Macro-descriptors for Alzheimer's disease detection using large language models: Supplementary Material

This document presents the supplementary material for the Interspeech 2024 paper [1].

1. Prompting strategies

Figures 1 to 7 show the prompts used to query the LLMs. The examples in prompt strategy 4.1 are from DementiaBank corpus [2], but do not integrate the ADReSS corpus. These examples correspond to the manual transcriptions. Similar prompts were defined for the ASR generated transcriptions.

P1.1:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD).
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure 1: *Prompt strategy P1.1.*

P1.2:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD).
Notice that the speech of a person suffering from AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure 2: *Prompt strategy P1.2.*

P1.3:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD).
Notice that the speech of a person suffering from AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion.
The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). For example, with reference to the first concept, "the mother is drying a plate" or "lady do dishes" is considered "accurate and complete; "lady with dishes" or "the mother is standing by the sink" is considered accurate but incomplete; "the woman is washing clothes" is considered inaccurate.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure 3: *Prompt strategy P1.3.*

P1.4:

USER:

You are an expert evaluator that detects if a person suffers from Alzheimer's Disease, from their language.
Person A is describing the "Cookie Theft" image using spontaneous speech. Given the transcription of the person's description, your task is to predict if the person suffers from Alzheimer's Disease (AD). Notice that the speech of a person suffering from AD is characterized by word-finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, and inappropriate use of pronouns. Furthermore, the discourse of AD patients is described as fluent but not informative, characterized by incomplete and short sentences, and lacking coherence and cohesion.
The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). For example, with reference to the first concept, "the mother is drying a plate" or "lady do dishes" is considered "accurate and complete; "lady with dishes" or "the mother is standing by the sink" is considered accurate but incomplete; "the woman is washing clothes" is considered inaccurate.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

No other output should be provided.

DESCRIPTION: www . the boy's getting cookies out of the cookie jar . he's handing one to a girl . the lad the stool he's standing on is falling . the lady's drying dishes . the sink is running over . the water's turned on full . um cups are sitting on the counter , plates sitting on the counter . puddle of water's on the floor . little girl is saying . shh don't tell anybody . and the cookie jar looks like it's ready to fall out . and the cookie jar is full, clear full . that's about all I see that's going on .

ASSISTANT:

{'comments': 'The description seems to be complete, mentioning the main entities in the picture: the kitchen, the girl, the boy, the lady or mother, the cookie jar and the overflowing sink. The description is also coherent. lexically diverse, and does not seem to evidence wording finding difficulties.',
'alzheimers_prediction': NO ,
'confidence_in_prediction': high }

USER:

DESCRIPTION: mm dishes are being dried . and the child is getting some cookies out of the jar . the the uh plant stand or the stand he's on is looks as though it's crooked . and the water's going over in the sink from the sink . is that all ? and then the boy was getting the cookies out of the jar and the cover is off . but she he's giving her a cookie and the stool is turning over and the water in the sink is boiling over or flowing over . and she's drying dishes .

ASSISTANT:

{'comments': 'The description seems confusing, without clearly identifying the subjects, and simply mentioning "she". The description has a poor lexically diversity. It seems to evidence wording finding difficulties, eg. "the the uh plant".',
'alzheimers_prediction': YES ,
'confidence_in_prediction': high }

USER:

DESCRIPTION: <picture_description>

Figure 4: *Prompt strategy P1.4.*

P1.5:

USER:

You are an expert fluency evaluator. Your task is to evaluate the description of an image provided in spontaneous speech by a person. The evaluation should focus on identifying word finding difficulties, repetitions, reduced vocabulary, an overuse of indefinite and vague terms, inappropriate use of pronouns and lack of coherence and cohesion.
The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). For example, with reference to the first concept, "the mother is drying a plate" or "lady do dishes" is considered "accurate and complete; "lady with dishes" or "the mother is standing by the sink" is considered accurate but incomplete; "the woman is washing clothes" is considered inaccurate.
Provide the output using the following json format:

{'comments': step-by-step explanation, with maximum 300 tokens.
'issues': YES/NO,
'confidence': high/low}

No other output should be provided.

DESCRIPTION: *<picture_description>*

Figure 5: *Prompt strategy P1.5.*

P2.1:

USER:

You are an expert fluency evaluator.
Person A is describing the "Cookie Theft" image using spontaneous speech. The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). Given the transcription of the person's description, your task is to evaluate the text in terms of coherence, lexical diversity, sentence length and word finding difficulties, using scores between 0 and 1.
Provide the ratings in a json format such as the example below. No other outputs.

{'text_coherence': number between 0 and 1,
'lexical_diversity': number between 0 and 1,
'sentence_length': number between 0 and 1,
'word_finding_difficulties': number between 0 and 1}

DESCRIPTION: *<picture_description>*

Figure 6: *Prompt strategy P2.1.*

P2.2:

USER:

You are an expert fluency evaluator, that works in the medical domain to support medical screening.
Person A is describing the "Cookie Theft" image using spontaneous speech. The image can be described using seven concepts (woman doing dishes, sink overflowing, boy on stool, children stealing cookies, girl reaching for cookie, stool falling, woman not noticing). Given the transcription of the person's description, your task is to evaluate the text in terms of coherence, lexical diversity, sentence length and word finding difficulties, using scores between 0 and 1. Then you evaluate if the person is likely to suffer from Alzheimer's Disease.
Provide the ratings in a json format such as the example below. No other outputs.

{'text_coherence': number between 0 and 1,
'lexical_diversity': number between 0 and 1,
'sentence_length': number between 0 and 1,
'word_finding_difficulties': number between 0 and 1,
'alzheimers_prediction': YES/NO,
'confidence_in_prediction': high/low}

DESCRIPTION: *<picture_description>*

Figure 7: *Prompt strategy P2.2.*

2. Task 1: LLMs as AD predictors

In this section, we present table 1 and figure 8, where we include the results obtained with *Llama-2-13B*, besides the results included in the main body of the paper. We have also performed preliminary experiments with *Llama-2-7B*, but the output frequently failed to comply with the requested format, thus not allowing the automatic analysis of the results.

Table 1: *AD classification based on LLM predictions. #Fail denotes the number of examples for which the model failed to follow the output instruction (identified on train/test). Reported results for accuracy, Acc, are presented in %.*

	Llama 13B			**Mistral 7B**			**Mixtral 8x7B**			**GPT-3.5**			**Mean**
	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	#Fail	Acc_{train}	Acc_{test}	**Acc**
Manual transcriptions													
P1.1	0/0	56.5	54.2	9/4	61.1	64.6	10/5	55.6	54.2	0/0	65.7	54.2	58.8
P1.2	0/0	48.1	56.3	21/4	63.9	64.6	4/2	62.0	64.6	0/0	51.9	60.4	58.0
P1.3	0/0	55.6	54.2	15/6	61.1	75.0	8/3	**72.2**	60.4	0/0	54.6	56.3	61.1
P1.4	4/1	50.9	47.9	56/28	62.0	70.8	17/7	61.1	66.7	0/0	59.3	66.7	59.8
P1.5	18/5	51.9	56.3	2/2	55.6	60.4	0/0	53.7	50.0	0/0	50.0	50.0	53.2
P2.2	0/0	50.0	50.0	1/0	54.6	54.2	0/0	60.2	54.2	0/0	67.6	70.8	57.9
Whisper transcriptions													
P1.1	0/0	55.6	50.0	13/4	67.6	58.3	14/5	60.2	56.3	0/0	65.7	64.6	60.7
P1.2	0/0	49.1	54.2	17/3	64.8	70.8	5/7	63.0	54.2	0/0	65.7	72.9	61.4
P1.3	0/0	54.6	52.1	14/9	67.6	60.4	11/4	60.2	68.8	0/0	63.9	66.7	61.7
P1.4	1/1	53.7	52.1	54/21	61.1	70.8	11/4	66.7	66.7	0/0	63.9	72.9	**62.7**
P1.5	22/11	53.7	50.0	6/1	62.0	72.9	1/1	55.6	58.3	0/0	50.0	50.0	56.1
P2.2	0/0	50.0	50.0	3/2	55.6	52.1	0/0	62.0	56.3	0/0	68.5	**75.0**	58.8
Wav2vec transcriptions													
P1.1	0/0	45.4	33.3	21/7	58.3	52.1	18/8	56.5	52.1	0/0	54.6	52.1	51.8
P1.2	0/0	49.1	39.6	8/8	58.3	54.2	8/7	56.5	62.5	0/0	50.9	47.9	52.9
P1.3	0/0	52.8	47.9	16/6	56.5	54.2	10/3	61.1	58.3	0/0	50.9	47.9	54.3
P1.4	0/0	50.9	47.9	56/20	64.8	58.3	12/5	66.7	66.7	0/0	58.3	56.3	59.3
P1.5	3/2	50.0	47.9	0/0	49.1	45.8	0/0	51.9	47.9	0/0	50.9	47.9	49.5
P2.2	0/0	50.0	50.0	2/3	55.6	56.3	0/0	63.0	60.4	0/0	63.9	62.5	57.9
Mean	3/1	51.5	49.7	17/7	60.0	60.9	7/3	60.4	58.8	0/0	58.7	59.7	

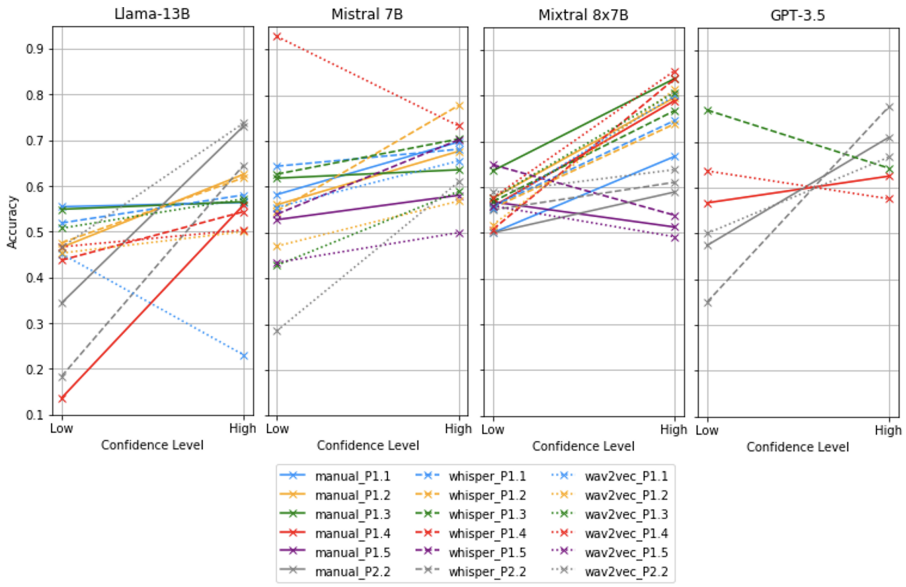

Figure 8: *Task 1: Combined train and test accuracy per confidence level. We require a minimum of 10 instances per confidence level to plot the prompting strategy.*

3. Task 2: LLMs as extractors of macro-descriptors

This section presents additional results for task 2, where we explore the potential of LLMs as extractors of macro-descriptors, for posterior AD classification. Table 2 reports the results for the five classifiers – SVM, LDA, 1NN, DT, and RF – that were used in the ADReSS baseline [3] , and for the four LLMs – Llama, Mistral, Mixtral, and GPT.

Table 2: *AD classification based on macro-descriptors.*

		Llama2 13B					**Mistral 7B**					**Mixtral 8x7B**					**GPT-3.5**					
		SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	SVM	LDA	1NN	DT	RF	**Mean**
Train set: 10-Fold CV																						
Manual	P2.1	65.7	66.7	50.9	67.6	65.7	63.9	63.0	65.7	67.6	67.6	70.4	73.1	69.4	68.5	75.0	70.4	69.4	63.9	70.4	70.4	67.3
	P2.2	67.6	66.7	54.6	67.6	67.6	63.0	65.7	57.4	59.3	70.4	71.3	73.1	67.6	75.9	75.9	67.6	68.5	54.6	65.7	70.4	66.5
Whisper	P2.1	68.5	68.5	63.9	66.7	67.6	70.4	70.4	58.3	66.7	72.2	69.4	70.4	73.1	67.6	70.4	66.7	66.7	63.0	68.5	68.5	67.9
	P2.2	64.8	65.7	56.5	69.4	69.4	73.1	74.1	61.1	**78.7**	**78.7**	75.0	69.4	70.4	75.9	75.9	69.4	70.4	57.4	61.1	71.3	69.4
Wav2vec	P2.1	62.0	60.2	52.8	63.0	63.0	64.8	64.8	53.7	67.6	65.7	63.9	62.0	51.9	58.3	63.0	63.9	63.9	56.5	70.4	70.4	62.1
	P2.2	57.4	57.4	57.4	60.2	59.3	68.5	69.4	63.0	66.7	66.7	62.0	62.0	61.1	55.6	63.9	63.9	61.1	53.7	66.7	61.1	61.9
Mean				63.1					66.6					68.1					65.5			
Mean (ASR)				62.7					67.7					66.1					64.7			
Test set																						
Manual	P2.1	68.8	72.9	54.2	68.8	68.8	70.8	66.7	45.8	68.8	68.8	66.7	77.1	66.7	64.6	79.2	58.3	60.4	58.3	66.7	62.5	65.7
	P2.2	72.9	70.8	68.8	72.9	72.9	68.8	60.4	54.2	64.6	70.8	77.1	75.0	70.8	75.0	75.0	68.8	68.8	60.4	70.8	68.8	69.4
Whisper	P2.1	66.7	62.5	64.6	58.3	58.3	68.8	68.8	43.8	72.9	72.9	64.6	66.7	64.6	72.9	72.9	66.7	66.7	70.8	68.8	68.8	66.0
	P2.2	54.2	64.6	54.2	64.6	64.6	**81.3**	77.1	64.6	**79.2**	**79.2**	70.8	62.5	64.6	75.0	75.0	75.0	72.9	66.7	77.1	72.9	69.8
Wav2vec	P2.1	64.6	66.7	58.3	66.7	66.7	62.5	62.5	70.8	62.5	66.7	72.9	68.8	60.4	70.8	77.1	70.8	70.8	62.5	66.7	66.7	66.8
	P2.2	64.6	64.6	41.7	52.1	64.6	62.5	62.5	60.4	62.5	68.8	64.6	72.9	58.3	68.8	72.9	62.5	62.5	54.2	58.3	62.5	62.1
Mean				63.8					66.3					70.1					66.3			
Mean (ASR)				61.1					67.5					68.9					67.2			

4. References

[1] C. Botelho, J. Mendonça, A. Pompili, T. Schultz, A. Abad, and I. Trancoso, "Macro-descriptors for Alzheimer's disease detection using large language models," in *Interspeech*, 2024.

[2] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[3] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge," in *Proc. Interspeech 2020*, 2020, pp. 2172–2176.