowledge-based methods are particularly important when the domain involves safety considerations that cannot be left to a statistical model. As Chapter 3 will explain, the system presented in this thesis combines a content-based core with explicit, rule-based safety logic written in deterministic code, precisely because some decisions in a health-adjacent domain must not be entrusted to a learned model. Table 4 summarises the four families.

Table 4: Comparison of the main families of recommendation systems.

Approach	How it works	Strengths	Weaknesses
Content-based	Matches item attributes to user profile	Works for new users with a profile; transparent	Limited diversity; needs rich item attributes

Approach	How it works	Strengths	Weaknesses
Collaborative filtering	Uses preferences of similar users	Discovers unexpected items; no need for item features	Cold-start problem; needs large interaction data
Hybrid	Combines content-based and collaborative methods	Mitigates weaknesses of each individually	More complex to design and tune
Knowledge-based / rule-based	Uses explicit rules from experts	Predictable; safety rules can be enforced	Rules must be maintained by humans

1.3.3 Recommendation Systems in Health and Skincare

Recommendation systems have been applied to a number of health and wellness contexts, including nutrition, fitness, medication adherence, and skincare. In the skincare context specifically, several commercial mobile applications now offer some form of automated routine recommendation, usually based on a short questionnaire and sometimes on a self-uploaded photograph. The depth of personalisation in these applications varies widely. Some offer little more than a fixed routine based on broad skin-type categories, while others use machine learning models to analyse user images.

One feature is largely absent from most existing skincare recommendation tools: explicit safety logic. The handling of conflicts between active ingredients, the recognition of conditions that warrant referral to a professional, and the enforcement of contraindications are not consistently implemented. The system presented in this thesis aims to address this gap by separating the recommendation task itself, which is handled by a combination of language-model perception and content-based matching, from the safety-critical decisions, which are handled by explicit rule-based logic in deterministic code. Table 5 summarises the main approaches taken by existing tools and how the system presented in this thesis differs from them.

Table 5: Approaches to skincare recommendation in existing tools, compared with the system proposed in this thesis.

Approach in existing tools	Personalisation depth	Safety logic	Typical input
Questionnaire-based fixed routine	Low (broad skin-type buckets)	Generally absent	Short multiple-choice form
Image-based skin analysis	Medium (per-image features)	Inconsistently implemented	Self-uploaded photograph
Hybrid (questionnaire + image)	Medium to high	Rarely explicit	Photograph + short form
System proposed in this thesis	High (multi-agent profile)	Explicit, rule-based, deterministic	Photograph + structured form

1.3.4 Evaluation of Recommendation Systems

Recommendation systems are typically evaluated using metrics that measure how well the recommended items match what the user would actually choose or rate highly [12]. The most common metrics are precision, which measures the proportion of recommended items that are relevant; recall, which measures the proportion of relevant items that are recommended; and coverage, which measures the proportion of the catalogue that the system is able to recommend.

These metrics, while useful, do not capture every aspect of the quality of a recommendation in a health-adjacent context. A skincare recommender that frequently recommends products which are statistically popular but unsuitable for the specific user's skin might still score well on precision and recall, while being unsafe in practice. For this reason, the evaluation reported in Chapter 3 of this thesis does not rely on standard recommender-system metrics alone, but also includes specific tests of safety logic, conflict detection, and behaviour across different Fitzpatrick skin types.

1.4 Conclusion

This chapter has introduced the technical background on which the rest of this thesis depends. It started with a general definition of artificial intelligence and a brief history of the

field, from the Turing test and the Dartmouth workshop to the recent rise of large multimodal models. It then narrowed to machine learning, the branch of AI concerned with systems that learn from data, and to deep learning, the family of methods that uses neural networks with many layers and that has driven most of the visible progress of the last decade. It then introduced large language models and multimodal models, the most recent generation of deep learning systems, and described both their main capabilities and the central limitation, hallucination, that any responsible application built on them must take into account.

The chapter then turned to recommendation systems, the family of techniques to which the system developed in this thesis ultimately belongs. The main types of recommendation systems were described, with particular attention to content-based filtering and knowledge-based methods, which are the two approaches most relevant to the design presented in Chapter 3. Standard evaluation metrics for recommendation systems were also briefly discussed, with a note that they do not capture every aspect of quality in a health-adjacent context.

With this background established, the next chapter turns from the technology to the domain, introducing the skincare industry, the consumer-information gap that motivates this work, and the existing landscape of related work in automated skincare guidance.

Chapter 2: Skincare and Related Work

2.1 Introduction

The previous chapter introduced the technical background on which this work is built: artificial intelligence, machine learning, deep learning, large multimodal models, and recommendation systems. This chapter turns from the technology to the domain. The system presented in this thesis does not exist in a vacuum: it addresses a specific need that arises from the way the skincare industry operates today, and it sits in a research landscape where other groups have approached related problems in different ways. Both of these contexts have to be established before the system itself can be described in Chapter 3.

The chapter is organised into two main parts. Section 2.2 describes the skincare domain. It begins with the industry itself, then moves to the dermatological background needed to understand the system, including the anatomy of human skin, the Fitzpatrick classification, the most common skin conditions, and the active ingredients used in modern skincare products. Then it describes the practical problem that motivates this work, the gap between the complexity of modern skincare and the guidance that consumers actually have access to, and it surveys the channels through which skincare guidance is currently delivered.

Section 2.3 then turns to related work. It reviews two different lines of research and practice that touch on the same problem space: the academic literature on artificial intelligence in dermatology, which has so far focused largely on the diagnosis of skin diseases from clinical images, and the much larger and less documented landscape of commercial skincare applications, which offer some form of automated personal guidance but rarely include explicit safety reasoning. The section closes with a synthesis that identifies the specific gap this thesis attempts to fill.

Section 2.4 closes the chapter with a brief summary and a transition to Chapter 3, in which the system itself is presented.

2.2 The Skincare Domain

2.2.1 The Global Skincare Industry

Skincare is one of the largest segments of the global personal care industry. Estimates of the precise size of the market vary considerably between research organisations, depending on

which product categories are counted as skincare and which are placed under broader categories such as cosmetics or personal care. Recent industry reports place the global skincare market in 2024 at somewhere between 115 billion and 185 billion USD, with annual growth rates in the range of roughly 4% to 8% [14], [15], [16]. The wide range itself is informative, since it reflects the fact that the boundaries of the category are not standardised, but in every published estimate the market is large and growing.

Modern skincare products are usually grouped into a small number of functional categories. Cleansers remove dirt, oil and residue from the skin surface. Toners are liquid preparations applied after cleansing, with effects ranging from re-balancing the skin's pH to delivering mild active ingredients. Moisturisers are intended to maintain the hydration of the epidermis and support the skin barrier. Sunscreens, applied during the day, protect the skin against ultraviolet radiation. Treatments and serums are higher-concentration products designed to deliver specific active ingredients, such as those discussed in Subsection 2.2.5, and are typically used to address a very particular skin concern. A given consumer routine is normally a sequence of products from several of these categories, applied in a defined order.

Three broad trends shape the modern skincare industry. The first is the shift, over roughly the last fifteen years, from simple two-step or three-step routines toward complex regimens involving multiple active ingredients applied in a particular order. Products that were once available only through dermatologists, such as retinoids and alpha-hydroxy acids, are now sold over the counter in many countries, and consumers are increasingly expected to choose and combine these ingredients themselves. The second is the rise of online and social-media-driven product discovery, which has increased the visibility of niche brands and ingredient-led marketing, but has also created an environment in which advice of widely varying quality is published alongside products [14]. The third is the growing interest among the major industry players in personalisation, often delivered through some form of digital diagnostic tool. Several of the largest global cosmetics companies have, in recent years, deployed in-store kiosks or mobile applications that use computer vision to analyse a customer's skin and propose products from the company's catalogue [16]. The system presented in this thesis sits in this last category, though, as Chapter 3 will explain, it differs from existing commercial offerings in several important ways.

Table 6 summarises a selection of recent industry estimates for the global skincare market. The differences between sources are presented honestly, because the precise market size is

less important to this thesis than the order of magnitude, which is consistent: skincare is a large industry that is growing faster than the global economy as a whole.

Table 6: Selected industry estimates for the global skincare market.

Source	2024 estimate (USD bn)	Forecast year	Forecast value (USD bn)	CAGR
Fortune Business Insights	~115	2032	~194	~6.7%
Roots Analysis	~185	2035	~407	~7.4%
Straits Research	~171	2033	~261	~4.8%
Market Data Forecast	~193 (2025)	2034	~393	~8.2%
SkyQuest	~137 (2023)	2032	~214	~4.6%

2.2.2 Human Skin: Structure and Function

A short overview of human skin is useful before describing the conditions that skincare products address and the ingredients that act on them. The skin is the largest organ of the human body and serves several functions at once: it acts as a barrier against the external environment; it regulates body temperature, contains sensory nerve endings, and helps to maintain hydration by preventing excessive water loss [18]. From a structural point of view, the skin is conventionally divided into three main layers, illustrated in Fig. 9.

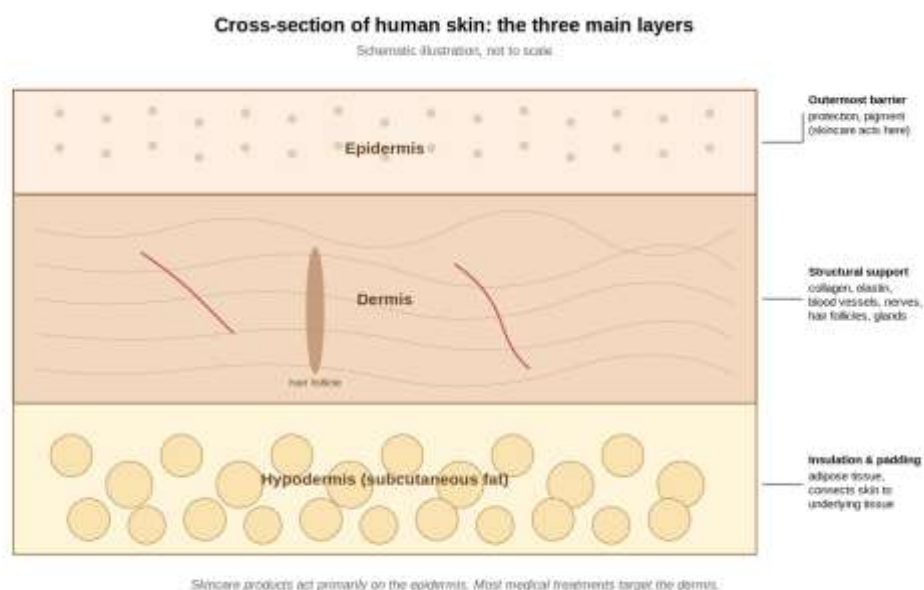


Fig. 9: Cross-section of human skin, showing the three main layers and their main functions.

The epidermis is the outermost layer of the skin. It is composed mainly of keratinocytes, the cells that produce the protein keratin, organised into several thin sublayers. The deepest part of the epidermis also contains melanocytes, the cells that produce melanin, the pigment responsible for skin and hair colour [18]. The outermost sublayer of the epidermis, called the stratum corneum, is made of flattened, dead keratinised cells and functions as the principal physical barrier of the skin. Most cosmetic skincare products act on the epidermis, and in particular on the stratum corneum.

The dermis lies underneath the epidermis and is considerably thicker. It is composed of dense connective tissue rich in collagen and elastin, the two structural proteins responsible for the mechanical strength and elasticity of the skin [18]. The dermis also contains blood vessels, lymphatic vessels, sensory nerve endings, sweat glands, sebaceous (oil-producing) glands, and the bases of hair follicles. The dermis is where most medical dermatological treatments act, and where signs of aging such as wrinkles and loss of elasticity originate.

The hypodermis, also called the subcutaneous layer, lies beneath the dermis. It is composed mainly of adipose tissue and loose connective tissue, and it serves to insulate the body, store energy, cushion underlying muscles and bones, and connect the skin to the deeper structures of the body [18]. Strictly speaking, the hypodermis is not always considered part of the skin itself, but it is functionally inseparable from the layers above.

Two further concepts will be referred to in the remainder of this chapter. The first is the skin barrier, which refers to the function of the stratum corneum and the lipid film that covers

it: together they prevent water from leaving the body and prevent harmful substances from entering. A compromised skin barrier, whether due to disease, harsh products, or environmental damage, is associated with dryness, irritation, and increased reactivity to ingredients that the skin would otherwise tolerate. The second is the classical division of skin types, distinct from the Fitzpatrick scale, into oily, dry, combination, and sensitive types, based on sebum production and reactivity. These categories are coarse but useful, and they appear on the packaging of most skincare products as an indication of the type of skin the product is intended for.

2.2.3 The Fitzpatrick Skin Phototype Classification

Beyond the structural anatomy described above, human skin varies between individuals, particularly in its colour and in the way it responds to sun exposure. The most widely used classification of these differences is the Fitzpatrick scale, introduced by the dermatologist Thomas B. Fitzpatrick of Harvard Medical School in 1975 and formally described in a 1988 paper in the *Archives of Dermatology* [17]. The scale divides skin into six types according to its reaction to ultraviolet light, as illustrated in Fig. 10. Type I corresponds to very pale skin that always burns and never tans; Type II to white skin that burns easily and tans only minimally; Type III to slightly darker white skin that burns moderately and tans uniformly; Type IV to lightly tanned or beige skin that burns minimally and tans easily; Type V to moderate brown skin that rarely burns and tans profusely; and Type VI to dark brown or black skin that does not burn and tans deeply [17]. The classification was originally developed to help dermatologists choose appropriate doses of ultraviolet light for treating psoriasis, but it has since become the standard descriptive scale used throughout dermatology and, more recently, in computer vision research on skin imagery [11].

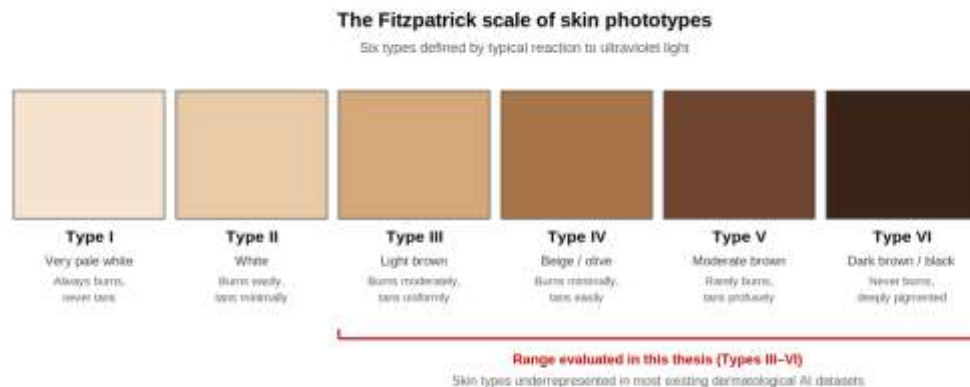


Fig. 10: The Fitzpatrick scale of skin phototypes.

The Fitzpatrick scale has well-known limitations. It was developed in the United States in the 1970s and is generally agreed to be more precise for lighter skin types than for darker ones. The boundaries between types are not always clear in practice, and the same individual may be classified differently by different observers. Despite these limitations, the scale remains the most widely used framework in both clinical dermatology and the research literature on skin AI, and it is the framework used to evaluate the system presented in this thesis. The system is evaluated specifically on Fitzpatrick types III through VI, the range covering most of the population of North Africa and the rest of the world outside northern Europe, since these are the skin types on which existing dermatological AI systems have been shown to perform least well [11].

2.2.4 Active Ingredients in Skincare

Modern skincare products list their full composition on the packaging using the International Nomenclature of Cosmetic Ingredients (INCI) system, which uses Latin and chemical names that are not generally understood by non-specialists. Many of the ingredients listed are inert and serve only to give the product its texture, stability or preservation. A smaller subset, usually referred to as "actives", has measurable biological effects on the skin and is the ingredients that most clearly distinguish one product from another. The most widely used actives, the conditions they address, and a citation to a recent review of each, are summarised in Table 8.

Table 7: Active ingredients commonly used in modern skincare products and their primary effects.

Active ingredient	Primary effect on skin	Typical use	Reference
Retinoids (retinol, tretinoin)	Promote keratinocyte turnover; increase epidermal and dermal thickness; reduce wrinkles	Photoaging, acne	[20]
Alpha-hydroxy acids (AHA: glycolic, lactic)	Chemical exfoliation; stimulate collagen synthesis; improve texture	Photoaging, hyperpigmentation, acne	[21]

Active ingredient	Primary effect on skin	Typical use	Reference
Beta-hydroxy acid (BHA: salicylic)	Oil-soluble exfoliation; penetrates pores; mild anti-inflammatory	Acne, oily skin	[21]
Vitamin C (ascorbic acid)	Antioxidant; cofactor for collagen synthesis; mild depigmenting effect	Photoaging, hyperpigmentation	[22]
Niacinamide (vitamin B3)	Strengthens skin barrier; reduces hyperpigmentation; anti-inflammatory	Hyperpigmentation, sensitive skin, rosacea	[23]
Benzoyl peroxide	Antibacterial against acne-causing bacteria; mild keratolytic	Acne	[19]

Several of these active ingredients interact with each other in ways that are not obvious from the packaging. Retinoids and alpha-hydroxy acids, both popular ingredients in anti-aging products, can each produce skin irritation on their own, and they generally should not be applied in the same routine without medical supervision [20], [21]. Vitamin C in its most common form, ascorbic acid, is most effective at a low pH that conflicts with the higher pH at which niacinamide is normally formulated, although the practical importance of this conflict in modern, well-buffered formulations is contested [22], [23]. Strong exfoliating acids can compromise the skin's natural barrier function if used too frequently, particularly when combined with retinoids [21]. None of these interactions are difficult to summarise; they are well documented in the dermatological literature. The difficulty, from the consumer's point of view, lies in identifying which ingredients are present in which products and in deciding which combinations are safe to use together. Fig. 11 illustrates some of the most commonly cited interactions.

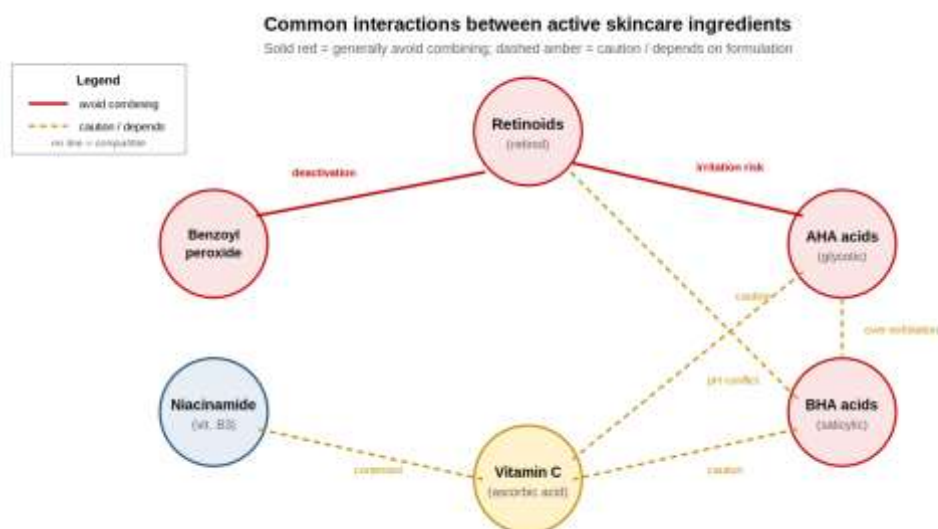


Fig. 11: Common interactions between active skincare ingredients.

2.2.5 The Consumer Information Problem

The combination of an increasingly complex product landscape, an expanding range of active ingredients, and an increasingly varied set of advice sources has produced a situation in which an ordinary consumer who wishes to assemble a skincare routine faces a non-trivial information problem. The previous subsection introduced the main active ingredients and their interactions. The point of the present subsection is to describe what happens when an ordinary person, without dermatological training, tries to apply this information to their own skin.

The information channels through which skincare questions are now most commonly answered include search engines, social media platforms, and, increasingly, general-purpose AI assistants. Each of these channels has known limitations. Search results are typically optimised for engagement rather than accuracy and frequently return contradictory advice from different sources. Social media content is heavily influenced by sponsored material, and the most visible voices are not always the most informed. General-purpose AI assistants, as discussed in Chapter 1, are subject to hallucination and can produce confident but incorrect statements about ingredient safety [9]. None of these channels provides personalised guidance that takes the user's specific skin type, condition history, and visible features into account in a structured way.

The motivation for the system presented in this thesis can be stated in these terms. The technical components needed to give a non-specialist user accurate, personalised, and safety-checked skincare guidance now exist, but they have not been assembled into a single system

that the user can actually consult. The remainder of this chapter examines how existing systems, both academic and commercial, have approached parts of this problem, and identifies the parts that have not been adequately addressed.

2.2.6 How Skincare Guidance Is Currently Delivered

Five main channels currently deliver some form of skincare guidance to consumers, with very different levels of reliability and accessibility. Understanding these channels is necessary for understanding what the system presented in this thesis can and cannot replace. Fig. 12 summarises the five channels in terms of three dimensions: reliability of the advice, accessibility to the average consumer, and depth of personalisation.



Fig. 12: The five channels through which skincare guidance is currently delivered, compared on reliability, accessibility, and personalisation.

The first channel is the dermatologist. A consultation with a board-certified dermatologist remains the gold standard for personalised skincare guidance. A dermatologist examines the skin directly, takes a full medical history, has the training to recognise both skin conditions and broader medical issues that manifest on the skin, and can prescribe medical-grade treatments that are not available over the counter. The limitation is one of access. Dermatologists are scarce and expensive in many countries, and even in countries where they are more available, a routine consultation about cosmetic skincare is generally not covered by public health systems. For most people in the world, regular dermatologist consultations for non-medical skincare questions are not realistic.

The second channel is the pharmacist or trained skincare consultant, often found in pharmacies and dedicated skincare stores. The advice available through this channel varies widely with the training of the individual, but it is generally limited to the products sold in the particular store, and the consultant has typically had no opportunity to observe the customer's skin over time.

The third channel is online media in its various forms: search engines, dedicated skincare websites, forums, and the social media accounts of professionals and laypeople. The volume of information available through this channel is enormous and growing, but as discussed in the previous subsection, the quality is highly variable and the user is left to evaluate the reliability of each source individually.

The fourth channel is the in-store digital kiosk or mobile application offered by some of the larger skincare brands. These tools usually take a short questionnaire and sometimes a photograph, and propose products from the brand's own catalogue. They are convenient and accessible, but they are limited to the products of a single brand, and the recommendations are visibly aligned with the commercial interests of the brand that operates the tool.

The fifth and most recent channel is the general-purpose AI assistant, which can be asked questions in natural language and will produce confident-sounding answers about any ingredient, product, or routine. As discussed in Chapter 1, this channel is subject to hallucination and does not have a reliable mechanism for verifying its own claims [9]. The system presented in this thesis can be seen, in part, as an attempt to provide the natural-language convenience of this channel together with the safety and reliability that a general-purpose assistant cannot offer on its own.

2.3 Related Work

2.3.1 Artificial Intelligence in Dermatology: Research Systems

The academic literature on artificial intelligence applied to dermatology is large and growing rapidly. The work that has received the most attention has focused on the diagnosis of skin diseases from clinical or dermoscopic images, using deep convolutional neural networks of the kind described in Chapter 1. A small number of representative studies are summarised here to characterise the state of this research and to clarify how it differs from the problem addressed by this thesis. Table 9 summarises these studies.

Table 8: Selected studies in the academic literature on artificial intelligence in dermatology.

Study	Year	Task	Dataset	Main outcome
Esteva et al. [10]	2017	Skin lesion classification (benign vs malignant)	129,450 clinical images	Performance comparable to 21 dermatologists
Haenssle et al. [24]	2018	Melanoma classification on dermoscopy	100-image test set, 58 dermatologists	CNN outperformed average dermatologist
Tschandl et al. [25]	2018	Release of public benchmark dataset	10,015 dermoscopic images (HAM10000)	Enabled wide reproducible research
Tschandl et al. [26]	2020	Human-AI collaboration for diagnosis	Multiple dermoscopic datasets	AI assistance improved accuracy, most for less experienced clinicians
Groh et al. [11]	2021	Classification across skin types	Fitzpatrick17k (16,577 images)	Significant accuracy gap across Fitzpatrick types

The work of Esteva et al., published in *Nature* in 2017, has already been discussed in Chapter 1 [10]. To briefly recall, the authors trained a deep convolutional neural network on 129,450 clinical images of skin lesions and demonstrated that, on a benchmark set of classification tasks, the network performed at a level comparable to that of board-certified dermatologists. The result is generally regarded as one of the first convincing demonstrations that deep learning could match expert clinical performance in a medical imaging task.

A complementary result was published the following year by Haenssle et al. in the *Annals of Oncology* [24]. The authors trained a Google Inception v4 convolutional neural network on dermoscopic images of melanocytic lesions and evaluated it on a 100-image test set against 58 dermatologists from 17 countries. The network achieved sensitivity and specificity values that, on this test set, were equal to or better than the average performance of the human dermatologists. The study was widely reported as a further demonstration that, for narrowly defined image-classification tasks in dermatology, deep learning could match or exceed average expert performance.

A practical limitation of much of this early work was the small size and limited diversity of the publicly available dermatological image datasets, which made it difficult to compare different methods on a common basis. This was directly addressed by the work of Tschandl, Rosendahl, and Kittler, who in 2018 released the HAM10000 dataset, a collection of 10,015 dermoscopic images of common pigmented skin lesions drawn from multiple sources [25]. The dataset was made publicly available through the ISIC archive and has since become one of the most widely used benchmarks in the field. A subsequent study by Tschandl et al., published in *Nature Medicine* in 2020, examined the practical implications of AI-assisted dermatological diagnosis [26]. The authors found that providing good-quality AI-based support to clinicians improved diagnostic accuracy over either the AI or the physician alone, and that the largest gains were observed for the least experienced clinicians. This result is important because it shifts the framing of the research question, from whether the AI can replace the dermatologist to how the AI can be most effectively combined with one.

A separate line of research, increasingly important in the last few years, has focused on the fairness of these systems across skin types. The study of Groh et al., already discussed in Chapter 1, annotated 16,577 clinical images with Fitzpatrick skin type labels and showed that a deep neural network trained on this dataset was significantly more accurate on the skin types best represented in the training data, which were the lighter Fitzpatrick types [11]. The general finding, replicated by several subsequent studies, is that dermatological AI systems trained on datasets dominated by lighter skin tend to perform worse on darker skin, with implications for both fairness and clinical safety.

A characteristic of this research literature, taken as a whole, is its focus on diagnosis rather than guidance. The task that these systems are designed to perform is to look at an image of a skin lesion and produce a probability that the lesion belongs to a particular diagnostic category. They are not designed to produce a skincare routine, to recommend products, or to

reason about ingredient interactions. The task addressed by this thesis is therefore not the same as the task addressed by most of the existing AI-in-dermatology literature, even though it shares many of the same technical foundations.

2.3.2 Commercial Skincare AI Applications

Alongside the academic literature, a substantial number of commercial mobile applications and in-store tools now claim to use artificial intelligence to provide personal skincare guidance. The technical details of these systems are generally not publicly documented, which is a limitation that has to be acknowledged at the outset of any review. What is publicly known about them comes from their consumer-facing descriptions, from press materials, and, in some cases, from independent journalism. The discussion below is therefore necessarily approximate.

The simplest commercial offerings consist of a questionnaire that asks the user about their skin type, age, climate, and main concerns, and produces a recommended routine assembled from a fixed catalogue of products. The personalisation here is shallow: it amounts to bucketing the user into one of a small number of pre-defined categories and returning the routine associated with that category.

A second category of tools accepts a self-uploaded photograph in addition to the questionnaire. These tools typically apply some form of image analysis to extract visible features such as the presence of redness, the apparent pore size, or the presence of acne lesions, and combine these features with the questionnaire responses to produce a recommendation. The image analysis is, in most cases, performed by a computer vision model whose architecture and training data are not publicly disclosed. Several of the largest global cosmetics companies operate tools of this kind, sometimes branded under names that reference artificial intelligence or skin diagnosis [16].

A third category, which has emerged in the last two years, integrates large language models into the recommendation process. These tools typically use a multimodal model to generate a natural-language description of the user's skin from a photograph, and then use the same or a different model to produce written recommendations. The convenience of this approach is considerable, but as discussed in Chapter 1, it is also exposed to the hallucination problem [9], and there is no public documentation that any of these tools currently mitigate that problem in a systematic way.

The features that distinguish the system presented in this thesis from these commercial offerings are summarised in Table 10 and discussed in the synthesis subsection below.

Table 9: Categories of commercial AI-based skincare tools, compared with the system proposed in this thesis.

Category of tool	Typical input	Typical mechanism	Limitations
Questionnaire-only	Short multiple-choice form	User bucketed into pre-defined skin type; fixed routine returned	Shallow personalisation
Questionnaire + image	Form + self-uploaded photo	Computer vision extracts features (redness, pores, acne) and combines with form	Model architecture and training data undisclosed
LLM-based natural-language tool	Photo + free-text questions	Multimodal LLM generates description and recommendations	Subject to hallucination; no public safety logic
System proposed in this thesis	Photo + structured form	Multimodal LLM for perception + content-based matching + explicit rule-based safety	Evaluated on Fitzpatrick17k; documented limits in Chapter 3

2.4 Conclusion

This chapter has provided the domain background needed to understand the system presented in the next chapter. It introduced the global skincare industry and its main product categories, summarised the anatomy of human skin and the Fitzpatrick classification of skin phototypes, described the most common visible skin conditions that consumer skincare addresses, and reviewed the main families of active ingredients used in modern products together with their most important interactions. It then characterised the consumer information problem that motivates this work and surveyed the five channels through which skincare guidance is currently delivered. The second half of the chapter reviewed two relevant lines of related work, namely academic research on AI in dermatology and commercial AI-

based skincare applications, and identified three specific gaps: the focus of academic work on diagnosis rather than guidance; the absence of explicit safety logic in most commercial systems; and the absence of evaluation across the full range of skin types in either body of work.

The next chapter turns from this background to the system itself. It describes the architecture of the proposed system, the dataset used to evaluate it, the implementation, and the experimental results, with particular attention to the three design choices identified at the end of the previous section.

Chapter 3: Implementation and Results

3.1 Introduction

The two preceding chapters established the ground on which this work stands. Chapter 1 set out the technical background that the system relies on: artificial intelligence, machine learning, deep learning, large multimodal models, and recommendation systems. Chapter 2 turned to the domain, describing the skincare industry, the dermatological background needed to reason about skin, and the consumer information problem that motivates the work, before surveying the related work in both academic dermatological AI and commercial skincare applications. That survey closed, in Subsection 2.3.3, by identifying three gaps that together define the contribution of this thesis: the academic literature concentrates on diagnosis rather than guidance; the commercial applications rarely expose any explicit, auditable safety logic; and neither body of work reports performance across the full range of skin types.

This chapter presents the system that was built to address those gaps, together with the experiments carried out to measure how well it does so. The system is a guidance system rather than a diagnostic one: its purpose is not to name a disease but to assemble a safe, personalised skincare routine from a real product catalogue. It is organised as a pipeline of seven cooperating agents, each with a single, well-defined responsibility, through which a user's photograph and questionnaire are progressively turned into a structured skin profile and then into a set of justified product recommendations.

Two design decisions run through the whole of that pipeline and are returned to repeatedly in this chapter. The first is a seed-then-overwrite treatment of perception: an initial vision pass produces a first, provisional description of the skin, which a second pass then refines in the light of the information the user has supplied about themselves. The second, and the more important of the two, is a strict separation between the learned components of the system and its deterministic components. Perception, refinement, and the generation of natural-language explanations are delegated to a learned model; the safety-critical decisions, such as the avoidance of incompatible active ingredients and the recognition of conditions that warrant referral to a professional, are taken by explicit, rule-based logic that never passes through that model. This separation is the design choice that Subsection 2.3.3 identified as the contribution of the thesis, and Section 3.4 describes how it is realised in practice.

The chapter is organised as follows. Section 3.2 describes the tools and technologies used to build the system. Section 3.3 describes the dataset on which the system is evaluated, and makes explicit that this dataset is a labelled test set used to measure a frozen, pretrained model rather than training data, since no model was trained in the course of this work. Section 3.4 presents the implemented system in detail, agent by agent, and follows a single request through the pipeline from end to end. Section 3.5 reports the experimental results, covering both the accuracy of the system's perception and the behaviour of its deterministic safety layer. Section 3.6 concludes the chapter and leads into the general conclusion of the thesis.

3.2 Tools

3.2.1 Overview of the Technology Stack

The system is implemented as a client-server web application. A browser-based interface, built with React and the Vite build tool, handles the interaction with the user: the capture of photographs, the skin questionnaire, the progress display, and the presentation of the final report. A Python backend, built on the FastAPI framework, orchestrates the agent pipeline, exposes it through a small set of REST endpoints, and enforces the structured data contract that the agents exchange. The perception and language tasks within the pipeline are delegated to GPT-4o, a pretrained multimodal model, which is used without any task-specific training. The product catalogue against which recommendations are made is held as a structured file that the backend reads at startup. Table 10 summarises the role of each component, and Figure 13 shows how they are connected.

Table 10: Principal components of the implemented system and their roles.

Component	Technology	Role in the system
User interface	React, Vite	Photo capture, questionnaire, progress display, report rendering
Backend and API	Python, FastAPI	Orchestrates the agent pipeline; exposes REST endpoints; enforces the data contract
Perception and reasoning	GPT-4o (multimodal, zero-shot)	Two vision passes (seed and refinement) and natural-language generation
Product catalogue	Structured data file	Source of candidate products and their attributes

3.2.2 Backend: Python and FastAPI

The backend is written in Python and built on FastAPI, a modern web framework for building APIs on top of standard Python type hints [27]. FastAPI was chosen for three properties that matter directly to this system. First, it is asynchronous, which allows the backend to wait on the relatively slow calls to the multimodal model without blocking other requests. Second, it uses Pydantic to validate data against declared models, which is what allows the data contract exchanged between the agents (described in Section 3.4.2) to be enforced rather than merely hoped for: a malformed object produced by one agent is rejected before it can reach the next. Third, it generates interactive API documentation automatically, which was useful during development for inspecting the pipeline at each stage. The development server is run through Uvicorn, the ASGI server with which FastAPI is normally deployed.

3.2.3 Frontend: React and Vite

The user interface is a single-page application written in React, an open-source JavaScript library for building user interfaces from reusable components [28]. The component model maps naturally onto the structure of the interface, in which each stage of the user's journey, the photo-capture panel, each question of the skin quiz, the progress display, and the final report, is a self-contained component with its own state. The project is built and served with Vite, a frontend build tool that provides a fast development server over native ES modules and an optimised production build [29]. The interface supports both English and French, in keeping with the bilingual product catalogue described in Section 3.4.7.

3.2.4 The Multimodal Model: GPT-4o

The perceptual and linguistic work of the pipeline is carried out by GPT-4o, the multimodal model released by OpenAI in May 2024, which accepts any combination of text and image as input and produces text as output [8]. GPT-4o is used in this system in two distinct ways. As a vision model, it examines the user's photograph and returns a structured description of the visible characteristics of the skin; this is done twice, once to seed the description and once to refine it, as described in Sections 3.4.4 and 3.4.5. As a language model, it generates the natural-language explanation that accompanies the recommendations, described in Section 3.4.9.

One point about the use of GPT-4o is central to the rest of the chapter. The model is used exactly as supplied, in a zero-shot manner: it is not fine-tuned, and no weights are changed at any point in this work. Where the model is required to return structured data, it is instructed to emit JSON conforming to a fixed schema, which the backend then validates. The consequence for the evaluation, taken up in Section 3.3, is that what is being measured is the behaviour of a frozen, pretrained model placed inside a particular system, not the behaviour of a model that has been trained on the evaluation domain.

3.3 Dataset

3.3.1 A Test Set, Not Training Data

The dataset described in this section serves a single purpose: to measure the behaviour of the system, and in particular of its perceptual components, against a known ground truth. It is important to be clear from the outset about what kind of dataset this is. No model is trained, fine-tuned, or otherwise adjusted anywhere in this work. The multimodal model at the centre of the pipeline, GPT-4o, is used exactly as supplied by its provider, in a zero-shot setting (Section 3.2.4). The dataset is therefore a labelled test set, used to evaluate a frozen, pretrained model, and not a training set. This distinction governs the interpretation of every result in Section 3.5: the figures reported there describe how a general-purpose model performs when placed inside this particular system and prompted in a particular way, not how a model that has learned the evaluation domain performs on held-out examples of it.

A consequence of this design is that the dataset does not need to be large in the sense that a training set must be. Its role is to provide enough labelled, representative examples to estimate the system's accuracy and, just as importantly, to expose the conditions under which that accuracy degrades. The remainder of this section describes where the images come from, what was labelled, and how the evaluation was carried out.

3.3.2 Source of the Images

The images are facial photographs collected to represent the consumer-photograph setting that the system targets: front-facing portraits taken under varied natural lighting, without clinical controls. The collection was restricted to Fitzpatrick types III to VI from the outset, reflecting the population the system is intended to serve and concentrating the evaluation on the darker skin types that are most often under-represented in dermatological datasets and least often reported on by the commercial tools surveyed in Chapter 2. Each image was hand-

labelled for the two attributes described below; no existing label from any external dataset was used.

3.3.3 The Two Ground-Truth Attributes

Each image in the test set was hand-labelled for two attributes. The first is the Fitzpatrick skin type, recorded on the four-point scale from III to VI. The second is the presence or absence of each of eight visible skin concerns, drawn from a fixed taxonomy. The two attributes correspond directly to the two things the system's perception must get right: an estimate of skin tone, which conditions much of the downstream reasoning about sun sensitivity and pigmentation, and a reading of which visible concerns are actually present, which drives the selection of products. A concern was recorded as present only where it was visible in the photograph itself, a point that becomes important when interpreting the results for concerns whose physical signature is difficult to capture photographically.

3.3.4 The Concern Taxonomy

The eight-category concern taxonomy is the same vocabulary that the system uses internally and that the user is asked about in the final question of the skin questionnaire (Section 3.4.10). Table 11 lists the eight categories together with the criterion used to decide, during labelling, whether each was present. The taxonomy deliberately mixes concerns of different kinds: some are textural (oiliness, enlarged pores), some are lesional (active acne), and some are pigmentary (post-inflammatory hyperpigmentation, dark spots, uneven tone). This mixture is representative of the concerns a real user brings to a skincare consultation, but it also means that the concerns are not equally easy to detect from a single photograph, which the results in Section 3.5 reflect.

Table 11: The eight-category concern taxonomy and the criterion for judging each present.

Concern	Criterion for presence
Active acne	Visible inflammatory lesions (papules, pustules) at the time of the photograph
Acne marks / PIH	Flat, darkened marks left by resolved acne lesions
Oiliness	Visible surface shine or greasy sheen across the T-zone or face
Dark spots	Discrete areas of concentrated pigmentation darker than the surrounding skin

Concern	Criterion for presence
Enlarged pores	Visibly dilated pores, typically across the nose and cheeks
Fine lines	Visible fine creasing, most often around the eyes or on the forehead
Dryness	Visible flaking, rough texture, or dull, dehydrated-looking skin
Uneven tone	Diffuse, non-localised variation in skin colour across the face

3.3.5 Composition of the Test Set

The test set spans the four Fitzpatrick types from III to VI, with deliberate emphasis on the darker types so that the evaluation is not dominated by the lighter end of the range. The eight concerns are not present in equal numbers across the images; as in any naturally occurring set of faces, some concerns, such as oiliness and active acne, appear frequently, while others, such as uneven tone assessed as a diffuse whole-face property, appear comparatively rarely. This imbalance is not a defect of the set so much as a property of the phenomenon, but it does mean that the per-concern results in Section 3.5 are estimated more reliably for the common concerns than for the rare ones, and the discussion there is careful to say so.

3.3.6 Evaluation Protocol

The evaluation proceeds image by image. Each photograph is passed through the perceptual stages of the pipeline exactly as it would be for a real user, and the structured description that the system produces is compared, attribute by attribute, against the hand-applied ground truth. For the Fitzpatrick attribute, the comparison yields two measures: an exact-match accuracy, which counts a prediction as correct only if it names the same type as the ground truth, and a within-one-type accuracy, which counts a prediction as correct if it falls within one step of the ground truth on the Fitzpatrick scale. The second measure is reported because the Fitzpatrick scale is ordinal and known to be noisy even among expert human raters, so that a prediction one step away is a qualitatively different kind of error from one several steps away. For each of the eight concerns, the comparison is treated as a binary detection problem and summarised with precision, recall, and the F1 score, defined in Section 3.5.1.

One practical limitation of the evaluation should be recorded here rather than left implicit. The perceptual stages depend on calls to a hosted model endpoint that is subject to rate limiting. A number of requests did not return a usable response within the evaluation run

because of this rate limiting, and the affected images were excluded from the analysis rather than retried under different conditions, so that every reported figure is computed only over images for which the full pipeline completed. This is a limitation of the measurement apparatus, not of the system's design, and Section 3.5.7 returns to its implications.

3.4 Implemented System

3.4.1 Architecture Overview

The system is organised as a pipeline of seven agents, shown in Figure 13. A request enters at the top as a face photograph together with the answers to a short questionnaire, and leaves at the bottom as a personalised routine accompanied by a natural-language explanation. Between those two points, each agent performs one well-defined transformation and hands its output to the next through a shared, validated data structure described in Section 3.4.2.

The first idea embodied in the pipeline is the seed-then-overwrite treatment of perception. Rather than asking the model to read everything from the photograph in a single pass, the system takes an initial reading (the seed, Agent 2) and then deliberately revisits it in the light of what the user has said about themselves (the overwrite, Agent 3). A first pass that sees only pixels cannot know that the user reports a retinoid allergy or describes their skin as sensitive; the second pass can, and can revise the provisional description accordingly. Separating the two passes keeps each one simple and makes the point at which user context enters the perception explicit rather than tangled into a single prompt.

The second and more important idea is the strict separation between the learned and the deterministic parts of the pipeline, indicated in Figure 13 by the two shadings. Only three stages are learned, in the sense that they delegate a judgement to GPT-4o: the two vision passes (Agents 2 and 3) and the final generation of prose (Agent 7). Everything that bears on the safety of the recommendation is deterministic. The consolidation of the profile (Agent 4), the selection of products (Agent 5), and above all the safety validation (Agent 6) are carried out by explicit, rule-based logic that never passes through the model. This is the design decision identified in Subsection 2.3.3 as the contribution of the thesis. Its consequence is that the recommendations the user receives are not produced by a model that might, on a given run, hallucinate an unsafe combination; they are produced by code whose behaviour can be read, tested, and audited, with the model confined to the tasks of perceiving and explaining.

The agents are described in turn in the following subsections, in the order in which a request encounters them. The deterministic image-quality gate is described first because it runs before any call to the model is made.

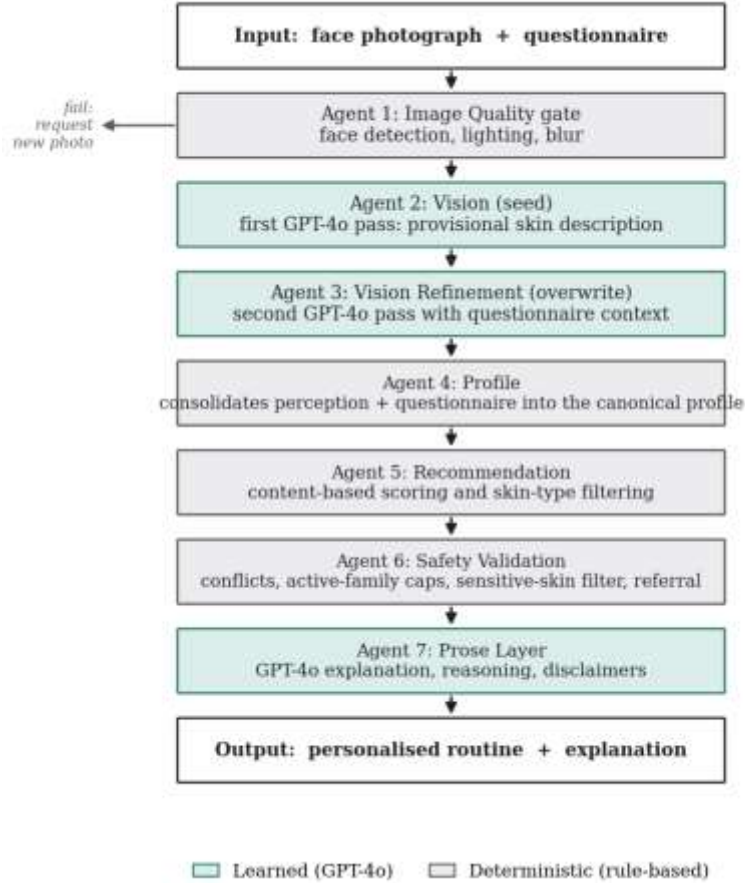


Fig. 13: The seven-agent pipeline. Learned stages (GPT-4o) are shaded; all safety-bearing stages are deterministic.

3.4.2 The Data Contract

The agents communicate through a single structured object that grows as it passes down the pipeline. Rather than each agent inventing its own format, the object conforms to a schema declared on the backend and validated by Pydantic at each boundary (Section 3.2.2), so that an agent that produces a malformed or incomplete object is detected immediately, before its output can corrupt a later stage. Table 12 lists the principal fields of this object and the agent responsible for populating each. The fields fall into three groups: the perceptual fields, written by the vision agents and describing what is seen; the questionnaire fields, supplied by the user; and the decision fields, written by the recommendation and safety agents and describing what is to be advised and why.

Table 12: Principal fields of the data contract that passes between the agents.

Field	Written by	Description
fitzpatrick_type	Agents 2, 3	Estimated skin tone on the Fitzpatrick scale (III to VI)
visible_concerns	Agents 2, 3	The subset of the eight-category taxonomy judged present
sensitivities	Agents 2, 3	Visible indications of sensitivity or barrier compromise
skin_type	Questionnaire	Self-reported skin type (oily, dry, combination, normal, sensitive)
reactivity	Questionnaire	Whether the skin reacts easily to products
age_range	Questionnaire	Self-reported age band
primary_concern	Questionnaire	The concern the user most wants addressed
ranked_concerns	Agent 4	Consolidated, ordered list of concerns driving selection
recommended_products	Agent 5	Ordered list of selected catalogue products
safety_flags	Agent 6	Conflicts removed, caps applied, referral indications
explanation	Agent 7	Natural-language justification presented to the user

3.4.3 Agent 1: The Image Quality Gate

Before any call to the model is made, the photograph passes through a deterministic quality gate whose purpose is to reject images from which no reliable reading could be taken, and to do so cheaply. The gate performs three checks in sequence. It first verifies that the image file is well-formed and that its dimensions fall within an acceptable range, ensuring the photograph is large enough for skin features to be visible. It then applies two lightweight image-statistics checks: an overall brightness measure, to reject photographs that are too dark or too washed out, and a focus measure based on the variance of the Laplacian, which falls sharply for blurred images. Finally, it checks for the presence of a face region of sufficient

size in the frame, so that a photograph of a hand or a blank wall is rejected before reaching the model. If any check fails, the pipeline halts and the interface asks the user for another photograph, as shown later in Section 3.4.10.

Placing this gate first, and implementing it deterministically, serves two ends. It protects the user from receiving confident-sounding advice based on an unreadable image, and it protects the system from spending a call to the model, and the latency and cost that accompany it, on a photograph that should never have reached the model in the first place. The gate is the first illustration of the general principle that decisions which can be made by cheap, predictable rules are not delegated to the model.

3.4.4 Agent 2: The Vision Agent (Seed)

The first perceptual pass sends the photograph to GPT-4o with an instruction to return a structured description of the visible characteristics of the skin. The model is asked for three things: an estimate of the Fitzpatrick type, the subset of the eight-category taxonomy that it judges to be present, and any visible indications of sensitivity. The model is instructed to emit its answer as JSON conforming to a fixed schema rather than as free text, so that the result can be validated and passed on without parsing prose. Listing 3.2 shows the shape of the object returned.

This first description is provisional. It is formed without any knowledge of the user beyond what the photograph shows, and it is expected to be revised. The decision to take it anyway, rather than waiting until the questionnaire is available and reading everything at once, is what makes the second pass a refinement rather than a fresh start, and is discussed next.

3.4.5 Agent 3: The Vision Refinement Agent (Overwrite)

The second pass sends the photograph to GPT-4o a second time, now together with the seed description and the user's questionnaire answers, and asks the model to refine the description in the light of that context. The questionnaire supplies information that no photograph can convey: the user's self-reported skin type, whether their skin reacts easily to products, their age band, and any known conditions or allergies. The refinement pass uses this context to correct and qualify the seed. A seed that recorded no sensitivities may be overwritten to record sensitivity when the user reports that their skin reacts easily; an estimate of fine lines may be tempered or reinforced by the user's age band; a concern the model was

uncertain about in isolation may be confirmed or dropped once the user's primary concern is known.

The name seed-then-overwrite is literal: where the two passes disagree, the refined pass takes precedence, because it is strictly better informed than the first. The output of this stage has the same schema as the seed (Listing 3.2), so that the rest of the pipeline is indifferent to how many passes produced it. This is the last point in the pipeline at which the model contributes to what the system believes about the skin; from here on, the description is treated as fixed and is reasoned over deterministically.

3.4.6 Agent 4: The Profile Agent

The profile agent consolidates the refined perceptual description and the questionnaire answers into the single canonical profile that the downstream agents act upon. This is a deterministic step: it performs no new perception and makes no call to the model. Its work is to reconcile the two sources of information into one ordered structure. In particular, it produces the ranked list of concerns that drives selection, combining the concerns the vision agents detected with the concern the user named as primary, and giving the user's stated priority appropriate weight. It also resolves the user's self-reported skin type and reactivity into the flags that the safety agent will later consult, such as a flag marking the profile as sensitive. The output is the canonical profile against which products are scored.

3.4.7 Agent 5: The Recommendation Agent

The recommendation agent selects products from the catalogue by content-based filtering, the technique introduced in Chapter 1, in which items are matched to a user by comparing the attributes of the items with the attributes of the user rather than by relying on the behaviour of other users. Here the user is represented by the canonical profile and the items are the products of the catalogue. The catalogue is held as a structured file in which each product carries nine fields; these are listed in Table 12. Four of the nine fields drive the selection logic, each playing a distinct role.

The primary match is made on the `concerns_addressed` field. Each product records the concerns it is formulated to treat, and the agent scores every product by how many of the profile's ranked concerns it addresses, awarding more weight to the concerns the user ranked highest. This score is what selection is built on: a product that treats what the skin actually has scores highly, and a product that treats nothing the user needs scores zero. The

skin_types_suitable field acts as a filter applied before scoring, so that a product is eligible only if it is suitable for the profile's skin type and products formulated for a different skin type are removed from consideration rather than merely scored low. The key_ingredients field is consulted by the safety logic described in the next subsection, and the conflicts_with field records the products and active families with which a product must not be combined. The remaining field used in the logic, price_dzd, serves only as a tie-breaker when two products are otherwise equally suitable; the descriptive and display fields play no part in selection.

Table 13: The nine fields carried by each product in the catalogue and their role in the logic.

Field	Role
name	Display
url	Display (link to the product page)
price_dzd	Tie-breaker between equally scored products
key_ingredients	Safety logic: active-family caps, sensitive-skin filter, sunscreen and barrier checks
skin_types_suitable	Eligibility filter applied before scoring
concerns_addressed	Primary match: scored against the profile's ranked concerns
description_fr	Display (French)
description_en	Display (English)
conflicts_with	Conflict avoidance in safety validation

3.4.8 Agent 6: The Safety Validation Agent

The safety validation agent is the deterministic core of the system and the stage that most distinguishes it from the commercial tools surveyed in Chapter 2. It takes the scored, filtered set of candidate products and applies a sequence of explicit rules whose effect is to guarantee that the routine finally recommended is internally consistent and safe to follow. None of these rules consults the model.

Four kinds of check are applied. A pairwise conflict check examines every pair of candidate products and removes combinations that are known to be incompatible, such as pairing a high-strength vitamin C with niacinamide, by consulting the conflicts_with field. An active-family cap ensures that the routine does not stack multiple products from the same

active family, admitting for example only one niacinamide product and only one salicylic acid product, so that the cumulative concentration of any one active stays within a safe range. A sensitive-skin filter, applied when the profile is flagged as sensitive, removes high-strength actives that are inappropriate for reactive skin, such as the stronger salicylic acid and niacinamide formulations. Finally, a referral check recognises presentations that warrant the attention of a professional rather than an over-the-counter routine, and flags them accordingly.

3.4.9 Agent 7: The Prose Layer Agent

The final agent generates the natural-language explanation that accompanies the routine. It sends GPT-4o the canonical profile and the safety-validated set of recommendations and asks it to explain, in the user's chosen language, why each product has been suggested and how the routine should be followed, together with the appropriate disclaimers. The important constraint is that this agent explains decisions that have already been made; it does not make them. The set of products is fixed before the prose agent is invoked, and nothing the agent generates can add a product, remove a safety flag, or override a conflict ruling. The model's fluency is therefore used where it is valuable, in producing a clear and reassuring explanation, and withheld from the place where it would be dangerous, in deciding what is safe to recommend.

3.4.10 The User-Facing Flow

The interface presents the pipeline to the user as a short, guided sequence. The user is first asked to provide one or more photographs. The capture screen, shown in Figure 14, requests a front view and optionally the two cheek profiles, and states the conditions for a usable photograph (natural light, no flash, hair pulled back) that the quality gate of Section 3.4.3 then enforces. One photograph is enough to proceed, but additional angles improve the reading.



Fig. 14: The photo-capture screen, requesting a front view and optional cheek profiles.

The user then answers a short questionnaire, shown in Figure 15. Four questions establish the questionnaire fields of the data contract: the self-reported skin type, whether the skin reacts easily to products, the age band, and the primary concern, the last drawn from the same eight-category taxonomy used throughout the system. While the analysis runs, the interface shows the progress screen of Figure 16, which names the stages of the pipeline as they execute, from the photo-quality check through the visual analysis and the formulation-safety check to the construction of the routine. The interface is bilingual throughout, in keeping with the bilingual catalogue.

Fig. 15: The four questions of the skin questionnaire: skin type, reactivity, age band, and primary concern.

Fig. 16: The progress screen, naming the stages of the pipeline as they execute.

3.4.11 An End-to-End Example

To make the operation of the pipeline concrete, this subsection follows a single request from photograph to recommendation. Figure 17 traces the data contract as it grows through the pipeline. The user uploads a front-facing photograph and answers the questionnaire as follows: combination skin, does not report reacting easily to products, age band 25 to 35, and primary concern dark spots, matching the questionnaire shown in Figure 15.

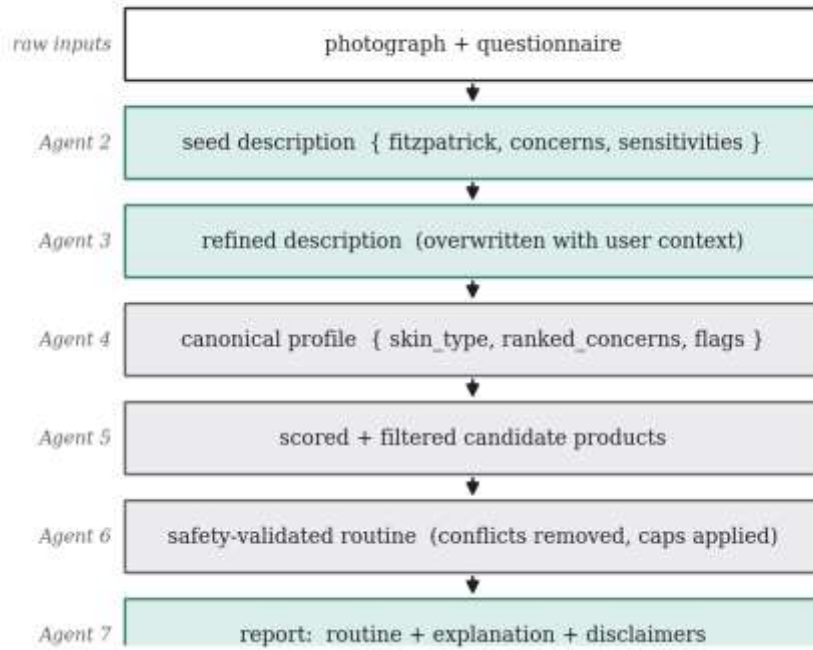


Fig. 17: The data contract as it grows through the pipeline for a single request.

The quality gate accepts the photograph. The seed pass returns a provisional reading (Listing 3.2): Fitzpatrick type IV, with oiliness, active acne, and dark spots judged present. The refinement pass, given the questionnaire, retains the tone estimate and the concerns but adjusts the sensitivity reading in line with the user's answers. The profile agent consolidates these into the canonical profile, in which dark spots, the user's stated priority, is promoted to the head of the ranked concerns.

The recommendation agent scores the catalogue against this profile, retaining products suitable for combination skin and ranking them by how well they address the ranked concerns, with products addressing dark spots scoring highest. The safety agent then prunes the result: where two highly scored products would stack the same active or carry a known conflict, one is removed, and the active-family caps are applied so that the final routine contains at most one product from each active family. The prose agent receives the surviving, safety-validated routine and produces the explanation the user reads, naming each product, relating it to the user's concerns, and adding the standard disclaimer that the advice is cosmetic guidance and not a medical diagnosis. The user thus receives a short, ordered routine in which every product can be traced back, through the contract of Figure 16, to a concern in the profile and a rule in the safety layer.

3.5 Results

3.5.1 Evaluation Metrics

The system's perception is evaluated against the ground truth of Section 3.3 using standard measures. For the Fitzpatrick attribute, two accuracies are reported, as defined in Section 3.3.6: exact-match accuracy and within-one-type accuracy. For each of the eight concerns, detection is treated as a binary problem and summarised by precision, recall, and the F1 score [30]. Precision is the proportion of the images for which the system reported a concern that the concern was in fact present; it answers the question of how far the system's positive reports can be trusted. Recall is the proportion of the images on which the concern was actually present that the system reported it; it answers the question of how much the system misses. The F1 score is the harmonic mean of the two and summarises the balance between them in a single number. The harmonic mean is used rather than the arithmetic mean because it penalises a large imbalance between precision and recall, so that a detector cannot score well by being cautious in one direction at the expense of the other.

3.5.2 Fitzpatrick Classification

Table 14 reports the accuracy of the system's Fitzpatrick estimates, and Figure 19 shows the same figures. The system names the exact Fitzpatrick type in a majority of cases, and falls within one type of the ground truth in the large majority of cases. The gap between the two figures is the more informative quantity: it shows that almost all of the system's tone errors are errors of one step on an ordinal, noisy scale, rather than gross misreadings. Given that the Fitzpatrick scale is known to produce disagreement even among trained human assessors, a within-one-type accuracy approaching nine in ten is a strong result for a zero-shot, general-purpose model, and it is achieved across the III to VI range rather than only on the lighter types.

Table 14: Fitzpatrick classification accuracy on the test set.

Measure	Accuracy
Exact match	0.62
Within one type	0.897

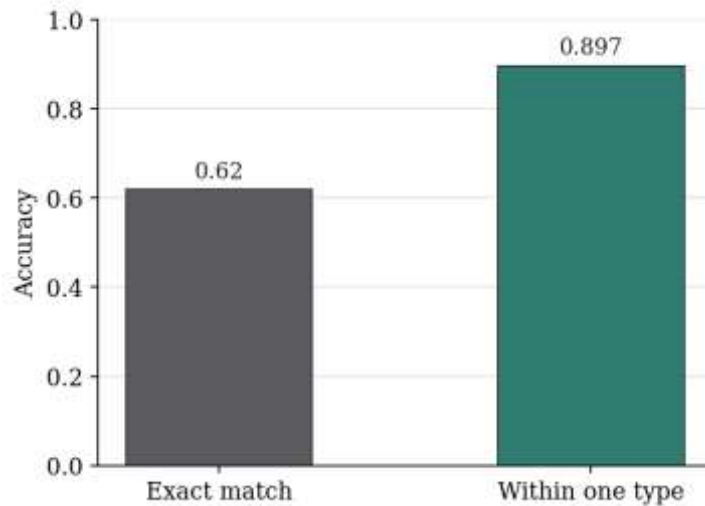


Fig. 18: Fitzpatrick accuracy: exact match against within-one-type.

3.5.3 Per-Concern Detection

Table 14 reports precision, recall, and F1 for each of the eight concerns, and Figure 19 shows the F1 scores ranked. The pattern across the concerns is coherent and interpretable. The concerns that present as clear, localised, high-contrast features in a photograph are detected best: oiliness, active acne, and post-inflammatory marks all reach F1 scores in the upper half of the range, driven by high precision, which means that when the system reports them it is usually right. The concerns that present as diffuse, low-contrast, or texture-dependent features are detected less well: dryness and uneven tone sit at the bottom of the ranking.

Table 15: Per-concern detection performance: precision, recall, and F1.

Concern	Precision	Recall	F1
Active acne	0.85	0.50	0.62
Acne marks / PIH	0.65	0.52	0.58
Oiliness	0.85	0.55	0.65
Dark spots	0.45	0.50	0.47
Enlarged pores	0.55	0.58	0.56
Fine lines	0.75	0.40	0.52
Dryness	0.60	0.15	0.24
Uneven tone	0.06	0.67	0.11

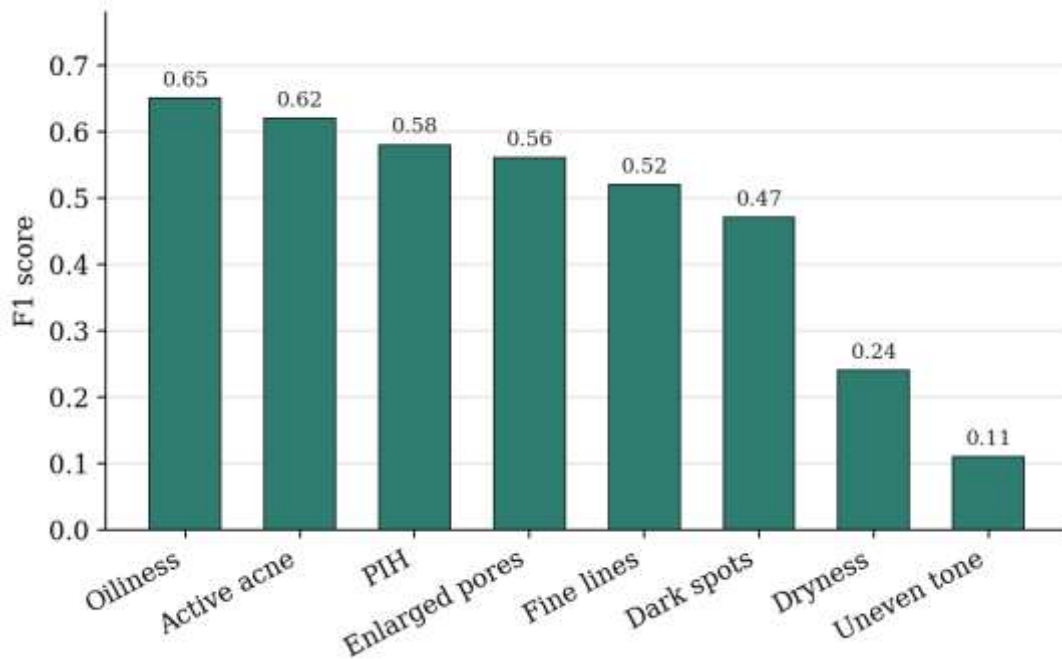


Fig. 19: F1 score for each of the eight concerns, ranked.

A recurring shape in the table is worth naming: for several concerns precision is markedly higher than recall, most sharply for dryness, and also for active acne and fine lines. The system, in other words, tends to under-report rather than over-report these concerns: when it does flag them it is usually correct, but it misses a proportion of the cases in which they are present. For a guidance system this is the safer of the two failure modes, since a missed concern leads to a routine that is merely incomplete rather than to one that recommends an inappropriate active for a concern that is not there. Uneven tone is the exception, and is discussed next.

3.5.4 Where the System Works and Where It Struggles

Dryness is the clearest case of a concern that is real on the skin but faint in the photograph. Its visible signature, fine flaking and a loss of surface sheen, is exactly the kind of low-contrast, lighting-dependent texture that a single consumer photograph captures poorly, and the system's low recall for dryness reflects this photographic limit rather than a flaw in its reasoning. It is detected reliably only when the flaking is pronounced.

Uneven tone shows the opposite imbalance, with high recall but very low precision, which produces the lowest F1 of the eight. Assessed as a diffuse, whole-face property, uneven tone is easily confused with the ordinary, healthy variation in facial colouring and with the appearance of pigmentation under uneven lighting, so the system reports it often but is

frequently wrong to do so. This is the one concern for which the system's reports should be treated with caution, and it is a candidate for a more tightly specified definition in future work.

Taken together, the per-concern results support a measured claim. The system perceives the concerns that a photograph can carry, the lesional and textural ones, with usable accuracy and a conservative bias, and it perceives the concerns that a photograph carries poorly, the diffuse pigmentary ones, less well and with a known failure mode. This is a more honest and more useful characterisation than a single headline accuracy would give, and it tells a deploying brand exactly which of the system's readings to trust.

3.5.5 Behaviour of the Safety Layer

Because the safety layer is deterministic, its behaviour is not estimated statistically but verified by construction, and this is itself a result. For any profile and any set of candidate products, the conflict check, the active-family caps, the sensitive-skin filter, and the referral check produce the same output every time, and that output can be read directly from the rules. In testing, the layer behaved exactly as specified: candidate sets that contained a known conflicting pair had one member removed; candidate sets that would have stacked two products of the same active family were reduced to one; and profiles flagged as sensitive never received the high-strength actives excluded by the sensitive-skin filter. The contrast with the perceptual stages is the point of the architecture: the parts of the system that can only be estimated are estimated and reported with their error rates, while the parts that bear on safety are made exact and need no error rate, because they cannot, by construction, do anything other than what their rules say.

3.5.6 Comparison with Commercial Approaches

A direct numerical comparison against the commercial tools surveyed in Chapter 2 is not possible, because those tools do not publish per-concern or per-skin-type performance figures, a gap that was itself one of the findings of Section 2.3. The comparison that can be made is qualitative and structural, and it is set out in Table 16. The point of the comparison is not that the present system perceives skin more accurately than a mature commercial product, a claim the evidence here would not support, but that it differs from those products on the two axes that Chapter 2 identified as gaps: it exposes an explicit, auditable safety layer, and it reports its perceptual accuracy across the Fitzpatrick range, including the darker types, rather than leaving both unstated.

Table 16: Structural comparison with a typical commercial skincare tool.

Property	Typical commercial tool	This system
Personalised recommendation	Yes	Yes
Explicit, auditable safety logic	Rarely disclosed	Yes, deterministic (Agent 6)
Conflict and active-stacking checks	Not disclosed	Yes, rule-based
Performance reported across skin tones	Rarely	Yes, Fitzpatrick III to VI
Safety decisions kept out of the model	Not disclosed	Yes, by design

3.5.7 Threats to Validity

Several limitations qualify the results and are stated plainly. First, the evaluation was reduced by the rate-limiting of the model endpoint described in Section 3.3.6: figures are computed only over the images for which the full pipeline completed, and a different or larger run might shift them. Second, the concerns are not equally represented in the test set, so the per-concern figures for the rarer concerns are estimated less precisely than those for the common ones and should be read as indicative rather than definitive. Third, the recommendation and safety layers were exercised against a single brand's catalogue; their logic is general, but their behaviour with a much larger or differently structured catalogue has not been measured here. Fourth, the ground truth was applied by inspection of the photographs, which bounds the evaluation to what is visible in a photograph and cannot speak to concerns that require touch or instrumentation to assess. None of these threatens the central architectural claim of the thesis, which rests on the design of the safety layer rather than on any particular accuracy figure, but each marks a direction in which the evaluation could be strengthened.

3.6 Conclusion

This chapter has presented the implementation of the system and the results of evaluating it. The system was described as a pipeline of seven agents that turns a photograph and a short questionnaire into a safe, personalised skincare routine with an explanation, built on a React and Vite frontend, a Python and FastAPI backend, and the GPT-4o multimodal model used zero-shot. Two design decisions were shown to run through the implementation: the seed-

then-overwrite treatment of perception, in which an initial reading of the photograph is refined in the light of the user's own account of their skin, and the strict separation of the learned components from the deterministic ones, which confines the model to perceiving and explaining and places every safety-critical decision in explicit, auditable rules.

The evaluation measured the system where it can only be estimated and verified it where it can be made exact. The perception of skin tone was accurate to within one Fitzpatrick type in the large majority of cases across the III to VI range, and the perception of visible concerns was accurate, with a conservative bias, for the concerns that a photograph can carry, and weaker, with a clearly characterised failure mode, for the diffuse pigmentary concerns that it cannot. The deterministic safety layer behaved exactly as specified, removing conflicts and capping active families by construction rather than by estimate. Against the commercial tools of Chapter 2, the system's distinguishing properties are precisely the two that Subsection 2.3.3 identified as missing from the existing work: an auditable safety layer, and performance reported across skin tones.

The general conclusion of the thesis, which follows, draws these findings together with those of the earlier chapters and sets out the directions, foreshadowed in the threats to validity above, in which the work could be carried further.

General Conclusion

This thesis set out to address a problem that is at once ordinary and consequential: the difficulty an ordinary consumer faces in deciding what to put on their skin. The market offers an overwhelming number of products, the advice that accompanies them is fragmented and frequently commercial in motive, and the tools that promise to simplify the choice tend to obscure the reasoning behind their recommendations. Chapter 2 examined this consumer information problem in detail and, in its survey of the related work, identified three gaps that together framed the contribution of the work. The academic literature on skin and artificial intelligence concentrates on diagnosis rather than on the everyday guidance a consumer actually needs; the commercial applications that offer such guidance rarely expose any explicit, auditable account of how they keep their recommendations safe; and neither body of work reports how its performance holds up across the full range of skin tones. The work was built around closing those three gaps at once.

The argument was developed over three chapters. Chapter 1 established the technical foundations on which the work rests, covering artificial intelligence, machine learning, deep learning, large multimodal models, and recommendation systems. Chapter 2 turned to the domain itself, setting out the structure of the skincare industry, the dermatological background needed to reason responsibly about skin, and the consumer information problem, before surveying the academic and commercial work that surrounds it and naming the gaps above. Chapter 3 presented the system that was built in response, described its implementation agent by agent, and reported the results of evaluating it.

The system is a guidance tool rather than a diagnostic one: its purpose is not to name a disease but to assemble a safe, personalised skincare routine, with an explanation, from a real product catalogue. It is organised as a pipeline of seven cooperating agents, built on a React and Vite interface, a Python and FastAPI backend, and the GPT-4o multimodal model used without any task-specific training. Two design decisions defined it. The first is the seed-then-overwrite treatment of perception, in which an initial reading of the photograph is deliberately revisited in the light of what the user reports about their own skin. The second, and the more important, is a strict separation between the learned components of the system and its deterministic ones. Perception and the generation of explanations are delegated to the model; every decision that bears on the safety of a recommendation, the avoidance of incompatible ingredients, the limiting of active families, and the recognition of cases that warrant a

professional, is taken by explicit, rule-based logic that never passes through the model. This separation is the central contribution of the thesis. Its consequence is that the advice a user receives is produced by code whose behaviour can be read, tested, and audited, while the model is confined to the two tasks for which it is genuinely well suited: seeing and explaining.

The system was evaluated against a hand-labelled test set of facial images spanning Fitzpatrick types III to VI, used to measure the behaviour of the frozen, pretrained model placed inside the pipeline rather than to train anything. No model was trained in the course of this work; a system was built around a pretrained one and measured. The results were estimated where they could only be estimated and verified where they could be made exact. The system's reading of skin tone fell within one Fitzpatrick type of the ground truth in the large majority of cases across the III to VI range, a strong result for a general-purpose model on an ordinal scale that is noisy even among trained human assessors. Its detection of visible concerns was accurate, and conservatively biased, for the concerns that a photograph can carry, and weaker, with a clearly characterised failure mode, for the diffuse pigmentary concerns that a single photograph carries poorly. The deterministic safety layer, by contrast, did not need to be estimated at all: because it is rule-based, its behaviour is correct by construction and was confirmed to remove conflicting combinations, cap active families, and protect reactive skin exactly as specified.

Taken together, these results answer the three gaps that motivated the work. Against the diagnostic emphasis of the academic literature, the system delivers actionable guidance. Against the opacity of the commercial tools, it exposes a safety layer that anyone can read and check. And against the silence of both on the question of skin tone, it reports its perceptual accuracy openly across the darker types that are most often under-served. The contribution is therefore not a claim that the system perceives skin more accurately than a mature commercial product, which the evidence would not support, but a demonstration that a guidance system can be both genuinely personalised and demonstrably safe, and that the safety can be made the property of auditable code rather than a hope placed in a model.

The work has clear limitations, stated here as plainly as they were in the results. The evaluation was carried out at a modest scale and was further reduced by the rate-limiting of the model endpoint, so the reported figures are best read as well-founded indications rather than final measurements. The recommendation and safety logic was exercised against a single brand's catalogue; the logic is general, but its behaviour with a much larger or differently structured catalogue has not been measured. The ground truth was applied by inspection of

photographs, which bounds the evaluation to what a photograph can show and leaves aside the concerns that require touch or instrumentation to assess.

Several directions follow naturally from these limitations. The most immediate is a larger and more balanced evaluation, free of the endpoint constraints that shaped this one, which would tighten the per-concern figures and allow per-skin-type performance to be reported with confidence. The diffuse concerns that the system reads poorly, uneven tone in particular, would benefit from a more tightly specified definition and from evidence drawn from more than a single photograph. Extending the system to several catalogues would test the generality of the recommendation and safety logic and is the natural step towards deployment. The deterministic safety layer itself can be enriched, both by broadening its rule base and by submitting those rules to review by a dermatologist or pharmacist, which the very auditability of the design makes straightforward.

Beyond the particular system, the thesis offers a pattern that should generalise. In any application where a powerful but fallible model is asked to advise on matters that touch a person's wellbeing, the safe course is neither to trust the model with everything nor to abandon it, but to use it precisely where its fluency is valuable and to place every consequential decision in logic that can be inspected. Confining the model to perceiving and explaining, while keeping the safety of the advice a deterministic and auditable property of the system, is the principle this work was built to demonstrate, and it is the principle on which any responsible extension of it should continue to rest.

Bibliography

- [1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2021.
- [2] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950. doi: 10.1093/mind/LIX.236.433.
- [3] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the Dartmouth summer research project on artificial intelligence," Aug. 31, 1955. [Online]. Available: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
- [4] T. M. Mitchell, *Machine Learning*. New York, NY: McGraw-Hill, 1997.
- [5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. doi: 10.1038/nature14539.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, 2012, pp. 1097–1105.
- [7] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 5998–6008. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [8] OpenAI, "Hello GPT-4o," OpenAI, May 13, 2024. [Online]. Available: <https://openai.com/index/hello-gpt-4o/> (accessed: May 22, 2026).
- [9] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, Article 248, pp. 1–38, Mar. 2023. doi: 10.1145/3571730.
- [10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. doi: 10.1038/nature21056.
- [11] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1820–1828. [Online]. Available: <https://arxiv.org/abs/2104.09957>
- [12] F. Ricci, L. Rokach, and B. Shapira, Eds., *Recommender Systems Handbook*, 3rd ed. New York, NY: Springer, 2022. doi: 10.1007/978-1-0716-2197-4.

- [13] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Lecture Notes in Computer Science, vol. 4321. Berlin, Heidelberg: Springer, 2007, pp. 325–341. doi: 10.1007/978-3-540-72079-9_10.
- [14] Fortune Business Insights, "Skin Care Market Size, Share & Growth Report 2032," Fortune Business Insights, 2025. [Online]. Available: <https://www.fortunebusinessinsights.com/skin-care-market-102544> (accessed: May 22, 2026).
- [15] Roots Analysis, "Skincare Market Size, Share, Growth Trends, Report to 2035," Roots Analysis, 2024. [Online]. Available: <https://www.rootsanalysis.com/reports/skin-care-market.html> (accessed: May 22, 2026).
- [16] Market Data Forecast, "Skincare Market Size, Share, Trends & Analysis, 2034," Market Data Forecast, 2026. [Online]. Available: <https://www.marketdataforecast.com/market-reports/skincare-market> (accessed: May 22, 2026).
- [17] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Archives of Dermatology*, vol. 124, no. 6, pp. 869–871, Jun. 1988. doi: 10.1001/archderm.1988.01670060015008.
- [18] OpenStax, *Anatomy and Physiology 2e*, ch. 5: "The Integumentary System." Houston, TX: OpenStax / Rice University, 2022. [Online]. Available: <https://openstax.org/books/anatomy-and-physiology-2e/pages/5-1-layers-of-the-skin>
- [19] R. V. Reynolds, H. Yeung, C. E. Cheng, F. Cook-Bolden, S. R. Desai, K. M. Druby, et al., "Guidelines of care for the management of acne vulgaris," *Journal of the American Academy of Dermatology*, vol. 90, no. 5, pp. 1006.e1–1006.e30, May 2024. doi: 10.1016/j.jaad.2023.12.017.
- [20] S. Mukherjee, A. Date, V. Patravale, H. C. Korting, A. Roeder, and G. Weindl, "Retinoids in the treatment of skin aging: an overview of clinical efficacy and safety," *Clinical Interventions in Aging*, vol. 1, no. 4, pp. 327–348, Dec. 2006. doi: 10.2147/ciia.2006.1.4.327.
- [21] A. A. Almeman, "Evaluating the efficacy and safety of alpha-hydroxy acids in dermatological practice: a comprehensive clinical and legal review," *Clinical, Cosmetic and Investigational Dermatology*, vol. 17, pp. 1661–1685, Jul. 2024. doi: 10.2147/CCID.S453243.

- [22] J. M. Pullar, A. C. Carr, and M. C. M. Vissers, "The roles of vitamin C in skin health," *Nutrients*, vol. 9, no. 8, p. 866, Aug. 2017. doi: 10.3390/nu9080866.
- [23] Y. C. Boo, "Mechanistic basis and clinical evidence for the applications of nicotinamide (niacinamide) to control skin aging and pigmentation," *Antioxidants*, vol. 10, no. 8, p. 1315, Aug. 2021. doi: 10.3390/antiox10081315.
- [24] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018. doi: 10.1093/annonc/mdy166.
- [25] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, Aug. 2018. doi: 10.1038/sdata.2018.161.
- [26] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, et al., "Human–computer collaboration for skin cancer recognition," *Nature Medicine*, vol. 26, pp. 1229–1234, Aug. 2020. doi: 10.1038/s41591-020-0942-0.
- [27] S. Ramírez, "FastAPI," 2018. [Online]. Available: <https://fastapi.tiangolo.com/>. [Accessed: Jun. 1, 2026].
- [28] Meta Open Source, "React: The library for web and native user interfaces." [Online]. Available: <https://react.dev/>. [Accessed: Jun. 1, 2026].
- [29] E. You, "Vite: Next generation frontend tooling." [Online]. Available: <https://vite.dev/>. [Accessed: Jun. 1, 2026].
- [30] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

Abstract

This thesis presents an AI-powered system that produces personalised, safety-checked skincare guidance from a single facial photograph and a short questionnaire. The system is organised as a pipeline of seven cooperating agents built on a React and FastAPI stack, in which a multimodal large language model (GPT-4o) performs the visual perception and writes the final report, while every safety-bearing decision, the detection thresholds for skin concerns, the checking of conflicts between recommended products, and the referral of cases that require a professional, is handled by deterministic code. Recommendations are drawn from a real product catalogue, so that every suggestion can actually be obtained. The system is evaluated on a labelled test set spanning Fitzpatrick skin types III to VI, with particular attention to fairness across skin tones. The results show strong within-one-type classification accuracy alongside a systematic under-detection of concerns, reported openly rather than concealed.

المخلص

تقدّم هذه الأطروحة نظامًا مدعومًا بالذكاء الاصطناعي لإنتاج إرشادات مخصّصة للعناية بالبشرة، يتم التحقق من سلامتها، يضم سبعة وكلاء متعاونين، (Pipeline) اعتماداً على صورة واحدة للوجه واستبيان قصير. يتكوّن النظام من خط معالجة لتنفيذ مهام (GPT-4o) ويعتمد النظام على نموذج لغوي كبير متعدد الوسائط. React و FastAPI مبنياً باستخدام تقنيتي الإدراك البصري وإعداد التقرير النهائي، في حين تُدار جميع القرارات المتعلقة بالسلامة من خلال شيفرة برمجية حتمية بما في ذلك تحديد عتبات اكتشاف مشكلات البشرة، والتحقق من التعارضات بين المنتجات (Deterministic Code) الموصى بها، وإحالة الحالات التي تستدعي استشارة مختص.

تُستمد التوصيات من كتالوج حقيقي للمنتجات، مما يضمن إمكانية الحصول على جميع المنتجات المقترحة فعليًا. وقد جرى تقييم النظام باستخدام مجموعة بيانات اختبارية مُعنونة تشمل أنواع البشرة وفق تصنيف فيتزباتريك من النوع الثالث إلى مع التركيز بشكل خاص على تحقيق العدالة عبر درجات ألوان البشرة (Fitzpatrick Skin Types III–VI) السادس المختلفة.

أظهرت النتائج دقة عالية في تصنيف الحالات ضمن النوع الواحد من البشرة، إلى جانب وجود ميل منهجي إلى التقليل من اكتشاف بعض المشكلات الجلدية. وقد تم الإبلاغ عن هذا القيد بصورة شفافة وصريحة بدلاً من إخفائه.