Fostering Transparency of Judicial Decisions and Enhancing the National Implementation of the ECHR (TJENI)

AI tools for the automatic summarisation of judicial decisions in Greek: a testing methodology

This report has been edited based on the draft prepared by Mario Ragazzi.

The opinions expressed in this work are the responsibility of the authors and do not necessarily reflect the official policy of the Council of Europe.

The project "Foster Transparency of Judicial Decisions and Enhancing the National Implementation of the ECHR" (TJENI) is funded by Iceland, Liechtenstein and Norway through the EEA and Norway Grants Fund for Regional Cooperation

Table of Contents

1. Introduction....3
2. Automated summarisation....6
3. Testing ground....13
4. Risks and points for consideration....14
5. Tools to be tested....18
6. Financial considerations....19
7. Testing procedure....24
8. Conclusions of pilot testing for the Supreme Court of Cyprus....27
References....29
Annex I. Cost estimation with GPT-3.5 and GPT-4....32
Annex II. Training configuration diagram....33
Annex III. Systematic testing configuration diagram....34
Annex IV. Usage scenario in an editorial workflow....35

1. Introduction

The use of digital tools in nearly every domain is gaining more and more popularity with more advanced features introduced by the day.

Among them, Artificial Intelligence (AI) – a set of scientific methods, theories and techniques whose aim is to reproduce, by a machine, the cognitive abilities of human beings. Current developments seek to have machines perform complex tasks previously carried out by humans[1].

Using new technology in judiciary is no exception to this, and new IT tools available on the market (or yet to be developed) pose new significant challenges related to their applicability, efficiency, correctness, and ethicality.

The Project "Foster Transparency of Judicial Decisions and Enhancing the National Implementation of the ECHR" (hereinafter – "TJENI Project") works with the justice systems in the partner states of the Project with the aim to strengthen the quality of their judicial decision-making through the integration of specialised tools and solutions aimed at improving the consistency, quality and transparency of judicial decisions.

This paper is developed to primarily examine automatic summarisation of judicial decisions in Greek by (more or less) digital tools with machine learning element(s).

For the purposes of this methodology, by summarisation we mean an operation that transforms an initial text T1 into a new text T2:

$$T1 \rightarrow T2$$

where:

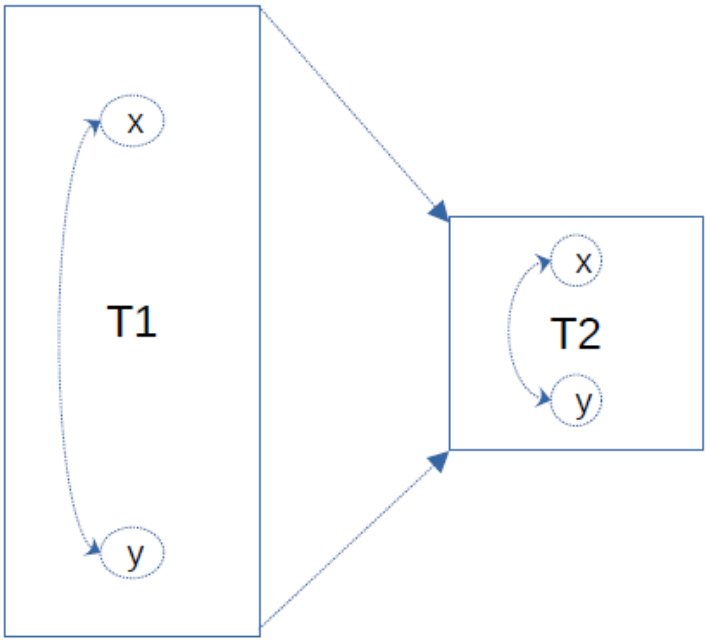
Length(T2) << Length(T1): Compression

Relevant information (T2) ~ Relevant information (T1): Referential invariance

The tested summarisation of judicial decisions represent mainly a form of action with a linguistic text that is transformed but shall keep its referential function.

[1] See CEPEJ European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, p. 69. Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

The question about the relevance of the information in the transferred text is decided by the person in change of the summarisation in each individual case. The scope of the relevant information that shall be preserved in the summarised text can be specified in the task part (prompt). Unlike other extractive operations on texts, however, summarisation carries the connotation of the preservation of the *whole* sense of the text, albeit in reduced form.

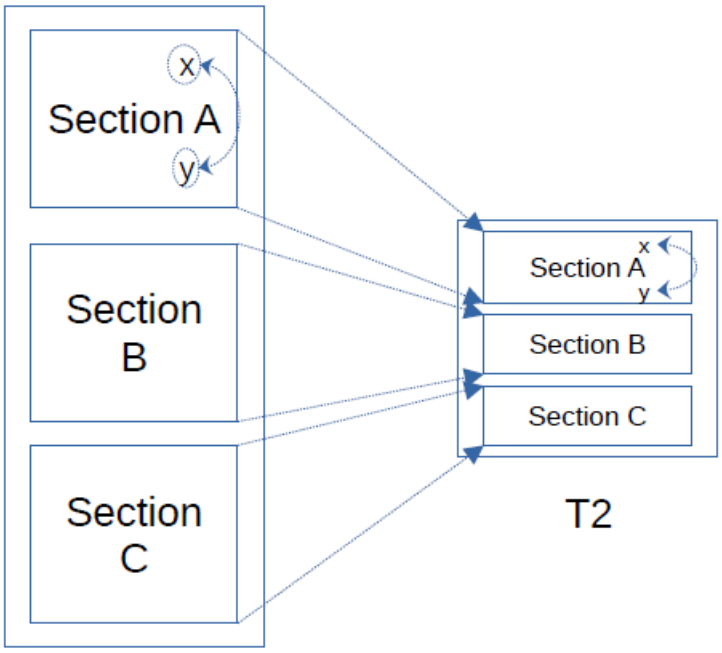


Drawing 1: A logical or referential connection between elements at the beginning and end of the text should be preserved in the summarised text.

In the summarisation process – the relevant information that should be preserved in compressed form – critically depends on the *intended audience* and their *use cases*.

What is equally important in the context of judicial decisions, is that in the summarised text the analysis is kept. It extracts relevant information from the original text with a holistic view and through a process of abstractive transformation, logic association and combination. In case of summarisation by human the theoretical doctrines and concepts are kept in the text to help reading and understanding of its main sense. Usually, the summarisation of judicial decisions is done by judges and legal experts of the same court that adopted the judgement.

When the summarised document is already structured in well-defined sections that is reflected in corresponding sections in the summary and significantly helps the summarisation process. It will be sufficient to summarise each section one at a time. In case of the summarisation with the assistance of algorithms the presence of the structures in text has an important implication on the efficiency of the summarisation.



Drawing 1: In a segmented text, a logical or referential textual connection must be preserved only within each section.

Recent availability of Large Language Models-based tools (one kind of so-called “Artificial Intelligence”) for effective Natural Language Processing operations has become an important enabling circumstance for automatic summarisation of texts, including judicial decisions.Such availability of relevant AI tools has marked a necessity to explore both:

- General (or "strategic") need
    - summarisation as a stepping stone to classification and analysis of potential unintentional divergency in jurisprudence (thus, reinforcing compliance with Article 6 of the European Convention on Human Rights)

        - e.g. Labelling for similarity analysis as in the French initiative developed in its Court of Cassation
- Contingent need
    - Legal Reports of the Supreme Court, for which the Supreme Court of Cyprus (SCC) explores different ways to improve the efficiency of its editorial processes in producing structured summaries of judicial decisions for their flagship publication

This testing methodology is, thus, tasked to examine whether it is possible or feasible to do automatic summarisation of judicial decisions in Greek by AI digital tools.

The testing is also enabling to collect systematic user feedback in a phase of fluid development; inform future procurement and contribute to standardisation of processes in judiciary.

2. Automated summarisation

The digital developments are appearing very fast, with new products announced on a weekly basis.

2.1. Main concepts and terminology

The latest IT developments on AI basis apply neural network approach: an information processing architecture with a great number of nodes ("neurons") that are placed in layers connected to each other, each of them able to carry out simple computations on the signals received from the connected nodes in the preceding layers and the internal status of each node, represented by a modifiable *weight*. The node passes on the result of its computation to more nodes in the following layer, and so on, until the network produces a result with a certain statistical regularity.

Machine Learning

The neural network approach turned out to provide a good basis for so-called Machine Learning (ML). This refers to a type of algorithm used by computers and computer programs for data analysis, prediction, classification, clustering tasks. It involves training an artificial intelligence model using large datasets or information from various sources such as news articles, social media posts etc., which then uses

the pre-programmed rules learned through machine learning algorithms (such as the thousands, or millions of weights associated to nodes in a neural network) to make predictions on new and unseen patterns of data in a specific field like image classification or finance analysis, for example, predicting stock prices.

> "*OpenAI's ChatGPT, Google's Bard and Microsoft's Sydney are marvels of machine learning. Roughly speaking, they take huge amounts of data, search for patterns in it and become increasingly proficient at generating statistically probable outputs — such as seemingly humanlike language and thought*." (Chomsky, 2023)

When processing natural language, the neural network is fed with a sequence of tokens, a sort of intermediate unit between individual letters and whole words used to map text to numbers (but this is not the case for Greek language, as described below). There are many kinds of neural networks depending on their architecture, how layers are connected to each other, and the kind of computation carried out by nodes (recurrent, adversarial, convolutional networks etc.) (Wind, 2023).

Transformer Neural Network (TNN)

A major breakthrough in the design of neural networks for natural language processing was the invention of the so-called Transformers by Google. The higher complexity of Transformer networks requires enormous amounts of computing power and large quantities of data for the effective definition of the associated model.

A Transformer Neural Network (TNN) or more commonly known as Transformers are an architecture that was originally proposed by Vaswani *et al.* (2017), to tackle the limitations of previous architectures in language translation tasks. TNN uses so-called *self-attention* mechanisms, which allow it to learn distributed representations for input sequences at scale without requiring explicit supervision on sequence order or length information, thus allowing the model to be more efficient while maintaining good performance (Raschka, 2023; Vaswani et al., 2017).

The core idea behind self-attention is to compute the importance of each word or token in a sequence by looking at the relationships between all the words in the sequence. This allows the model to capture dependencies and relationships between

different positions in the input sequence without relying on recurrent or convolutional structures. The self-attention mechanism computes attention scores between each pair of words in the input sequence. These scores represent the importance or relevance of one word to another. The attention mechanism allows the model to weigh the influence of different words when processing each word.

Large Language Models (LLMs)

LLMs, also known as transformers or autoregressive language model encoders/decoder networks, have made significant progress in natural language processing over the past decade. The most notable developments include neural machine translation models that outclass human translators on a wide range of text genres and tasks. In addition, large LLMs are becoming increasingly common for many natural language processing (NLP) applications. Despite their impressive performance, these models remain computationally expensive to train even on a moderate scale. This limitation is likely due in part to the very large number of parameters required by such LLMs – each encoder and decoder needs hundreds or thousands more neurons than an equivalent smaller model. Besides, training requires very large amounts of data. To address these limitations, several strategies have been proposed for fine-tuning (i.e., pre-process) the input text during inference: reducing computation overhead by using pre-trained models; optimising parameters to reduce computational demands and speed up processing time at scale; or increasing parameter efficiency through techniques like low rank adaptation.

The most recent versions of LLMs – for instance Generative Pre-trained Transformer (GPT) – use neural networks pre-trained over gigantic amounts of data and are able to generate human-level text.

2.2. Ways of Machine Learning training

Generally, there are three main ways of machine learning training.

1. Pre-training during the construction of the model itself
    - Without user's inputs, huge amounts of data needed, very expensive
2. Fine-tuning of an existing general model

- Not for the average individual end-user, but doable and affordable with limited costs by an organisation.

3. During its application through tasks (prompts) and context variation
   - Can be done by a user, inexpensive, easy at first glance, but subtle.
     - In-prompt conditioning:

       Creating a semantic context, e.g. "*You are a legal assistant, and you have to help judges to summarise judicial decisions*..."

     In-context learning (ICL):

     - examples of the task are provided as context.

Reinforcement Learning from Human Feedback (RLHF) is the last phase of development of advanced language models. It takes a pre-trained general language model and further trains it on additional data against a reward model with substantial human feedback in the loop. The result is a layer of conditioning hidden from the end-user that greatly improves the model's responses for a specific task (e.g. behaving as a chatbot). (Lambert, 2022)

In short, LLMs (Large Language Models) have the ability to predict the most likely next token – based on their initial extensive training on a large corpus of text – after a given sequence. Repeating the process hundreds of times, they generate human readable text. Hence, they are usually *prompted* with an initial text, and they *complete* it. Prompting involves providing specific instructions or text as input to a generalised (not fine-tuned) LLM, guiding the model's response to generate relevant output.

Fine-tuning involves updating the model's parameters with a dataset, allowing the LLM to adapt to specific contexts.

2.3. Summarisation techniques

When deciding on the use of automated summarisation, it is important to differentiate between extractive and abstractive summarising.

- Extractive summarisation: select (part of) phrases from the original text that convey important meaning and copy them as they are to the output. Comparable to highlighting when reading.

- Abstractive summarisation: generate a text that efficiently and comprehensively conveys the main facts and meaning of the original with a desired compression factor.

Abstractive summarising works better with a wider "peripheral" view of the original, that is: summarise this page, but take into account what is written in the preceding and following ones. This helps to establish a hierarchy of relevance within the text, about what is more or less relevant (this can also be guided by in-prompt instructions, see below).

The Economist quoted Dr Percy Liang, a professor at the Institute for Human-Centred Artificial Intelligence at Stanford University, saying that 'today's LLMs, which are based on the so-called "transformer" architecture developed by Google, have a limited "context window"—akin to short-term memory. Doubling the length of the window increases the computational load fourfold. That limits how fast they can improve. Many researchers are working on post-transformer architectures that can support far bigger context windows—an approach that has been dubbed "long learning" (as opposed to "deep learning").' (The Economist, 2023b)

2.4. The problem of the input limitation

Language models can deal with a limited amount of text (tokens) in each interaction. The user's prompt (task) and machine's completion (produced result) combined cannot exceed a certain number of tokens. Through the mechanism of self-attention, the machine has the ability to maintain a *context* of previous inputs (prompts), and this in turn shapes how the machine responds. But this is any case limited by the maximum token size accepted by the model, and in most cases, this is not enough to process the text of a whole judicial decision.

Current limitations on the length of the initial text that needs to be summarised imposed by different AI tools available in open access include:

- GPT-2 : max 2k tokens
- GPT-3 : max 4k tokens
- GPT-4 : max 8k tokens, or even 32k (but limited access so far)

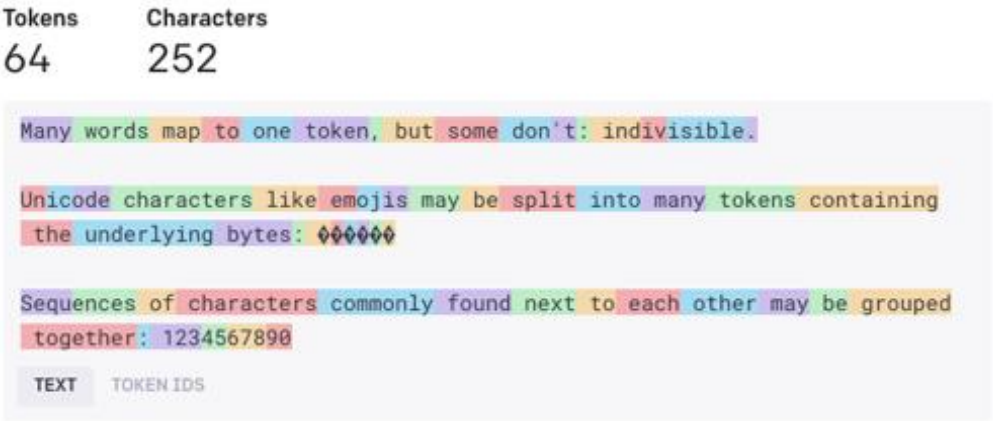
- Anthropic recently announced a window of 100k tokens for its *Claude* model (not clear when it will become available).

If the whole judicial decision cannot fit in one prompt window (as its length requires big number of tokens), there are some techniques to overcoming this limitation:

- Divide the text in (overlapping) chunks and create partial summaries of each.
- Collate the partial summaries into a new, shorter text. If this still does not fit a one prompt window, repeat the operation. Clearly, with every iteration the quality of the summary is likely to decrease.
- If the text is well-structured in sections, each of them can be summarised independently of the others and mapped to corresponding sections in the summary.
- More advanced strategies involve modifying the models at a deeper level, e.g. grafting a recurrent strategy on top of the transformer network or designing new kind of language models like in (Wind, 2023), which goes beyond the scope of this Methodology.

The problem of tokenisation for languages in non-Latin scripts (e.g. Greek)

Tokens are parts of words converted into numbers, e.g.



The conversion rate of token to characters remains different for different languages, for instance:

- English: 1 token ≈ 4 characters (1,500 words ≈ 2048 tokens)
- Greek:
  - 1 token ≈ 0.9 characters (*gpt2* encoder used in GPT-2 and GPT-3)
  - 1 token ≈ 1.15 characters (*cl100k_base* used in GPT-3.5)

In practical terms, it means that despite some improvement writing in Greek script means that less content fits in prompt window. In English, a whopping 20-thousand words long decisions would fit in a single 32k tokens prompt, with room to spare for a detailed prompt. It means that using AI tools in open access in Greek would be less efficient and more expensive (OpenAI charges a per-token fee –> summarising Greek is 4x more expensive than in English)

- Tokens are linked to bytes
- Bytes per character: depends on the encoding, less for Latin scripts, more for the Greek script.

An example of tokenisation with OpenAI

	Characters	Tokens (GPT-3)	Tokens (GPT-3.5)
Εφεσείοντες-Ενάγοντες	21	26	23
Appellants-Plaintiffs	21	7	7

App	ell	ants	-	Pl	aint	iffs
2213	616	1821	12	2169	1673	19383

E		φ	ε	σ	ε	ί	ο	ν	τ	ε	ς
138	243	86134	31243	45028	31243	55241	28654	34369	36924	31243	46742

Note:

E [Latin] : [36]

E [Greek] : [138, 243]

Hence, using standard OpenAI tokeniser (based on UTF-8 encoding) negatively affects Greek script. But it's not the only available tool and there might be an advantage is using alternative tokenisers that are efficient on the Greek script. Just as changing encoding system from GPT-2 or GPT-3 to GPT-3.5 (cl100k_base) slightly improved the C/T ratio for Greek, other encodings, perhaps native to Greek script, may do better.

Alternatively, it is advised to explore in practice the relative quality gains/losses of automatically translating a decision into English first, run the summariser, and translate the output back to Greek.

3. Testing ground

3.1. Existing dataset

The dataset of the Supreme Court of Cyprus includes judicial around 20 000 decision and their summaries in digital format (MS Word/PDF/HTML) for several past years. In particular it includes:

- Cyprus Law Reports (CLR): cases determined by the Supreme Court of Cyprus. The Court's flagship publication and main reference for legal practitioners in the country
- All decisions summarised by highly qualified legal experts that have the following sections:
    - Front matter with date, case number, parties, judges
    - Thematic keywords with context: a very brief description of the context in which the keywords (legal matters/questions of the case) appear in the judgement
    - The facts of the case
    - Grounds of appeal
    - Decision
    - Referred case law and statutory provisions (written law).

Since the Supreme Court switched the official language from English to Greek at the end of 1989 its decisions in Greek are summarised and published in the Cyprus Law Reports total of approximately 17 500 decisions. They are available in various formats: HTML on Cylaw website, MS Word or PDF in the Supreme Court's archives; although not necessarily all the formats exist for all decisions.

An additional number of approximately 3 500 decisions from 2017 to date have not been summarised yet, and they are available as HTML on Cylaw website[2] or MS Word in the Supreme Court's website.

3.2. Preparation of the dataset

The decisions for which a summary already exist may be used for the training of a model. The preparation of the dataset would involve some work of splitting summaries and decisions (usually they are found in the same file), format conversion (typically towards JSON as required by OpenAI) and data clean-up (e.g. "Α Π Ο Φ Α Σ Η" printed with spaces for emphasis, as it usually appears in the decisions, confuses the model. Other unnecessary information for the text shall also be removed (line and page numbers, headers and footers). Although this can be automatized with standard techniques, it will require additional resources.

4. Risks and points for consideration

There are several levels of concern regarding the use of AI-based tools.

The Council of Europe is working on the impact of such technology on human rights as part of its Digital Agenda 2022-2025 (Council of Europe, 2023). While it is commonly understood that AI can be of significant help in many areas, it also raises certain challenges, notably as regards the protection of privacy and personal data, risk of discrimination, lack of oversight of decision-making systems and the difficulty of applying existing legal frameworks to issues raised by AI.

In this light, the CoE's Committee on Artificial Intelligence is preparing a draft for a framework convention requiring signatory countries to take steps to ensure that AI

[2] A portal run by the Cypriot Bar Association that provides access to Cypriot law, secondary legislation, and court decisions. Available at: http://www.cylaw.org/

based tools are designed, developed, and applied in a way that protects human rights, democracy, and the rule of law (Committee on Artificial Intelligence (CAI), 2023).

In its turn, CEPEJ issued a European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment in which five principles are put forward:

1. Principle of respect for fundamental rights: ensure that the design and implementation of artificial intelligence tools and services are compatible with fundamental rights.
2. Principle of non-discrimination: specifically prevent the development or intensification of any discrimination between individuals or groups of individuals.
3. Principle of quality and security: with regard to the processing of judicial decisions and data, use certified sources and intangible data with models elaborated in a multi-disciplinary manner, in a secure technological environment.
4. Principle of transparency, impartiality and fairness: make data processing methods accessible and understandable, authorise external audits.
5. Principle "under user control": preclude a prescriptive approach and ensure that users are informed actors and in control of the choices made.

Another important document, Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: a proposal prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence, outlines a four-step process for assessing adverse impact on human rights, democracy and the rule of law generated by AI systems, addressing simultaneously the risks arising from the specific and inherent characteristics of AI systems and the impact of such systems on human rights, rule of law and democracy.

In view of the above, the major potential risks specific to AI-aided summarisation of judicial decisions are described in the following section.

4.1. Privacy and personal data protection

It is expected that all judicial decisions that need to be summarised are already in the public domain (in case of Cyprus: published on Cylaw and Legnet websites) and anonymised.[3] Nevertheless, the concerns about privacy and personal data protection during application of AI – based tools available in open access remain.

For example, a tool for summarising judicial decisions might be used by judges or legal officers on a draft text also *before* it has been anonymised and published.

Mounting concerns about feeding sensitive information to OpenAI or other companies are well-placed. Shortly after the Italian authority for the protection of personal data announced that it started an investigation about ChatGPT, it resulted in improved privacy policies of the ChatGPT (OpenAI, 2023), checking for the age of its users below 13 years, and introduction of an opt-out for personal information not to be included in their machine training process (Deutsche Welle, 2023). More recently, Apple, Samsung and other corporations have forbidden to their employees the use of ChatGPT (McGlauflin, 2023). In France, there exists legislation banning the use of predictive litigation AI.

4.2. "Black box" models

A LLM like ChatGPT presents itself to the user like a "black box", an opaque system that can be observed only in terms of its inputs and outputs, without any knowledge of its internal workings or the algorithms used to process those inputs. In particular, these models are difficult to understand or reverse-engineer due to their lack of transparency or the secrecy surrounding their design and operation. The knowledge built into the models, if any, is non-explanatory (Sobha Kartha, 2021) for reasons intrinsic in the way the models are built. Producing results that cannot be explained and justified poses a potential problem in particular to public institutions that have to be accountable to the public.

[3] The anonymisation of judicial decisions was introduced by the Cypriot publisher since June 2022. Following the entering into force of the GDPR the Supreme Court issued its first rules for the anonymization of judicial decisions on 19 July 2018. Until then, anonymisation was performed only to protect the identity of minors and victims of sexual crimes. The 2018 rules mandated to replace for any published decision the first name of parties, witnesses and other persons except for judges and counsel with 'XXX'."

Generated results are usually non-reproducible. The model does not answer to the same prompt with exactly the same text because of how the next-token likelihood is statistically calculated. Sometimes this is a feature, it seems that the model's answers are more "creative" or just less predictable, and there is a parameter called T that sets the degree or randomness in the model. For summarisation tasks, on the contrary, it is recommended to set T close to zero to reproduce as close as possible the facts and meaning of the original text.

In this light, it is noteworthy that according to a compilation called "Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law"[4] prepared by the CAHAI Secretariat, transparency is the most prevalent ethical principle in the current soft law spectrum. This typically involve the promotion of methodological approaches to "explainable AI", that is AI systems whose outputs and decisions can be understood by human experts. These methods and techniques contrast with "black box" approaches to machine learning where the steps through which an AI system arrived at a specific decision are unintelligible to human experts including the system's designers.

4.3. Unpredictable changes in API usage policy

Relying on a single major provider like OpenAI exposes developers and users to potential negative effects of a sudden changes in the provider's API usage policy. Unilateral changes cannot be ruled out, with unforeseen implications for the project costs along the way. This could be mitigated if the developer / user enters into a negotiated commercial agreement with OpenAI that establishes a more predictable technical perimeter within which to develop custom solutions.

4.4. Control over the tool

While AI tools available in open access have certain advantages in terms of accessibility and affordability, a careful consideration should be given to balancing all the risks when relying on private corporations for such an essential piece of the democratic institutions such as the administration of justice. Following an example set by

[4] Available at: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

the French Court of Cassation, it seems realistic to develop one's own AI-based tool for justice-related functions. As this approach is technically complicated and expensive, cross-border cooperation with judiciaries from other states having the same needs on the joint development might be considered.

4.5. Inter-passivity, or managing users' expectations

There is an inherent risk that we could name as the opposite of the inter-activity that a good technological support tool should facilitate in the best-case scenario. On the contrary, under inter-passivity (Žižek, 2006), the user like unconsciously expects the machine to act for them, and dealing with language and intelligence this might derive in a state like: "I think that the machine can think on my behalf, so I can stop thinking".

A similar idea is presented in the Opinion No.(2011)14 of the Consultative Council of European Judges "Justice and information technologies (IT)", stressing that IT must not prevent judges from applying the law in an independent manner and with impartiality (para. 8) and that IT cannot replace the judge's role in hearing and weighing the factual evidence in the case, determining the law applicable and taking a decision with no restrictions other than those prescribed by law (para. 31).

5. Tools to be tested

5.1. Off-the-shelf tools

The number of new AI-based tools with various functionality available on the market is rapidly growing. Preliminary research of the tools available on the market did not identify off-the-shelf tools ready for automated summarisation of judicial decisions in Greek. It is important to note, however, that the situation is rapidly evolving, and such tools can appear soon.

It appears that any off-the-shelf solution will have to involve some adjustment and development. Fortunately, there are many open, collaborative projects on automatic summarisation in general, from which to draw inspiration.

Selection criteria for commercial tools could include:

- Works natively with Greek language and script (input and output)

- Can accept for summarisation texts (or documents) with a minimum length of ~5000 characters (in Greek)

5.2. Tools and tests

Since there aren't off-the-shelf tools ready for the automated summarisation of judicial decisions in Greek, this methodology recommends developing them based on OpenAI's models (to establish a baseline reference of capabilities) and at the same time develop the tools for testing and scoring them. This will be done in two nested cycles (see the testing procedure below).

- Basic tool: in-prompt GPT-3.5.
  - Script for in-prompt conditioning + chunk splitting with GPT-3.5-turbo, simulated dialogue via APIs.
  - Same with GPT-4 (with the potential advantage of larger prompt windows)?
- Advanced: fine-tuned model
  - GPT-3 Davinci or Curie, respectively the most powerful model and second best of the series.
- Scoring tool

6. Financial considerations

6.1. Preliminary cost estimation

The following analysis will focus on Open AI's costs to establish a baseline against which to compare other possible solutions in the future.

OpenAI pricing, USD per 1k tokens

Model	Training	Usage	Prompt	Completion
GPT-3-Ada	0.0004	0.0016	n.a.	n.a.
GPT-3-Babbage	0.0006	0.0024	n.a.	n.a.
GPT-3-Curie	0.003	0.012	n.a.	n.a.
GPT-3-Davinci	0.03	0.12	n.a.	n.a.

					Ratio to GPT-3.5	
GPT-3.5-turbo	n.a.	0.002	n.a.	n.a.	Prompt	Completion
GPT-4-8k			0.03	0.06	15 x	30 x
GPT-4-32k			0.06	0.12	30 x	60 x

Cost of use of GPT-3 models (Ada...Davinci) does not differ between prompt ("question") and completion ("answer").

GPT-4 pricing in principle favours summarisation (long prompt, short completion).

GPT-3.5 and GPT-4 are models already tuned by OpenAI to function as chat bots and cannot be further fine-tuned.

Direct summarisation with in-prompt conditioning for GPT-3.5 and GPT-4

The Supreme Court of Cyprus provided the following examples for the average lengths (L, measured in characters) of the decisions (D) and corresponding summaries (S).

		L(D)	L(S)	L(D) / L(S)
P1	Civil appeal	15631	4379	3.57
P1	Civil appeal	14903	11665	1.28
P2	Criminal appeal	12694	3765	3.37
P2	Criminal appeal	4786	842	5.68
P3	Administrative appeal	10402	2044	5.09
P3	Administrative appeal	5556	2518	2.21

For a first order, rough approximation of the estimated length of the decisions, we take the average of the two data points provided by the Court for each part (civil, criminal, administrative cases).

The Court provided also data about the number of decisions to be published from the last three years:

	PART 1 (CIVIL)	Part 2 (CRIMINAL)	PART 3 (ADMINISTRATIVE)
2020	322	109	66
2021	418	145	75
2022	323	129	94

Combining the available information about the number of decisions for each kind with the first order approximations for the lengths of the decisions and related summaries, and OpenAI pricing per token we can get an initial cost estimation of using in-prompt conditioning with GPT-3.5 and GPT-4 models.

Year	**kTokens, all decisions**	**Ktokens, all summaries**	**Cost GPT-3.5 (4k)**	**Cost GPT-4-8k**	**Cost GPT-4-32k**
2020	5,461	2,549	$16	$317	$634
2021	7,043	3,295	$21	$364	$728
2022	5,814	2,650	$17	$327	$655

The GPT-3.5 model was meant as a chat-friendly but inexpensive version of Davinci and its low-cost combined with well-documented API calls that allowed to programmatically simulate a Q&A session immediately attracted the interest of developers (so much that OpenAI has recently introduced rate limitations on GPT-3.5 for free plans).

GPT-4's much higher running costs should be evaluated against its potentially better summarising quality, giving a possibility to fit most decisions in one context window (prompt=decision + completion=summary).

		L(D)	**T(D)**	**Fits in one context**		
		kChar	kTok	4k	8k	32k
P1	Civil appeal 60-2022 (longest)	43	36.7	No	No	No
P1	Civil appeal (avg.)	15	13.3	No	No	Yes

		L(D)	T(D)	Fits in one context		
		kChar	kTok	4k	8k	32k
P3	Administrative appeal 86-15	18	15.4	No	No	Yes
P1	Civil appeal 28-2022	5.2	4.4	No	Yes	Yes
P2	Criminal appeal 02-2022	4.6	3.9	almost*	Yes	Yes

* There's barely enough room for the prompt, but likely not enough for the completion.

6.2. Cost estimation for the Cyprus Law Reports

The issues of the Cyprus Law Reports that are already available in digital format (approximately one decade, 2008-2017) constitute a potential dataset ideally suited for fine-tuning a model (tool). This dataset has to be built from the original files (PDF, docx or HTML), but in principle it would include around 630 pairs (decision, summary) per year. All summaries have been drafted and validated by human experts in the past and therefore represent the gold standard of the output quality.

For the cost estimation the summary was considered as a single text, without breaking it into internal sections.

During the model fine-tuning both the prompt (the original decision) and the desired completion (the human-drafted summary) are fed to the model. As the Criminal Law Book (CLR) includes both summary and the original judgement for each case, the total number of characters in one issue is a good proxy of the size of the training dataset for each month.

	kChar.	**kTokens (GPT-3)**	**Davinci**	**Curie**
USD per kToken, training			0.03	0.003
Number of training epochs[5]			4	4
Fine-tuning costs for: one book	3150	3500	420	42

[5] Number of times the training dataset has to be fed to the model in order to fine-tune it.

	kChar.	kTokens (GPT-3)	Davinci	Curie
one year, or 5 books (3x civ, crim, adm)			2100	210
Full dataset (ten years)			21000	2100

Greek: 1 token ~	0.90	characters in GPT-3
English: 1 token ~	4	characters

These are the approximate costs for fine-tuning the model assuming that the full dataset is needed to achieve the desired standard of quality. It is possible that this could be accomplished by fine-tuning the model with a limited training dataset (e.g. data produced during three or five years, instead of ten).

There are additional costs for running a fine-tuned model, as OpenAI charges for both prompt and completion (at the same rate, unlike GPT-4 models that charge more for completions than prompts). In order to estimate the costs of running a fine-tuned model, we can take the same data for the number of decisions to be summarised from the years 2020-2022 and apply GPT-3 usage pricing for the model Davinci and Curie, as shown in the table below.

Year	kTokens, all decisions	kTokens, all summaries	Running costs GPT-3 Curie (4k)	Running costs GPT-3 Davinci (4k)
2020	5,461	2,549	$96	$960
2021	7,043	3,295	$124	$1,240
2022	5,814	2,650	$102	$1,020

A fine-tuned GPT-3 Davinci model – with a supposedly high quality of output requires an initial, one-off investment of 15 000 – 20 000 US dollars for the fine-tuning, plus one 1000 US dollars per year for its application for approximately 500 decisions that shall be summarised annually. GPT-3 Curie costs 1/10 of the costs of Davinci.

Even the most expensive GPT-4 model with 32k prompt window has lower running costs, zero fine-tuning costs (in fact, they are not fine-tuneable by design, so far), and very low training costs (limited to the inception phase while setting the system up, designing the best prompt and so on). Using the advantage of GPT-4 longer prompt window and well curated prompt to offset the lack of fine-tuning, assuming that the quality of output is comparable, the in-prompt conditioning seems the most cost-effective approach.

7. Testing procedure

The following process was followed for testing of the AI tools for summarisation of legal texts.

7.1. Step 1: Baseline for summarising: prompting with human expert feedback

At this stage preparatory work is done which is composed of the following steps.

- a. Develop (adapt) a simple summarising script including:
    i. Ingest the input file (decision) in either PDF or DOC/ODT or HTML to the AI tool
    ii. Set conditions of the task: describe the desired results
    iii. Split the text in (slightly overlapping) chunks and ingest them sequentially to the AI tool
    iv. Collate the responses received from the AI tool, and ingest them back to it as expanded context
    v. Task the AI tool to produce the summary according to an outline defined in the prompt
    vi. Ask further questions within-context (e.g. for clarification)

The following points shall be taken into consideration int his process:

Token-intensive task: each of the steps require tokens (as described above).

There are also limitations of the size of the window for the prompt ingestion, which are linked to the requirement of big amount of tokens as mentioned above.

So, for particularly long decisions it needs extra levels of nesting – additional layering of information.

- b. Run the script on a dataset composed of a series of texts pairs (D=decision, HS=human generated summary)
- c. Collect legal experts' feedback on the results (using the evaluation criteria proposed below)
    i. Legal experts compare the machine-generated summary with the human-generated one (existing in the dataset)
    ii. Legal experts score the machine-generated summary on a grid with the following criteria with specification of certain metric:
        - Total number of factual mistakes
        - (In)correct identification of relevant features:
            - Main conclusions of the court
            - Accuracy of the important facts
            - Correctness of the indication of the parties (including their referencing throughout the summary)
            - Thematic keywords
            - Grounds of appeal
            - Legal doctrine and precedent cited
            - Not confusing the lower Court decision with the Supreme Court's one
        - Relevant information missing from the machine-generated summary
- d. Improve the prompt (reformulate or adjust the task) on the basis of the score
- e. Reiterate until the human experts scoring of the results reach a constant level of sufficient quality[6] in the metric adopted.

[6] The level of the sufficient quality shall be agreed with the legal expert.

See Annex II. Training configuration diagram.

7.2. Step 2: Automated scoring tool development

Using the prompt and script developed at the previous stage, automatic testing running the tool on a dataset composed of the *same* pairs (D=decision, HS=human generated summary) as before

a. Produce machine generated summary (MS) using the tool above for each decision in the dataset
b. Compare the results received HS versus MS
c. Calculate a similarity score (e.g. semantic distance of embedding vectors of the whole text or after automated extraction of features) and perform a statistical analysis of the similarity scores
d. Compare the machine generated score with the evaluation of human experts
    i. If machine generated score < human generated score
        - Change the similarity score algorithm
    ii. If machine generated score = human generated score
        - The scoring algorithm has reached a stable functioning and can be used in the overall summary generation chain to evaluate the quality of the results.

7.3. Step 3: Run the summarising tool and the scoring tool on the dataset

This step shall allow validation of the automated scoring tool developed as step 2.

Human experts assess both the quality of the MS and scoring on a random sample of decisions.

The scoring tool can be applied to test various summarisation tools in order to find the most suitable one.

For more information, please see Annex III. Systematic testing configuration diagram.

8. Conclusions of pilot testing for the Supreme Court of Cyprus

8.1. Pondered comparison of the scoring of each tool produced during testing

The initial assessment focuses on the quality of the results, especially by using OpenAI's tools as a baseline against which to compare other tools.

Preliminary tests conducted before and during a TJENI workshop in Nicosia in May 2023 showed that GPT-3.5 capabilities in the summarisation of short decisions in English are close to human-like level. The width of the maximum context window remains the main constraint for the quality of summarisation. Judgments of the ECtHR that were used for testing are well-structured with a clear outline that maps to the corresponding sections of the summary: Facts, Law, Decision, Article 41 (satisfaction: damage and costs and expenses). This allows to partially relax the context length constraint, at least for decisions written in English.

The decisions of the Supreme Court of Cyprus are also implicitly well-structured, but the outline is not marked by headings, making automatic parsing (detection and reporting of errors or incomplete information) more difficult. The summary drafted by legal officers for the Cyprus Law Reports is usually longer than the ECtHR's. Overall, the negative combined effect of encoding and tokenization of the Greek script in the current OpenAI models (and probably also the smaller amount of Greek material originally used to build the models) reduces the context window to one fourth, thus greatly affecting the quality of the summaries even when the same structured section-by-section mapping is applied. Simple algorithmic approaches that divide the original text in chunks to be summarised one at a time while connecting them at a later stage has shown promise but overall inferior quality.

At the same time even with these limitations, OpenAI already sets a relatively high baseline quality against which to compare open-source alternative solutions.

8.2. Cost and quality analysis

While the section above focused on the costs of training and running OpenAI's models, it should not be assumed that alternative, open-source solutions would be free.

It is to be expected that while open-source solutions will have negligible running costs, they will entail substantial development costs, most probably hardware purchases needed to fine-tune open-source models.

In particular, the costs of fine-tuning an OpenAI's model like Davinci with the dataset based on the Cypriot Law Review materials – approximately 20 thousand euros – are of the same order of magnitude of the procurement of dedicated hardware (Graphics Processing Units/GPUs) for fine-tuning and running open-source models.

8.3. Open vs Proprietary

A tool developed for the use of a public institution should favour the open-source approach. In particular taking into account that "researchers in the open-source community, using free, online resources, are now achieving results comparable to the biggest proprietary models" (The Economist, 2023c). There are some sustainability concerns (Douglas Heaven, 2023), but this seems the way to go.

The summarisation of the judicial decision of the Supreme Court of Cyprus operates on documents that are already public and published online. This allows to eliminate the risks related to personal data. But in other use cases a stronger protection of personal data might be required, along with the need of locally developed and run AI applications.

8.4. Better tokenization

A tool that uses a custom tokenization algorithm based on more efficient encoding of the Greek script should be preferred *ceteris paribus* over a tool that uses standard, English-centred tokenization algorithms.

The use of more efficient tokenization for Greek should produce better results and as such it should be favoured.

References

Ananthaswamy, A. (2023, April 13). A New Approach to Computation Reimagines Artificial Intelligence. *Quanta Magazine*. https://www.quantamagazine.org/a-new-approach-to-computation-reimagines-artificial-intelligence-20230413/

Bashir, D. (2023, April 29). In-Context Learning, In Context. *The Gradient*. https://thegradient.pub/in-context-learning-in-context/

Chomsky, N. (2023, March 9). Opinion | Noam Chomsky: The False Promise of ChatGPT. *The New York Times*. https://web.archive.org/web/20230309052428/https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Committee on Artificial Intelligence (CAI). (2023). *Revised Zero Draft [framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Council of Europe. https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f

Council of Europe. (2023). *Council of Europe and Artificial Intelligence*. Council of Europe. https://www.coe.int/en/web/artificial-intelligence

Deutsche Welle. (2023, April 29). Italy lifts ban on ChatGPT after data privacy improvements. *Dw.Com*. https://www.dw.com/en/ai-italy-lifts-ban-on-chatgpt-after-data-privacy-improvements/a-65469742

Douglas Heaven, W. (2023, May 12). The open-source AI boom is built on Big Tech's handouts. How long will it last? *MIT Technology Review*. https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Harari, Y. N. (2023, April 28). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist*. https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

Heikkilä, M. (2023, May 23). Suddenly, everyone wants to talk about how to regulate AI. *MIT Technology Review*. https://www.technologyreview.com/2023/05/23/1073526/suddenly-everyone-wants-to-talk-about-how-to-regulate-ai/

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Lambert, N. (2022, December 9). Illustrating Reinforcement Learning from Human Feedback (RLHF). *Hugging Face*. https://huggingface.co/blog/rlhf

McGlauflin, P. (2023, May 19). Apple, Goldman Sachs, Samsung, and 10 others clamp down on ChatGPT at work. *Fortune*. https://fortune.com/2023/05/19/chatgpt-banned-workplace-apple-goldman-risk-privacy/

OpenAI. (2023, May). *How your data is used to improve model performance*. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

Raschka, S. (2023, April 22). *Understanding Large Language Models*. https://magazine.sebastianraschka.com/p/understanding-large-language-models

Sobha Kartha, N. (2021, November 12). Explain Yourself—A Primer on ML Interpretability & Explainability. *The Gradient*. https://thegradient.pub/explain-yourself/

The Economist. (2023a, April 20). How to worry wisely about artificial intelligence. *The Economist*. https://www.economist.com/leaders/2023/04/20/how-to-worry-wisely-about-artificial-intelligence

The Economist. (2023b, April 23). Large language models' ability to generate text also lets them plan and reason. *The Economist*. https://www.economist.com/science-and-technology/2023/04/19/large-language-models-ability-to-generate-text-also-lets-them-plan-and-reason

The Economist. (2023c, May 11). What does a leaked Google memo reveal about the future of AI? *The Economist*. https://www.economist.com/leaders/2023/05/11/what-does-a-leaked-google-memo-reveal-about-the-future-of-ai

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wind, J. (2023, March 23). *The RWKV language model: An RNN with the advantages of a transformer*. https://johanwind.github.io/2023/03/23/rwkv_overview.html

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond* (arXiv:2304.13712). arXiv. https://doi.org/10.48550/arXiv.2304.13712

Zewe, A. (2023, February 7). Solving a machine-learning mystery. *MIT News | Massachusetts Institute of Technology*. https://news.mit.edu/2023/large-language-models-in-context-learning-0207

Žižek, S. (2006). *How to read Lacan*. Granta.

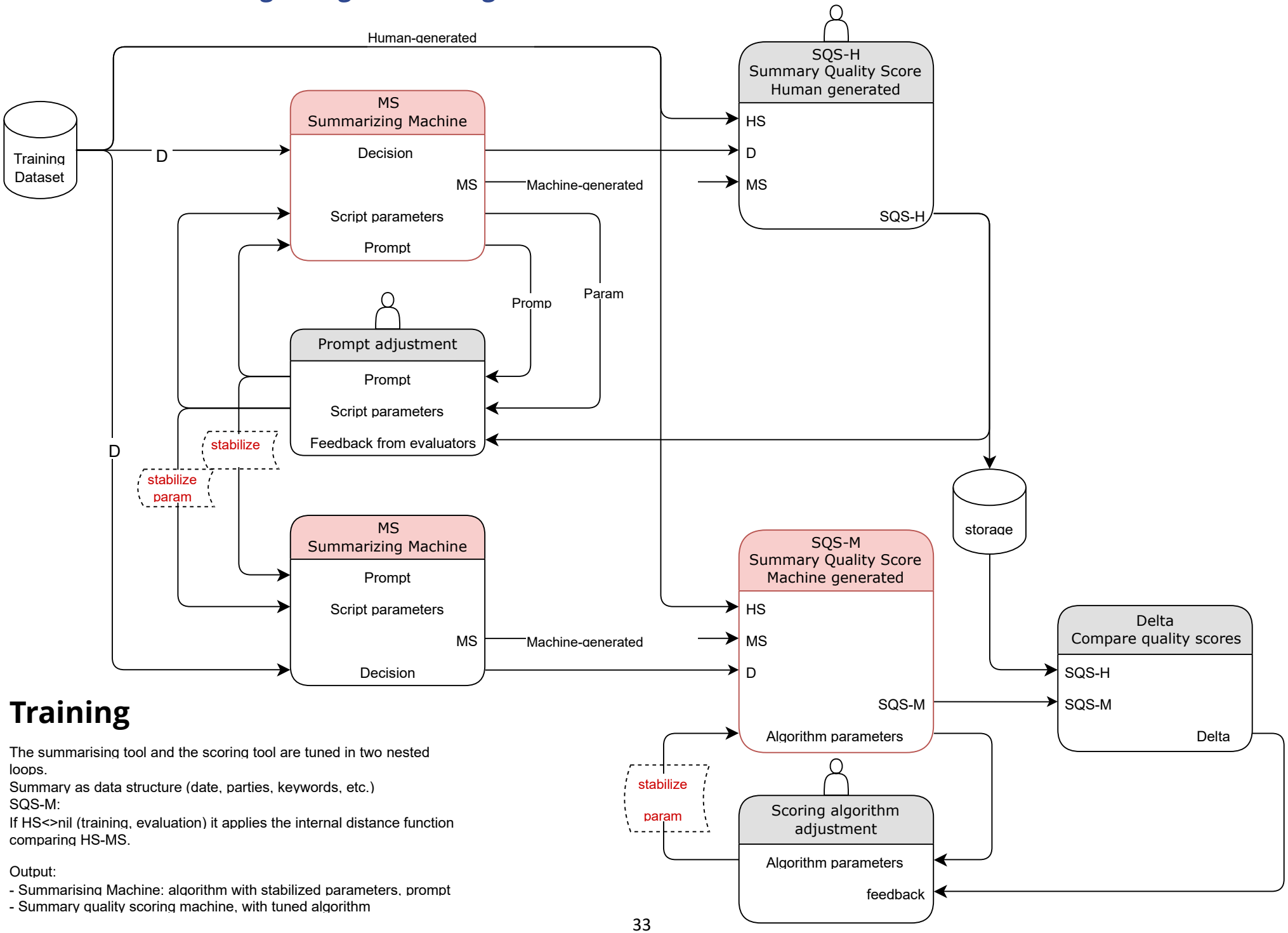
Annex I. Cost estimation with GPT-3.5 and GPT-4

Cost estimation of annual in-prompt summarisation of the Cyprus Law Reports with GPT-3.5 and GPT-4 (via API).

Year	Part 1 (CIVIL)			Part 2 (CRIM)			Part 3 (ADMIN)			All parts, per year					
	Cases	L(D)	L(S)	Cases	L(D)	L(S)	Cases	L(D)	L(S)	C/T	kTok. D	kTok. S	Cost GPT-3.5 (4k)	Cost GPT-4-8k	Cost GPT-4-32k
2020	322	15267	8022	109	8740	2303	66	7979	2281	1.17	5,461	2,549	$16	$317	$634
2021	418	15267	8022	145	8740	2303	75	7979	2281	1.17	7,043	3,295	$21	$364	$728
2022	323	15267	8022	129	8740	2303	94	7979	2281	1.17	5,814	2,650	$17	$327	$655

L: length
D: Decision
S: Summary
C: number of characters in the text
T: number of tokens in the text
C/T: estimated from a sample of six decisions in Greek using OpenAI's *tiktoken* library and the *cl100k_base* encoder.

Annex II. Training configuration diagram

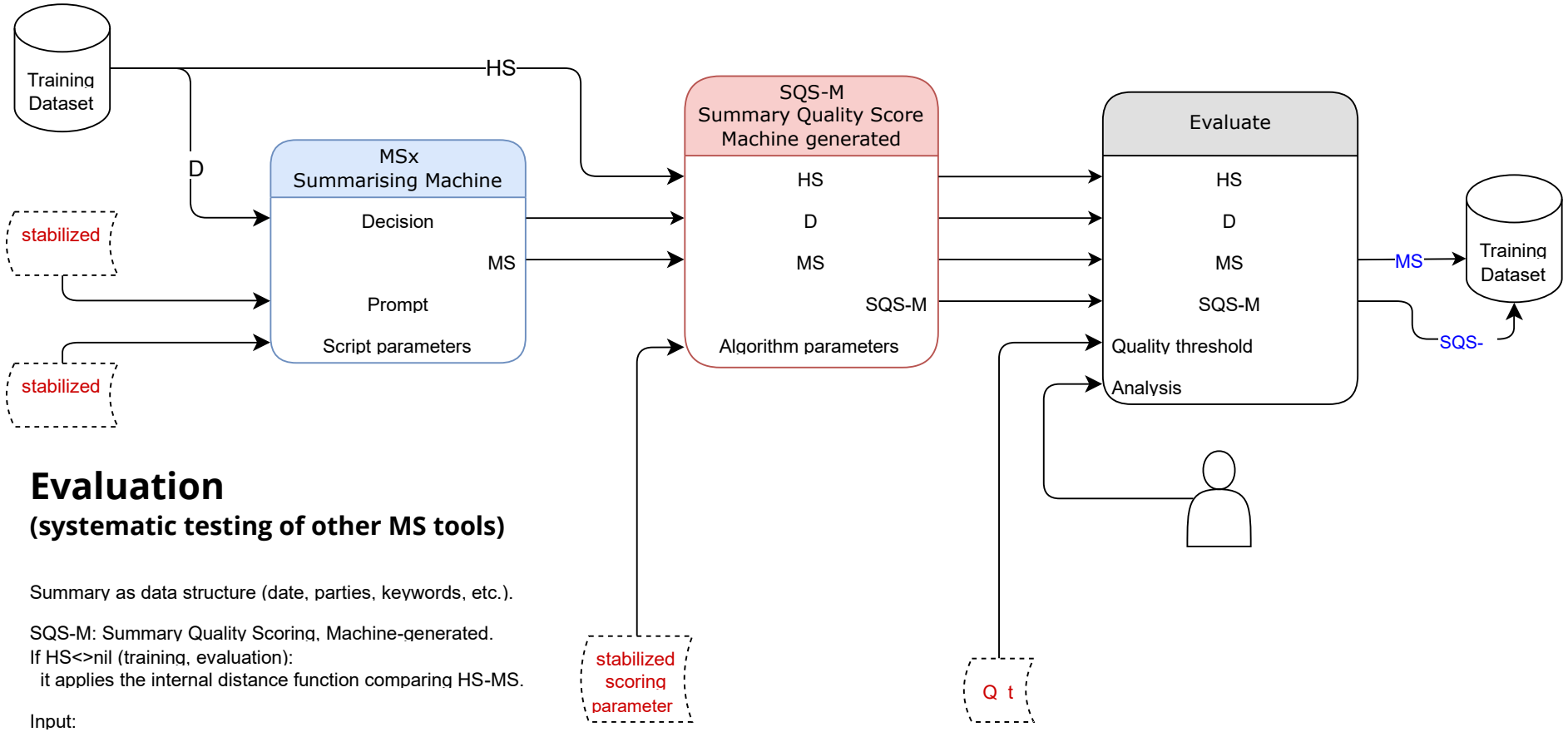


Training

The summarising tool and the scoring tool are tuned in two nested loops.
Summary as data structure (date, parties, keywords, etc.)
SQS-M:
If HS<>nil (training, evaluation) it applies the internal distance function comparing HS-MS.

Output:
- Summarising Machine: algorithm with stabilized parameters, prompt
- Summary quality scoring machine, with tuned algorithm

Annex III. Systematic testing configuration diagram



Evaluation

(systematic testing of other MS tools)

Summary as data structure (date, parties, keywords, etc.).

SQS-M: Summary Quality Scoring, Machine-generated.
If HS<>nil (training, evaluation):
it applies the internal distance function comparing HS-MS.

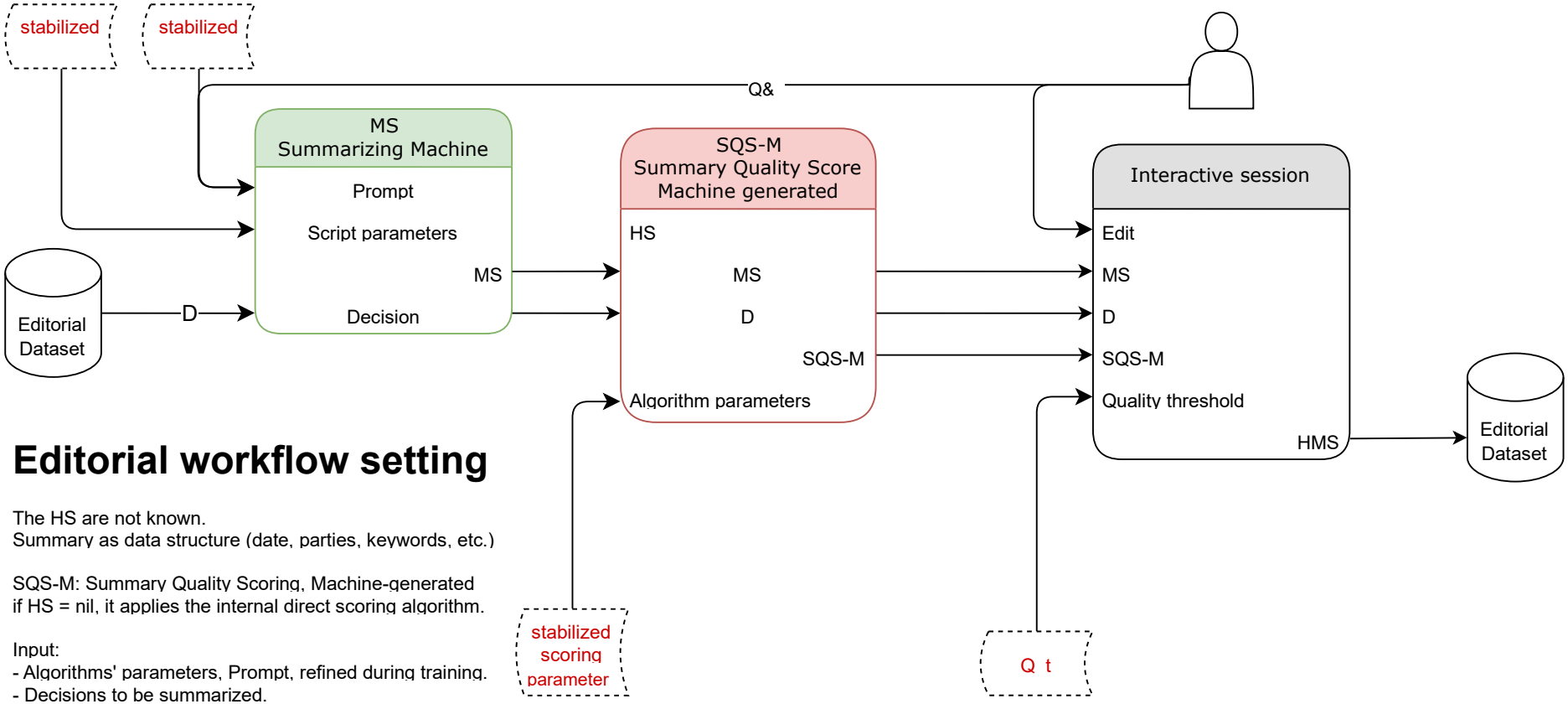
Input:
- Algorithms' parameters, Prompt, refined during training.
- Decisions to be summarised.
- Human-generated summaries.
- Quality threshold.
- Human supervision.

Output:
- Machine generated summaries.
- Scoring data on the tool being tested.

Annex IV. Usage scenario in an editorial workflow

Although out of the scope of this methodology focussed on testing candidate tools to do automated summarisation of judicial decisions in Greek, the final diagram illustrates the concept of a possible usage scenario one a tool has been tested, evaluated and chosen for an editorial workflow.

The legal officers would be able to control the process, supported by both the automatic summariser and scoring tools. Crucially, they would be able to interact with the language model via a custom user prompt in a Q&A session to complement the information already present in the machine-generated summary.



Editorial workflow setting

The HS are not known.
Summary as data structure (date, parties, keywords, etc.)

SQS-M: Summary Quality Scoring, Machine-generated
if HS = nil, it applies the internal direct scoring algorithm.

Input:
- Algorithms' parameters, Prompt, refined during training.
- Decisions to be summarized.
- Quality threshold.
- Human supervision and real-time interaction.

Output:
- Human-edited, Machine-generated Summaries