Supplementary Material

Table 10. Illustrative (simple and natural) prompts explored to semi-automate all the data wrangling batteries. When different alternatives are provided, those with the best results in our preliminary experiments are chosen for the rest of the experiments in the paper. The last two columns provide some tips on the application of GPT3 for the automation of different data wrangling tasks.

Battery	Prompt Template		Notes	When does it work well?	When does it work badly?
	Examples (n-shot)	Question			
Manipulation	Input: 'X'\nOutput: 'Y'\n\n	Input: 'Z'\nOutput:	No alternatives. We wanted a generalisation from the concept itself and we tested a varying number of n-shots. The semantic context about the domain is clear.	Tasks that can be solved by simple string transformation	Reasoning or calculation capabilities are required (e.g., identification of dimensions, coefficients, arithmetic operations, etc.). Domain information required (e.g., date format)
Types		What is the best domain that describes the values in $\{v_1, v_2, \ldots, v_n\}$?		Asking for the 'semantic type' of the values. Working with nominal variables	Asking for the 'domain' of the values. Working with numerical variables
		What is the best semantic type that describes the values in $\{v_1, v_2, \ldots, v_n\}$?			
		What type of data is $\{v_1, v_2, \ldots, v_n\}$?	Alternative. Not very specific. Worst results in the initial tests		
		What domain of data is $\{v_1, v_2, \ldots, v_n\}$?	Alternative. Not very specific. Worst results in the initial tests		
		What are $\{v_1, v_2, \ldots, v_n\}$?	Alternative. Not very specific. Worst results in the initial tests		
		What is the domain of the set $\{v_1, v_2, \ldots, v_n\}$?	Alternative. Not very specific. Worst results in the initial tests		
Ordinal	Is 'house' higher than 'apartment'? Yes Is 'apartment' higher than 'house'? No Is 'red' higher than 'blue'? No Is 'blue' higher than 'red'? No Is 'old' higher than 'young'? Yes Is 'young' higher than 'old'? No Is 'totally agree' higher than 'agree'? Yes Is 'agree' higher than 'totally agree'? No Is 'New York' higher than 'Chicago'? No Is 'Chicago' higher than 'New York'? No Is 'Heavy rain' higher than 'Showers'? Yes Is 'Showers' higher than 'Heavy rain'? No	Is "v_1" higher than "v_2"?	The context is always the same for all the examples (12-shot), while only the 13th line changes. This long context is added because it helps to frame the question and gives better results. We tried with some other terms different from 'higher', such as 'larger', 'more', etc., but we got worst results in the initial tests for these variations	Telling ordinal vs non-ordinal. Working with attributes that have a textual representation of numbers or intervals (e.g., age, tumor-size, etc.).	Recognising the actual order of complex attributes or dealing with time if the adjective is not adjusted. Dealing with absent values
	Is (low, medium, high) an ordinal? Yes Is (door, window, wheel) an ordinal? No	Is $\{v_1, v_2, \ldots, v_n\}$ an ordinal?	Simpler version: 2-shot (always the same) and the question.		
Anomalies		Are there any outliers in $\{v_1, v_2, \ldots, v_n\}$?		Dealing with anomalies of attributes with a low number of unique values	Dealing with attributes with many unique alphanumeric terms of different lengths
		Which is the odd-one-out in $\{v_1, v_2, \ldots, v_n\}$?	Alternative. Despite being equivalent to 'oulier', the initial tests showed worse results.		
Imputation	feature_1: v_1, feature_2: v_2, ..., feature_n: v_3 feature_1: v_4, feature_2: v_5, ..., feature_n: v_6 ... feature_1: v_7, feature_2: v_8, ..., feature_n: v_9	feature_1: v_n, feature_2: v_m, ..., feature_n:	Full prompt strategy using all the available features in the dataset and different n-shot strategies	Relevant attributes are provided. Imputation for both categorical and numerical data	Complex attributes (e.g., addresses)
	feature_x: v_1, feature_y: v_2 ... feature_z: v_3, feature_y: v_4	feature_x: v_5, feature_y:	1-Feature prompt using only the most relevant feature and different n-shot strategies		

Fig. 8. Overview of results obtained by the different versions of GPT-3 models disaggregated by learning strategy (from zero-shot to 10-shot learning), system, domain and data wrangling task. Tasks per domain sorted by average performance.

Table 11. Wrong answers provided by the GPT-3 DaVinci system setting (examples based on large sentences are represented in Table 12). 4-shot setting used for illustrative purposes.

Input 1	Input 2	Input 3	Output	Correct output	Problem	Domain
1990.18.01	-	-	01.18.1990	01/18/1990	changeFormat-1	dates
28-10-2001	-	-	10/28/2001	08/23/2010	changeFormat-2	dates
99/09/03	-	-	09/99/03	09/03/99	changeFormat-3	dates
06-04-2001	-	-	04/06/2001	06/04/2001	changeFormat-3	dates
08-07-2017	-	-	07-08-2017	08-07-2017	changeFormat-4	dates
31/05/17	-	-	17	31	getDay-1	dates
25-08-85	-	-	08	25	getDay-1	dates
31/03/75	-	-	31st	31th	getDayOrdinal-1	dates
09/11/53	-	-	11th	09th	getDayOrdinal-1	dates
05/09/2008	-	-	5th	05th	getDayOrdinal-2	dates
Thursday, October 31, 1985	-	-	October	Thursday	getMonthName-1	dates
11/06/2016	-	-	Jun	Nov	reduceMonthName-2	dates
8072017	-	-	07-08-2008	08-07-2017	setFormat-1	dates
2261993	-	-	93-22-1993	26-02-1993	setFormat-1	dates
8222000	-	-	22/02/2000	12/08/2000	setFormat-2	dates
11202001	-	-	20/12/2001	10/11/2001	setFormat-2	dates
casper	canal13	-	casper@canal13.cl	casper@canal13.com	generate-3	emails
caedgigo@garmendia.cl	-	-	gmail.com	garmendia.cl	getAfterAt-2	emails
1-845-456-7891	-	-	18454567819	18454567891	deletePunctuation-1	freetext
1-845-333-7891	-	-	18453337890	18453337891	deletePunctuation-1	freetext
1-7891	-	-	7891	17891	deletePunctuation-2	freetext
1-7892	-	-	7892	17892	deletePunctuation-2	freetext
Aliquam	-	-	a	A	firstCharacter-2	freetext
Computer and Communications Security	-	-	ACSAC	CCS	getCaps-2	freetext
Jones	1	-	Mr. Jones	Sr. Jones	addTitle-2	names
Dario Gag-Dorado	-	-	dari	daga	login-1	names
Paco Gabot Narale	-	-	paco	pagana	login-1	names
Daman Hivser-Kleiner	-	-	dakle	dahi	login-2	names
Giussepe Hindravtoks	-	-	giussepe	gihi	login-2	names
Agustino	Zimmann	-	Agustino, A.	Agustino, Z.	reduceName-5	names
618-4390	PAN	-	(7) 618-4390	(507) 618-4390	countryPrefix-3	phones
36-678-59-10	AUT	-	(9) 36-678-59-10	(43) 36-678-59-10	countryPrefix-7	phones
846-2730	AND	-	(846) 2730	(376) 846-2730	countryPrefix-7	phones
425-846-2730 425-425 425	-	-	425-425 425	425-846-2730	getNumber-2	phones
618 4390	425	-	425-618-4390	725-618-4390	setPrefix-1	phones
743-1650	425	-	425-743-1650	892-743-1650	setPrefix-1	phones
618-4390	425	-	(+425) 618-4390	(+725) 618-4390	setPrefix-5	phones
743-1650	425	-	(+425) 743-1650	(+892) 743-1650	setPrefix-5	phones
08	5	-	12	13	addTime-1	times
16:15:12	5	-	05:15:12	21:15:12	addTime-1	times
21:20	5	-	26:20	02:20	addTime-2	times
16:15:12	-	-	16:15:12:00	16:15:12	appendTime-1	times
16:15:12	30	-	16:15:12:30	16:15:12	appendTime-3	times
06:15:12	-	-	06:15:12	16:15:12	convert-1	times
08:40	EST	PST	20:40:00	05:40:00	convert-10	times
06:15	CET	MST	00:15:00	22:15:00	convert-10	times
14:10	12h	-	14:10	02:10 PM	convert-3	times
21:20	12h	-	21:20	09:20 PM	convert-3	times
08:40 UTC	24h	-	20:40	08:40	convert-4	times
06:15:12	24h	-	06:15:12	16:15:12	convert-4	times
14:10	-	-	14:10 AM	02:10 PM	convert-5	times
21:20	-	-	21:20 AM	09:20 PM	convert-5	times
08:40 UTC	-	-	03:40 PM	08:40 AM	convert-6	times
14:10	CET	UTC	14:10	13:10	convert-7	times
21:20	EST	UTC	03:20	02:20	convert-7	times
08:40	EST	PST	20:40	05:40	convert-8	times
06:15:12	CET	MST	18:15:12	22:15:12	convert-8	times
14:10	CET	UTC	14:10:00	13:10:00	convert-9	times
21:20	EST	UTC	07:20:00	02:20:00	convert-9	times
08	-	-	08	-	deleteTime-1	times
12.20dg	mg	-	12.200001	1220	convert-1	units
1854 dam	dm	-	18.540000	185400	convert-1	units
92.26 K	cK	-	92.26	9226	convert-2	units
12.2	dg	mg	12.200001	1220	convert-3	units
1854	dam	dm	18.540000	185400	convert-3	units
92.26	K	cK	92.26	9226	convert-4	units
81	hm	cm	0.081	8100	convert-4	units
1854 dam	-	-	Area	Length	getSystem-1	units
12.20dg	-	-	12.20	12.2	getValue-1	units

Table 12. Wrong answers provided by the GPT-3 DaVinci system for examples from the Manipulation Battery based in large sentences. 4-shot setting used for illustrative purposes.

		problem	**domain**
Input	Etiam dapibus tortor sed cursus blandit.		
Output	Etiamdapibustortorsedcursussed	deletePunctuation-3	freetext
Correct output	Etiamdapibustortorsedcursusblandit		
Input	Quisque mattis dolor ac metus fermentum consectetur.		
Output	Quismattisdoloracmetusfermentumconsectetur	deletePunctuation-3	freetext
Correct output	Quisquemattisdoloracmetusfermentumconsectetur		
Input	Integer malesuada elit et hendrerit fermentum.		
Output	Integermalesuadaelithendreritfermentum	deletePunctuation-4	freetext
Correct output	Integermalesuadaelitethendreritfermentum		
Input	Pellentesque tempor justo a dui varius maximus.		
Output	Pellentesque temporjustoaduivariusmaximus	deletePunctuation-4	freetext
Correct output	Pellentesquetemporjustoaduivariusmaximus		
Input	Mauris fringilla neque nec libero consequat quis facilisis nisi tristique		
Output	Maurisfringillanequeconsequatlibero	deleteSpaces-2	freetext
Correct output	Maurisfringillanequenecliberoconsequatquisfacilisisnisitristique		
Input	Nam cursus purus nec mi maximus fermentum		
Output	Namcursuspurusnecminimaximusfermentum	deleteSpaces-2	freetext
Correct output	Namcursuspurusnecmimaximusfermentum		
Input	Quisque mattis dolor ac metus fermentum consectetur.		
Output	QUISQUE MATTIS DOLOR AC METUS FERMENTUM CONSEQUETUR	toUpper-1	freetext
Correct output	QUISQUE MATTIS DOLOR AC METUS FERMENTUM CONSECTETUR		

Table 13. Illustrative examples of answers provided by the GPT-3 DaVinci system for some columns from the Types Battery.

Zillow	**Prompt1**
Input	What is the best domain that describes the values in {Palmyra, Defiance,. . . ,Verona}?
Output	The best domain that describes the values in {Palmyra, Defiance,. . . ,Verona} is ”small towns”
Correct type	city
Input	What is the best domain that describes the values in {Norman, Frackville, . . . , Corte Madera}?
Output	The best domain would be ”towns in the United States.”
Correct type	city
	Prompt2
Input	What is the best semantic type that describes the values in {Palmyra, Defiance,. . . ,Verona}?
Output	The best semantic type that describes the values in {Palmyra, Defiance,. . . ,Verona} is Place
Correct type	City
Input	What is the best semantic type that describes the values in {Norman, Frackville, . . . , Corte Madera}?
Output	The best semantic type that describes the values in the given set is ”cities.”
Correct type	City
MedicalNoShow	**Prompt1**
Input	What is the best domain that describes the values in {F, M, M, F, M, M, M, M, M, M}?
Output	The best domain would be ”Frequency of Males.”
Correct type	Gender
Input	What is the best domain that describes the values in {F, M, M, F, F, M, F, M, F, M}?
Output	The best domain would be ”Females and Males.”
Correct type	Gender
	Prompt2
Input	What is the best semantic type that describes the values in {F, M, M, F, M, M, M, M, M, M}?
Output	The best semantic type that describes the values in {F, M, M, F, M, M, M, M, M, M} is ”Gender.”
Correct type	Gender
Input	What is the best semantic type that describes the values in {F, M, M, F, F, M, F, M, F, M}?
Output	The best semantic type that describes the values in {F, M, M, F, F, M, F, M, F, M} is ”Gender.”
Correct type	Gender
SpeedDating	**Prompt1**
Input	What is the best domain that describes the values in {28, 20, 32, 28, 42, 34, 37, 34, 28, 30}?
Output	The best domain would be 28-42.
Correct type	Age
Input	What is the best domain that describes the values in {29, 55, 23, 55, 28, 21, 31, 33, 29, 34}?
Output	The best domain would be {21, 23, 28, 29, 31, 33, 34, 55}.
Correct type	Age
	Prompt2
Input	What is the best semantic type that describes the values in {28, 20, 32, 28, 42, 34, 37, 34, 28, 30}?
Output	The best semantic type that describes the values in {28, 20, 32, 28, 42, 34, 37, 34, 28, 30} is a set of whole numbers.
Correct type	Age
Input	What is the best semantic type that describes the values in {29, 55, 23, 55, 28, 21, 31, 33, 29, 34}?
Output	The best semantic type that describes the values in {29, 55, 23, 55, 28, 21, 31, 33, 29, 34} is a set of whole numbers.
Correct type	Age

Table 14. Illustrative examples of answers provided by the GPT-3 DaVinci system for numerical attributes from the Anomalies Battery.

Dataset (Attribute)		
BreastCancer (chromatin)	**Input**	Are there any outliers in {1, 3, 4, 1, 2, 6, 4 . . . }?
	Output	The outliers are 9, and 10.0.
	Boxplot Output	The outlier is 10
Ecoli (mcg)	**Input**	Are there any outliers in {0.0, 0.04, 0.06, 0.1, 0.11, 0.12, 0.16, 0.17, 0.18, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, . . . }?
	Output	The outliers are 0.0, 0.04, 0.06, 0.1, 0.11, 0.12, 0.16, 0.17, 0.18, 0.2, 0.21, 0.22, 0.23, 0.24, 0.25, . . .
	Boxplot Output	There are no outliers.
Wine (mcg)	**Input**	Are there any outliers in {3.06, 0.76, 3.24, 1.49, 4.69, 3.39, . . . }?
	Output	The outliers are 0.47, 0.61, 0.63, and 0.66. . . .
	Boxplot Output	There are no outliers.
Glass (Na)	**Input**	Are there any outliers in {10.73, 11.2, 14.3, 12.07, 13.3, . . . }?
	Output	The outliers are 10.73, 11.02, 11.03, 11.23, and 15.79.
	Boxplot Output	The outliers are 10.73, 11.02, 11.03, 11.23, 11.45, 15.79, 17.38
Iris (Sepal.Width)	**Input**	Are there any outliers in {3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9, . . . }?
	Output	The outliers are 3.0, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2 and 4.4.
	Boxplot Output	There are no outliers.
Mpg (hwy)	**Input**	Are there any outliers in {29, 29, 31, 30, 26, 15, 20, 17, 26, . . . }?
	Output	The outliers are 5, 12, 14, 18, and 17.
	Boxplot Output	The outliers are 44, 44, 41.

Table 15. Illustrative examples of correct and wrong answers provided by the GPT-3 DaVinci system for categorical attributes from the Anomalies Battery. Inserted synthetic anomalies highlighted in yellow.

Attribute		
model	**Input**	Are there any outliers in {a4, a4 quattro, a6 quattro, c1500 suburban 2wd, corvette, k1500 tahoe 4wd, malibu, caravan 2wd, dakota pickup 4wd, durango 4wd, ram 1500 pickup 4wd, expedition 2wd, **1.8**, explorer 4wd, f150 pickup 4wd, mustang, civic, sonata, tiburon, grand cherokee 4wd, range rover, navigator 2wd, mountaineer 4wd, altima, maxima, pathfinder 4wd, grand prix, forester awd, impreza awd, 4runner 4wd, camry, camry solara, corolla, land cruiser wagon 4wd, toyota tacoma 4wd, gti, jetta, new beetle, passat, . . . }?
	Output	There are no outliers in this data
	Correct Output	The outlier is 1.8
trans	**Input**	Are there any outliers in {auto(l5), manual(m5), manual(m6), auto(av), auto(s6), auto(l4), auto(l3), auto(l6), **honda**, auto(s5), auto(s4), . . . }?
	Output	There are no outliers in this data
	Correct Output	The outlier is honda
model	**Input**	Are there any outliers in {audi, chevrolet, dodge, ford, honda, hyundai, jeep, land rover, lincoln, mercury, nissan, **r**, pontiac, subaru, toyota, volkswagen, . . . }?
	Output	There are no outliers in this data
	Correct Output	The outlier is r
manufacturer	**Input**	Are there any outliers in {audi, chevrolet, dodge, ford, honda, hyundai, **f**, jeep, . . . }?
	Output	There are no outliers in this data
	Correct Output	The outlier is f
fl	**Input**	Are there any outliers in {p, **24**, r, e, d, c, . . . }?
	Output	The outlier is 24
	Correct Output	The outlier is 24
drv	**Input**	Are there any outliers in {"f" "4" "r", **auto(l5)** . . . }?
	Output	The outlier is auto(l5)
	Correct Output	The outlier is auto(l5)

Table 16. Generic instructions providing a basic skeleton for the incorporation of language models in the automation or assistance of this type of tasks. They should be customised depending on the problem to be solved, the capabilities required for its resolution, the ability of the user as a good prompter, the requirements and costs of the language models, and the evaluation criteria used to determine the efficiency of the whole pipeline.

Guidelines

1. Familiarise with language models using the interactive interface but also programmatically through an API. For instance, replicating the results for some of the batteries in this paper could be a good training for users.
2. Determine and understand the availability, quotas and any restriction about the language models for which regular access is available.
3. Identify the data wrangling problem, and, if possible, associate it with the taxonomy in Table 1.
4. Identify how the problem appears in the data processing pipeline, if it is a one-off process or it is going to be repetitive. This will quantify the scaling factor and the potential gain of automation.
5. Determine if language models are a good fit: if the task requires semantic content, only a few examples and simple reasoning, then language models may be an option. Always think of other better or cheaper alternatives.
6. Derive and try some prompts for the task at hand. These can be few-shot input-output prompts using several examples, zero-shot prompts describing the task to do for a single input, or a combination of both.
7. Decide the strategy to extract the results from the language model automatically and how their quality is going to be evaluated.
8. Run some preliminary tests with some samples, variations of prompts and extraction processes. Derive the quality metrics, the total human time invested and the number of tokens/compute used by the system.
9. If the result is sufficiently effective considering all the factors derived in the point above, then run the data wrangling process with all the data.
10. Document the process and the lessons-learned for the future, and disseminate the experience with other colleagues inside and beyond the organisation that regularly face data wrangling problems.