ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme

Claudiu D. Hromei[1], Danilo Croce[1], Valerio Basile[2] and Roberto Basili[1]

[1]Università degli Studi di Roma Tor Vergata, Italy
[2]Università di Torino, Italy

Abstract

This paper explores the potential application of a monolithic neural model for all tasks in EVALITA 2023. We evaluated two models: `extremIT5`, an encoder-decoder model, and `extremITLLaMA` an instruction-tuned Decoder-only Large Language Model, specifically designed for handling Italian instructions. Our approach revolves around representing tasks in natural language, where we provide instructions to the model using prompts that define the expected responses. Remarkably, our best-performing model achieved first place in 41% of the subtasks and showcased top-three performance in 64%. These subtasks encompass various semantic dimensions, including Affect Detection, Authorship Analysis, Computational Ethics, Named Entity Recognition, Information Extraction, and Discourse Coherence.

1. Introduction

In recent years, Large Language Models (LLMs) have garnered substantial attention due to their remarkable performance across a wide range of NLP tasks. In addition to achieving state-of-the-art results in individual tasks, LLMs such as T5 [1], mT5 [2], IT5 [3], and FlanT5 [4] have demonstrated exceptional capabilities in solving various tasks individually and collectively through multi-task training paradigms.

In parallel, the generative power exhibited by models like GPT [5] and GPT3 [6], as well as the recent development of LLaMA [7] foundational models, has opened up new avenues for leveraging the concept of "prompting". This approach allows for modeling inductive tasks by formulating them linguistically, as natural language queries or instructions, enabling the model to provide accurate responses based on it. By combining the power of these models and the prompting technique, complex tasks can be addressed using a straightforward and intuitive interaction paradigm, without the need for task-specific feature engineering or neural architectures.

This paper presents `ExtremITA`, our approach developed for the EVALITA challenge [8]. The purpose of this work is to investigate how the adoption of a Large Language Model (LLM) can be taken to its extreme consequences by proposing a single model capable of tackling a wide array of heterogeneous tasks. Our methodology leverages an Encoder-Decoder model and a Decoder-only model, both trained on the union of all the available datasets for the challenge. By adopting a multi-task learning framework, our objective is to evaluate the applicability of a single model in effectively solving multiple tasks at once. Notably, our approach offers a significant advantage as it enables the resolution of diverse tasks by employing a unified architecture and fine-tuning based on input-output pairs. The EVALITA challenge serves as a robust testing ground for assessing the capabilities of LLMs across various Italian linguistic tasks, without any specific architectural requirement. Instead, we will trigger the model with task-specific prompts, such as "*Is there any mention of a conspiracy in this text? Answer yes or no.*" or "*How much consistent is this sentence, on a scale of 0 to 5?*".

The complete list of tasks in which the `ExtremITA` approach participated is here reported, in a wide range of semantic dimensions, including Affect Detection, Authorship Analysis, Computational Ethics, Named Entity Recognition, Information Extraction, and Discourse Coherence: *i)* EMit – Categorical Emotion Detection in Italian Social Media [9]; *ii)* EmotivITA – Dimensional and Multi-dimensional Emotion Analysis [10]; *iii)* PoliticIT – Political Ideology Detection in Italian Texts [11]; *iv)* GeoLingIt – Geolocation of Linguistic Variation in Italy [12]; *v)* LangLearn – Language Learning Development [13]; *vi)* HaSpeeDe 3 – Political and Religious Hate Speech Detection [14]; *vii)* HODI – Homotransphobia Detection in Italian [15]; *viii)* MULTI-Fake-DetectiVE – MULTImodal Fake News Detection and VErification [16]; *ix)* ACTI – Automatic Conspiracy Theory Identification [17, 18, 19]; *x)* NERMuD - Named-Entities Recognition on Multi-Domain Documents [20]; *xi)* CLinkaRT – Linking a Lab Result to its Test Event in the Clinical Domain [21]; *xii)* WiC-ITA – Word-in-Context task for Italian [22]; *xiii)* DisCoTEX – Assessing DIScourse COherence in Italian TEXts [23].

The aforementioned 13 tasks comprised 22 subtasks, where the proposed models ranked first in 9 subtasks (41%), and achieved a top-three position in 14 subtasks (64%). The adopted LLMs (especially LLaMA-based) proposed solution strongly supports the viability and high

CEUR Workshop Proceedings (CEUR-WS.org)

performance of a single monolithic architecture, as it only requires modeling the tasks in natural language using prompts. This approach has been further reinforced by recent work [24], which indicates the same direction.

In the rest of the paper, Section 2 describes the adopted LLMs. Section 3 provides the results, accompanied by a brief error analysis. Finally Section 4 derives the conclusions.

2. Multi-task prompting in ExtremITA

The Transformer architecture [25] can be divided into two main components, each giving rise to distinct families of models. The encoder, exemplified by BERT [26], RoBERTa [27], and DeBERTa [28], is responsible for encoding input sequences and generating meaningful representations (embeddings) using the self-attention mechanism. On the other hand, the decoder, represented by models like GPT [5], GPT3 [6], and LLaMA [7], generates output sequences in an auto-regressive manner based on the input and previously generated output tokens. Additionally, another family of models, the Encoder-Decoder models, such as T5 [1] and BART [29], combine the strengths of both encoder and decoder components. These models maintain the integration of the two aforementioned blocks and they are usually used in tasks like machine translation, summarization, and question-answering, where complex input understanding as transduction is required.

A first effective application of an Encoder-Decoder architecture in a multi-task scenario is presented in [1]: in particular, the pre-training process of the so-called T5 involves training the model on a large corpus of diverse text data, which consists of a wide range of sources such as books, articles, and websites, but also texts involved in machine translation, classification and regression tasks. During pre-training, T5 utilizes a denoising objective, similar to other popular Transformer-based models like BERT and GPT. The model is trained to reconstruct masked or corrupted input text, which helps it learn meaningful representations and capture contextual information. One of the key strengths of T5 is its versatility. By casting various NLP tasks into a text-to-text format, it can be fine-tuned on a specific task simply by providing a prefix that serves as a description of the task and appropriate input-output pairs during fine-tuning. In practice, such an architecture can be triggered by concatenating the name of the task it is trained on with an input text, and it generates in output the expected solution to the task, e.g., a class label in a classification task or a text span that answers to a question. This flexibility eliminates the need for task-specific architectures or modifications, making it easier to apply T5 to different scenarios. Recently, this model was applied to hundreds of tasks in [24], while in [4] a systematic pre-training at large scale demonstrates the effectiveness within "zero-shot" or "few-shot" learning scenarios. In this paper, the first approach we adopted is based on T5, pre-trained on Italian texts, namely IT5 [3].

On the other hand, Decoder models are typically trained to be triggered by text, such as a natural language request or a piece of text intended for processing. These models generate text one word at a time, producing an output that can be an answer to a question or a solution to the given tasks or requests. Such models have the ability to essentially follow instructions, as exemplified by the recent release of ChatGPT. This characteristic holds a greater appeal, as tasks can be linguistically described using prompts, where the input sentence serves as contextual information. InstructGPT [30] is an extension of the GPT [6] language model explicitly designed to excel in multi-task scenarios when used with prompts. It combines the power of language models with the ability to follow instructions provided in the form of natural language prompts. Unlike conventional language models that generate text freely, InstructGPT is fine-tuned using human feedback to understand and generate text based on a given prompt and to select the best sequence that humans would have preferred. Another language model that adopts this instruction-tuning technique is Alpaca [31], which builds upon the LLaMA [7] foundational models. In the case of Alpaca, the authors created 175 sets of instructions, input sentences, and corresponding outputs. These were then used to generate variations using GPT 3.5, resulting in a collection of approximately $52,000$ instruction examples. The LLaMA model was further fine-tuned using this extensive dataset, a process referred to as instruction-tuning. The outcome of this effort was the Stanford Alpaca [31] as an instruction-following LLaMA model. More recently, an Italian counterpart called Camoscio [32] has overgone a similar intruction-tuning to Alpaca but on Italian data, essentially serving as the Italian equivalent. It is based on the same LLaMA model and it was instruction-tuned on the 52.000 instructions that were automatically translated into Italian using ChatGPT as in [32]. As the size of these models continues to grow, reaching trillions of parameters, there is a need for a way to fine-tune them effectively using modest GPU resources. The technique adopted in this paper is called Low-Rank Adaptation (LoRA [33]). LoRA involves freezing the weights of the pre-trained model and introducing trainable rank decomposition matrices into each layer of the Transformer architecture. This approach significantly reduces the number of trainable parameters for downstream tasks while avoiding additional inference latency.

To summarize, the ExtremITA approach for the EVALITA challenge focuses on efficiently modeling all

Task	Output Templates
EMit A	{"*Rabbia*", "*Anticipazione*", "*Disgusto*", "*Paura*", "*Gioia*", "*Amore*", "*Tristezza*", "*Sorpresa*", "*Fiducia*"}$^{+} \vee$ "*Neutrale*"
EMit B	{"*Direzione*", "*Argomento*", "*Entrambi*", "*Non specificato*"}
EmotivITA	"*Valenza: {0-5} Stimolo: {0-5} Controllo: {0-5}*"
PoliticIT	"*Gender*: {"*Uomo*", "*Donna*"} *PIB*: {"*Sinistra*", "*Destra*"} *PIM*: {"*Sinistra*", "*Destra*", "*Centro Sinistra*", "*Centro Destra*"}"
GeoLingIt	"*Regione*: {*Abruzzo*, .., *Veneto*} *Latitudine*: {} *Longitudine*: {}"
LangLearn*	{"*Corretto*", "*Non Corretto*"}
HaSpeeDe 3*	{"*Odio*", "*Non Odio*"}
HODI A*	{"*Omotransfobico*", "*Non Omotransfobico*"}
HODI B	<HOMOTRANSPHOBIA_MENTION>
MULTI-Fake -DetectiVE	{"*Certamente Falso*", "*Probabilmente Falso*", "*Probabilmente Vero*", "*Certamente Vero*"}
ACTI A*	{"*Cospirazione*", "*Non Cospirazione*"}
ACTI B	{"*Terrapiattista*", "*Covid*", "*Qanon*", "*Russia*"}
NERMuD	[<ENTITY_TYPE>] <TEXT_SPAN_THAT_EVOKES_ENTITY>
CLinkaRT	"[*BREL*] <RML_ENTITY_MENTION> [*SEP*] <EVENT_ENTITY_MENTION> [*EREL*]"
WiC-ITA*	{"*Uguale*", "*Differente*"}
DisCoTEX 1*	{"*Coerente*", "*Non Coerente*"}
DisCoTEX 2	{*0-5*}

Table 1
Output templates for ExtremITA models. In EMit A the model is requested to generate one or more labels from the first set ($^{+}$) or the text "*Neutrale*" if no emotion is expressed. In the tasks with * the extremITLLaMA model is requested to respond with {"*Si*", "*No*"}, for more details see Table 2.

available tasks using a single monolithic architecture, based on two independently tested models:

- extremIT5, An Encoder-Decoder model, based on IT5[1], consisting of approximately 110 million parameters. This model is trained by concatenating the name of the task and the input sentence/-paragraph in the input texts, each representing an example from a generic EVALITA task. Its purpose is to generate a piece of text that solves the target task.
- extremITLLaMA, an instruction-tuned Decoder-only model, built upon the LLaMA foundational models[2], with a total of 7 billion parameters. The initial model was trained using the LoRA technique on Italian translations[3] of Alpaca instruction data. This training enables the model to comprehend instructions in Italian. After training the adapters, they are merged into the original model to create an instruction-based model (using the "merge" procedure from [33]). Finally, this model is further fine-tuned using LoRA on instructions that reflect the EVALITA task. For each example from EVALITA, an input text is paired with a manually crafted question that simulates an instruction to be solved, accurately representing the specific task.

The next section describes how the 22 subtasks in EVALITA are encoded as prompts to fine-tune the above architectures.

Prompt Engineering in ExtremITA. The approach employed in this study draws inspiration from the original T5 and IT5 methodologies. Similar to those approaches, each training example is converted into a text-to-text format.

The model called extremIT5 is trained as a generic T5 model. In input, each example for an individual task is given to the neural architecture as concatenated after the task name. As an example, in the ACTI A task [17, 19] the input is just *"ACTI: Hanno votato tutti obbligo vaccinale, green pass, persecuzioni varie"*.

The output depends on the specific task. For a comprehensive compilation of outputs for the ExtremITA models, please refer to Table 1. In classification tasks involving only one label (such as EMit B, LangLearn, HaspeeDe 3, HODI A, MULTI-Fake-DetectiVE, ACTI A, ACTI B, WiC-ITA and DisCoTEX 1) the output is just the label of the target class. In the above example, the output would be "*Cospirazione*" as the input text reflects some conspiracy theory. In some tasks, such as PoliticIT [11], where a text is expected to be associated with the gender and the political inclination of the author, multiple labels reflecting such different dimensions are used, e.g., "*uomo sinistra centro-sinistra*". In EMit A [9] where multiple emotions can be triggered, these are provided as a sequence of labels. In regression tasks, such as EmotivITA [10] and DisCoTEX 2 [23], the output is the number to be

[1] https://huggingface.co/it5/it5-efficient-small-el32
[2] https://huggingface.co/decapoda-research/llama-7b-hf
[3] https://github.com/teelinsan/camoscio/tree/main/data

Task Name	Natural language instruction
EMit A	"*Quali emozioni sono espresse in questo testo? Puoi scegliere una o più emozioni tra 'rabbia', 'anticipazione', 'disgusto', 'paura', 'gioia', 'amore', 'tristezza', 'sorpresa', 'fiducia', o 'neutro'.*"
EMit B	"*Di cosa parla il testo, tra 'direzione', 'argomento', 'entrambi', 'non specificato'?*"
EmotivITA	"*Scrivi quanta valenza è espressa in questo testo su una scala da 1 a 5, seguito da quanto stimolo è espresso in questo testo su una scala da 1 a 5, seguito da quanto controllo è espresso in questo testo su una scala da 1 a 5.*"
PoliticIT	"*Scrivi se l'autore del testo è 'uomo' o 'donna', seguito dalla sua appartenenza politica tra 'destra', 'sinistra', 'centrodestra', 'centrosinistra'.*"
GeoLingIt	"*Scrivi la regione di appartenenza di chi ha scritto questo testo, seguito dalla latitudine, seguita dalla longitudine.*"
LangLearn	"*Questi due testi separati da [SEP] sono presentati nell'ordine in cui sono stati scritti? Rispondi sì o no.*"
HaSpeeDe 3	"*In questo testo si esprime odio? Rispondi sì o no.*"
HODI A	"*In questo testo si esprime odio omotransfobico? Rispondi sì o no.*"
HODI B	"*Con quali parole l'autore del testo precedente esprime odio omotransfobico? Separa le sequenze di parole con [gap].*"
MULTI-Fake -DetectiVE	"*L'evento riportato nel testo è 'certamente vero', 'probabilmente vero', 'probabilmente falso', o 'certamente falso'?*"
ACTI A	"*In questo testo si parla di una cospirazione? Rispondi sì o no.*"
ACTI B	"*Di quale teoria cospirazionista parla questo testo, tra 'Covid', 'Qanon', 'Terrapiattista', 'Russia'?*"
NERMuD	"*Scrivi le menzioni di entità nel testo, indicandone il tipo: [PER] (persona), [LOC] (luogo), [ORG] (organizzazione).*"
CLinkaRT	"*Trova i risultati dei test e delle misurazioni nel testo. Per ogni risultato, scrivi '[BREL]', seguito dal risultato seguito da '[SEP]', seguito dal test, seguito da '[EREL]'. Se non trovi nessun risultato, scrivi '[NOREL]'.*"
WiC-ITA	"*La parola compresa tra [TGTS] e [TGTE] ha lo stesso significato in entrambe le frasi? Rispondi sì o no.*"
DisCoTEX 1	"*Le due frasi precedenti, separate da '[SEP]', sono coerenti tra loro? Rispondi sì o no.*"
DisCoTEX 2	"*Quanto è coerente questa frase, su una scala da 0 a 5?*"

Table 2
List of the natural language instruction definition for all tasks for the `extremITLLaMA` model. Notice that these instructions have not been heavily optimized against individual tasks, also due to time constraints during the EVALITA challenge.

predicted within a specific range. In GeoLingIt [12], the models are requested to determine the region of origin of the tweet and the corresponding coordinates (latitude and longitude) based solely on the text. For instance, for the `extremIT5` model the task name ("GeoLingIt") is provided, while for `extremITLLaMA` a more in detail prompt is given: "*Scrivi la regione di appartenenza di chi ha scritto questo testo, seguito dalla latitudine, seguita dalla longitudine.*". For example, if the input sentence is "*Daje che je 'a famo!*", the model should provide the answer "*Lazio 41.8984164 12.54514535*", considering the use of the typical Roman dialect. This particular task combines both multi-label classification and regression, as it requires determining the region (classification) and providing the precise coordinates (regression) simultaneously. In HODI B [15] where the span of the offending text is expected to be extracted, it is simply provided as output. In NERMuD [20], the list of expected Named Entities is reported as a sequence of text spans, each associated with the corresponding entity type. CLinkaRT [21] focuses on extracting the names of medical tests performed on patients from an input text and linking them to the corresponding test results, treating it as a Relation Extraction problem. Here the relations are encoded with a slightly more complex form to summarize a list of relations, each associating an EVENT with a corresponding measure (or RML); as an example, the sentence "*CLinkaRT*: *Il PSA aumentava da 2 a 62 ng/ml.*" is associated with "[*BREL*] *2* [*SEP*] *PSA* [*EREL*] [*BREL*] *62 ng/ml* [*SEP*] *PSA* [*EREL*]" (where 2 and 62 reflect the RML while PSA is the test event).

In contrast, as `extremITLLaMA` is pre-trained to execute instructions, it leverages a structured prompt, which comprises the textual description of the task and the specification of the desired output format. For instance, when applied to the ACTI task, the instruction provided is "*In questo testo si parla di una cospirazione? Rispondi sì o no.*". The subsequent sentence to be evaluated is appended to this instruction. A comprehensive list of such instructions can be found in Table 2.

The decoder is thus expected to continue the sentence by generating the answer. In general, the same answers used in `extremIT5` are adopted. The only exception concerns the following binary classification tasks (LangLearn, HaSpeeDe 3, HODI A, ACTI A, WiC-ITA and DisCOTEX 1) where the instruction is only expected to answer *yes* or *no*, to reduce data sparseness.

3. Experimental Results

Experimental Setup. Models were trained using PyTorch, the Huggingface library and the Peft packages to implement the LoRA technique. Both models were trained on the unified dataset of all the tasks of EVALITA. Generally, one example in an EVALITA task corresponds to an example in our learning setting. Below are some exceptions. The dataset for the ACTI task was expanded by

incorporating some[4] sentences from dataset B and vice versa, resulting in an increase in the number of examples from 460 to 1,909 for ACTI A and from 300 to 777 for ACTI B.

Since in CLinkaRT only (long) documents were made available, these medical reports were segmented into smaller parts with a minimum of 50 characters and a maximum of 30 words using the Spacy library, respecting sentence boundaries. Moreover, we augmented this dataset with examples derived from the dataset made available in TESTLINK@IberLEF 2023[5] that contains medical reports in Spanish: although of a different language, these texts contain similar phenomena about events and measures that are generally language invariant and were useful to augment the dataset. This process significantly augmented the dataset, expanding it from 83 large documents to 3,903 shorter sentences. In general, this process recovered more than 95% of annotated relations. In the case of EMit, the dataset underwent a transformation where emoji representations were converted into textual descriptions, enhancing its compatibility with language models. GeoLingIt was modified to solve task A and task B simultaneously, enabling a single prediction for both tasks. For HODI B, only sentences expressing homotransphobia were considered, resulting in a reduction from 5,000 to 1,914 examples. The dataset of the LangLearn task was truncated into sentences with a maximum of 100 tokens, and additional examples with inverted sentence pairs (by flipping the label from positive to negative and vice versa), augmenting the dataset from 3,377 to 6,438 examples. In MULTI-Fake-DetectiVE we neglected images, and duplicate examples were removed (i.e., same text and different image), leading to a decrease from 1,058 to 860 examples. NERMuD was transformed into a sequence-to-sequence task from its original token classification format. In PoliticIT, each text was divided into sentences with a maximum length of 200 tokens, enabling more manageable input for language models. At classification time, a voting strategy was applied to select the final class about gender and political ideas, grouping all sentences written by the same author. Lastly, the WiC-ITA dataset was expanded by including examples[6] with inverted sentence pairs while preserving the same label, resulting in an increase from 5,610 to 6,600 examples. Overall, the entire dataset is composed of a total of 134,018 examples.

The `extremIT5` model underwent 10 epochs of training with a learning rate of $2 \cdot 10^{-5}$, while the `extremITLLaMA` model underwent 2 epochs of training with a learning rate of $3 \cdot 10^{-4}$. The models employed a batch size of 64 for `extremIT5` and 32 for `extremITLLaMA`. To optimize the models' performance, a linear scheduler with warmup was applied, utilizing a warmup ratio of 0.1. The `extremITLLaMA` model's training process utilized LoRA to refine the W_q, W_k, W_v and W_o modules of the transformer (for more details please refer to the original paper [33]), incorporating a matrix rank $R = 8$ and a parameter $\alpha = 16$ for the LoRA matrices. The decoding strategy in the generation phase used a beam search equal to 4, temperature of 0.2, with a top probability of 0.75 amongst the first 40 candidates. Two Tesla T4 GPUs with 16GB of memory each were used in parallel. This was particularly beneficial for the `extremITLLaMA` model, as its training duration exceeded 144 hours. The training data was divided into a 95% training set and a 5% validation set initially for hyper-parameter optimization. We release the source code on GitHub[7] for reproducing the experiment and dataset generation.

Results Discussion. The experimental results are reported in Table 3. We presented the tasks categorized by sub-task, followed by the Evaluation Metric, and the scores and ranks achieved by our `extremIT5` model, `extremITLLaMA` model, and the best competitor. The best-performing method for each subtask is highlighted in bold. Our systems, particularly `extremITLLaMA`, ranked first in 9 out of 22 subtasks (i.e., the 41% of subtasks) in EVALITA 2023. Additionally, it ranks in the top-three position in 14 subtasks, i.e., 64% of all tasks. However, we faced challenges in tasks such as GeoLingIt, LangLearn, and WiC-Ita, where our monolithic architectures demonstrated its limitations. These tasks specifically require a system to detect and analyze changes in the author's writing style or the contextual meaning of words. Our models are primarily designed for sentence classification or rewriting spans of input text to justify previous decisions (e.g., HODI).

There are also important considerations regarding the computational cost of both training and inference. Training `extremIT5` (made of "only" 110 million parameters) required approximately 12 hours on the entire EVALITA dataset, while `extremITLLaMA` (made of 7 billion parameters) took over 144 hours. In terms of inference, `extremITLLaMA` processes only 2 or 3 sentences per second, whereas `extremIT5` handles over almost one hundred sentences per second. This significant difference in processing speed makes the `extremITLLaMA` model less practical, despite its superior performance across a wide range of tasks. Additionally, the number of parameters between the two models differs by one order of magnitude, with `extremITLLaMA` having 7 billion parameters compared to `extremIT5`'s 110 million.

Overall, the above results are quite impressive, espe-

[4] Only the positive examples, i.e. the ones that involved any conspiracy theory, are added from the dataset A to B or viceversa.

[5] https://e3c.fbk.eu/testlinkiberlef

[6] Only the positive examples underwent sentence order flipping in order to rebalance the class distribution.

[7] https://github.com/crux82/ExtremITA

Task	SubTask	Eval metric	extremIT5		extremITLLaMA		Best Competitor	
			Score	R	Score	R	Score	R
Emit	A	F1	0.5086	2	**0.6028**	**1**	0.4994	3
	B	F1	0.6331	2	**0.6459**	**1**	0.6184	3
EmotivITA	B	Pears Val	0.7080		**0.8110**		0.8110	
		Pears Aro	0.4300	4	**0.6330**	**1**	0.6520	2
		Pears Dom	0.5480		**0.6300**		0.6540	
PoliticIT	-	F1	0.7034	7	0.7719	3	**0.8241**	**1**
GeoLingIt	A	F1	0.3999	10	0.3818	11	**0.6630**	**1**
	B	Avg Km	126.1	7	145.15	9	**97.74**	**1**
LangLearn	COWS	F1	0.1600	10	0.5500	8	**0.7500**	**1**
	CITA	F1	0.4100	10	0.6100	8	**0.9300**	**1**
HaSpeeDe 3	A	F1 - text.	0.9079	2	0.9034	3	**0.9128**	**1**
		F1 - context.	0.9079	2	0.9034	3	**0.9128**	**1**
	B	F1 - xRel.	0.5921	4	**0.6525**	**1**	0.6461	2
		F1 - xPolitic.	0.9079	2	0.9034	3	**0.9128**	**1**
HODI	A	F1	0.7431	10	0.7942	5	**0.8108**	**1**
	B	F1	0.6598	4	**0.7228**	**1**	0.7051	2
Multi-Fake-Detective	A	F1	na	na	0.5070	2	**0.5120**	**1**
	ATD	F1	0.3480	3	**0.4640**	**1**	0.4600	2
ACTI	A	F1	0.8183	7	0.8565	2	**0.8571**	**1**
	B	F1	0.8057	7	0.8556	5	**0.9123**	**1**
NERMUD	DAC	F1	0.8300	2	**0.8900**	**1**	na	3
CLinkaRT	-	F1	0.3382	4	0.5916	2	**0.6299**	**1**
Wic-Ita	A	F1 it-it	0.6100	5	0.5100	10	**0.7300**	**1**
		F1 it-en	0.6200	4	0.5400	8	**0.7400**	**1**
	B	F1 all	0.6100	5	0.5100	10	**0.7300**	**1**
DisCoTEX	1	Acc	0.7000	3	**0.8150**	**1**	0.7200	2
	2	HM*	0.0600	4	**0.6500**	**1**	0.6300	2

Table 3
ExtremITA ranks and results. Here each task is divided into the subtasks we participated in. Our models reported are `extremIT5` and `extremITLLaMA` and as a comparison the best competitor (either that won or placed higher in the ranking). In bold the rank and the scores of the winning systems. The HM* measure for the Discotex task refers to the Harmonic Mean between Pearson's and Spearman's correlations.

cially when considering that no task-specific architectural designs were applied. Instead, a single LLM was utilized, demonstrating competitive performance across almost all tasks. The key to achieving such results seems to lie in properly prompting the model with natural language requests or employing task-specific encoding techniques for the outputs. We can expect higher results to be achieved using larger LLMs such as LLaMA 65B. To conduct a more comprehensive evaluation and optimization, it would have been beneficial to explore a broader range of architectures and thoroughly investigate all the hyper-parameters of the models. The estimation of these parameters was done hastily due to the time constraints imposed by the EVALITA deadlines and the extensive commitment required for the parallel completion of all 13 tasks.

Error Analysis. Since our team participated in all the tasks, it would be unfeasible to provide a deeper analysis of each individual result in this report. However, in order to gain some insight into the inner working of the two models we employed, here we present some error analysis carried out on two tasks. We selected a task where our systems ranked very high, and one where they ranked very low. In the EmIt task A, `extremITLLaMA` ranked first in the official ranking, and `extremIT5` was second. The task is a multi-label classification problem, where the labels are eight emotions defined by Plutchik [34] plus "love" and a label for neutral texts. Table 4 reports the performance of the two `ExtremITA` systems broken down by labels. It is interesting to notice that the advantage shown by `extremITLLaMA` on the aggregated result comes from a skewed distribution over the labels. In particular, `extremIT5` is hardly capable of modeling Fear, which is also the least represented label in the test set. An inverse correlation between the number of positive instances in the test set and the gain in performance of `extremITLLaMA` with respect to `extremIT5` is indeed present. This indicates that `extremITLLaMA` is better suited than `extremIT5` for the classification of sparser phenomena. Moreover, `extremITLLaMA` shows superior capability in modeling and correctly predicting every emotion, besides "Trust", where `extremIT5` results in a better performance.

In the LangLearn task, our systems ranked quite low, respectively 8th place for `extremITLLaMA` and 10th place for `extremIT5`. LangLearn is a text pair classi-

Label	extremIT5			extremITLLaMA			Δ			Support
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	
Anger	0.500	0.464	0.481	0.759	0.393	0.518	0.259	-0.071	0.037	56
Anticipation	0.690	0.471	0.559	0.675	0.612	0.642	-0.015	0.141	0.083	85
Disgust	0.554	0.594	0.573	0.674	0.588	0.628	0.120	-0.006	0.055	165
Fear	1.000	0.077	0.143	0.636	0.538	0.583	-0.364	0.461	0.440	13
Joy	0.684	0.520	0.591	0.648	0.590	0.618	-0.036	0.070	0.027	100
Love	0.708	0.330	0.450	0.745	0.398	0.519	0.037	0.068	0.069	103
Neutral	0.705	0.614	0.656	0.657	0.757	0.704	-0.048	0.143	0.047	210
Sadness	0.584	0.474	0.523	0.750	0.537	0.626	0.166	0.063	0.103	95
Surprise	0.344	0.539	0.420	0.632	0.422	0.506	0.288	-0.117	0.086	102
Trust	0.679	0.699	0.688	0.698	0.673	0.685	0.019	-0.026	-0.003	272

Table 4
Performance in terms of Precision, Recall and F1-measure of our systems on the EmIt A task, where the Δ column is the difference between `extremITLLaMA` and `extremIT5`.

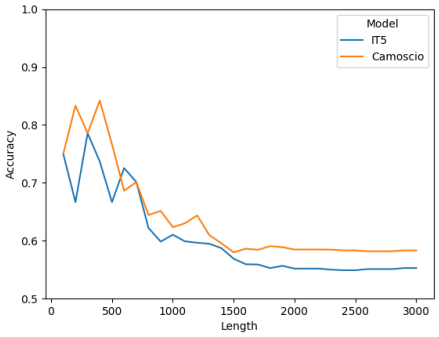

Figure 1: Accuracy of our systems on the LangLearn test set, with texts removed that are longer than an increasing threshold (horizontal axis).

fication task where the most informative features are expected to be stylistic, rather than semantic, to capture the development in language learning of the author of the texts. With this premise, we were anticipating a subpar performance by our transformer-based models from the beginning. However, another relevant characteristic of this task is the length of the texts. For computational reasons, we had to cut the texts to 100 tokens or less, therefore leaving out a significant portion of the data — we retained exactly 24.6% of the tokens from the two training sets combined. We checked the impact of the text size on the accuracy of the prediction, under the hypothesis that longer texts in the test set (which were cut by our systems to a greater extent) are penalized. The plot in Figure 1 shows the accuracy of our systems against portions of the test set where the texts were filtered by size. The number on the horizontal axis is a threshold on the minimum size in terms of characters of the two texts forming an instance of the test set. Indeed, the downward trend indicates that the predictions of our systems are more accurate on shorter pairs of texts, while more and more errors are made by both systems on longer texts.

4. Conclusions

In a recent position paper with a provocative title, Basile [35] asks himself "is EVALITA done?", referring to the mounting trend of LLMs and zero-shot approaches in NLP and their impact on the evaluation campaign. Judging by the results presented in this report, the answer is still the same as the original paper, i.e., *no*. The variety and challenge offered by the tasks of EVALITA continue to represent a fundamental resource to understand and develop language resources and tools for the Italian language, as shown, for instance, by the variability of the ranking obtained by our transformer-based models. However, the raw performance of `extremIT5` and `extremITLLaMA`, with minimal adaptations and tuning, is undoubtedly pushing the limits of some tasks, especially text classification tasks with roots in text semantics. In any case, these results once again confirm the huge potential of LLMs and their applicability in real-world scenarios. It is important to note that this experiment, while not conclusive, used the smallest available models due to their size limitations. Additionally, it would be worthwhile, from a sustainability standpoint, to explore the results that can be achieved by significantly reducing the amount of annotated data available through zero or few-shot learning approaches.

Acknowledgments

We would like to thank the "Istituto di Analisi dei Sistemi ed Informatica - Antonio Ruberti" (IASI) for supporting the experimentations. Claudiu Daniel Hromei is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on *Health and life sciences*, organized by the Università Campus Bio-Medico di Roma.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67. URL: http://jmlr.org/papers/v21/20-074.html.

[2] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 483–498. URL: https://doi.org/10.18653/v1/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[3] G. Sarti, M. Nissim, IT5: large-scale text-to-text pretraining for italian language understanding and generation, CoRR abs/2203.03759 (2022). URL: https://doi.org/10.48550/arXiv.2203.03759. doi:10.48550/arXiv.2203.03759. arXiv:2203.03759.

[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, CoRR abs/2210.11416 (2022). URL: https://doi.org/10.48550/arXiv.2210.11416. doi:10.48550/arXiv.2210.11416. arXiv:2210.11416.

[5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, CoRR abs/2005.14165 (2020). URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[8] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[9] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[10] G. Gafà, F. Cutugno, M. Venuti, Emotivita at EVALITA2023: Overview of the dimensional and multidimensional emotion analysis task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] D. Russo, S. M. Jiménez-Zafra, J. A. García-Díaz, T. Caselli, M. Guerini, L. A. Ureña-López, R. Valencia-García, Overview of PoliticIT2023@EVALITA: Political Ideology Detection in Italian Texts, 2023.

[12] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the task on geolocation of linguistic variation in Italy, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[13] C. Alzetta, D. Brunato, F. Delll'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, Langlearn at evalita 2023: Overview of the language learning development task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[14] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[15] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[16] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task, in: Proceed-

ings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[17] P. Russo, N. Stoehr, M. Horta Ribeiro, Subtask a- conspiratorial content classification, 2023. URL: https://kaggle.com/competitions/acti-subtask-a.

[18] P. Russo, N. Stoehr, M. Horta Ribeiro, Subtask b - conspiracy category classification, 2023. URL: https://kaggle.com/competitions/acti-subtask-b.

[19] G. Russo, L. Verginer, M. H. Ribeiro, G. Casiraghi, Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 742–753.

[20] A. Palmero Aprosio, T. Paccosi, Nermud at evalita 2023: Overview of the named-entities recognition on multi-domain documents task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[21] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain , in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[22] P. Cassotti, L. Siciliani, L. Passaro, M. Gatto, P. Basile, Wic-ita at evalita2023: Overview of the evalita2023 word-in-context for italian task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[23] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, Discotex at evalita 2023: Overview of the assessing discourse coherence in italian texts task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. arXiv:2210.11416.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[26] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the NAACL 2019, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[28] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[30] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[31] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[32] A. Santilli, Camoscio: An italian instruction-tuned llama, https://github.com/teelinsan/camoscio, 2023.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[34] R. Plutchik, H. Kellerman, Theories of emotion 1 (1980).

[35] V. Basile, Is EVALITA done? on the impact of prompting on the italian NLP evaluation campaign, in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022), volume 3287 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 127–140. URL: https://ceur-ws.org/Vol-3287/paper13.pdf.