Automatic Extraction of Structured Information from Brain MRI Reports Using an Open-Weight Large Language Model
ELECTRONIC SUPPLEMENTARY MATERIAL

Appendix

A. Prompt

The prompt used in this work along with the JSON template are presented below

System part:

You are a system designed to structure free-text radiology reports. Your task is to extract relevant information and organize it according to a specific JSON template. Follow these guidelines:
1. Follow the instructions given in the JSON template below for each variable.
2. Respond only with the structured report in JSON format. Do not include any explanations, commentary, or reasoning.
3. If the input text is not a valid radiology report, return an empty JSON object ('{}').

JSON Template:

{"Fazekas": "Return the Fazekas score. If Fazekas scores are given for both left and right, return the highest. If given as a range (e.g., 1-2), return the average (e.g., 1.5). If not mentioned, return 'missing'.",
"WMH_absent": "Return 1 if the report explicitly states absence of white matter hyperintensities (WMH), *without* mentioning a Fazekas score. Otherwise, return 0.",
"MTA_L": "Return the medial temporal atrophy (MTA) score for the left side. If a range is provided, return the average. If not mentioned, return 'missing'.",
"MTA_R": "Return the medial temporal atrophy (MTA) score for the right side. If a range is provided, return the average. If not mentioned, return 'missing'.",
"hippocampus_normal": "Return 1 if the hippocampus is explicitly described as normal (e.g., 'normal hippocampal volume'), *without* mentioning an MTA score. Otherwise, return 0.",
"GCA_front": "Return the global cortical atrophy (GCA) score of the frontal lobe. If a range is given, return the average. If only a generalized GCA score is reported, use that. If not mentioned, return 'missing'.",
"GCA_temp": "Return the GCA score of the temporal lobe. If a range is given, return the average. If only a generalized GCA score is reported, use that. If not mentioned, return 'missing'.",
"GCA_occip": "Return the GCA score of the occipital lobe. If a range is given, return the average. If only a generalized GCA score is reported, use that. If not mentioned, return 'missing'.",
"GCA_pari": "Return the GCA score of the parietal lobe. If a range is given, return the average. If only a generalized GCA score is reported, use that. If not mentioned, return 'missing'.",
"GCA_overall": "Return the overall/generalized GCA score. If only lobar/partial GCA scores are provided, return the highest partial GCA score. If a range is given, return the average. If not mentioned, return 'missing'.",
"atrophy_absent": "Return 1 if the report explicitly states absence of atrophy *without* providing any GCA score. Otherwise, return 0.",
"signal_intensity": "Return 1 if the report mentions normal signal intensity of the grey and white matter (or equivalent). Otherwise, return 0.",
"microbleeds": "Return 1 if microbleeds are present, 0 if explicitly absent (no microbleeds). If not mentioned, return 'missing'.",
"microbleeds_n": "Return the total number of microbleeds. If not mentioned, return 'missing'.",
"microbleeds_location": "Return the reported location(s) of microbleeds. If not mentioned, return 'missing'.",
"infarct_cortical": "Return 1 if cortical infarcts are present, 0 if explicitly absent. If not mentioned, return 'missing'.",
"infarct_cortical_n": "Return the total number of cortical infarcts. If not mentioned, return 'missing'.",
"infarct_lacunar": "Return 1 if lacunar infarcts (lacunes) are present, 0 if explicitly absent. If not mentioned, return 'missing'.",

"infarct_lacunar_n": "Return the total number of lacunar infarcts. If not mentioned, return 'missing'.",
"infarct_cerebellar": "Return 1 if cerebellar infarcts (of any type) are present, 0 if explicitly absent. If not mentioned, return 'missing'.",
"infarct_cerebellar_n": "Return the total number of cerebellar infarcts. If not mentioned, return 'missing'.",
"infarct_splinter": "Return 1 if splinter infarcts are present, 0 if explicitly absent. If not mentioned, return 'missing'.",
"infarct_any": "Return 1 if any type of infarct is present, 0 if explicitly absent (no infarcts). If not mentioned, return 'missing'.",
"infarct_any_n": "Return the total number of infarcts (of any type) if specified. If not mentioned, return 'missing'.",
"infarct_location": "Return the reported location(s) of infarcts. If not mentioned, return 'missing'.",
"DWI": "Return 1 if diffusion-weighted imagine (DWI) shows diffusion restriction. Return 0 if DWI is explicitly normal (no diffusion restriction). If not mentioned, return 'missing'.",
"SWI": "Return 1 if susceptibility-weighted imaging (SWI) shows abnormalities. Return 0 if SWI explicitly stated as normal. If not mentioned, return 'missing'.",
"siderosis": "Return 1 if superficial siderosis is mentioned. Otherwise, return 0.",
"artifact": "Return 1 if imaging artifacts (e.g., motion) affecting interpretation are mentioned. Otherwise, return 0.",
"quantib": "Indicate 1 if Quantib segmentation (or quantitative evaluation) is mentioned. otherwise return 0."

User part:

Radiology Report:

{the report to be structured}

B. Variable definitions and distributions

Table S1: Variable definitions and distributions for the total dataset and the 100 double-annotated reports.

Variable	Definition	Distribution		
Categorical variables		Value	N: total dataset (n=947)	N: double-annotated subset (n=100)
MTA Left	Medial temporal atrophy score of the left side	missing	285	33
		0	135	10
		0-1	16	2
		1	232	22
		1-2	43	2
		2	160	18
		2-3	16	1
		3	52	12
		3-4	1	0
		4	7	0
MTA Right	Medial temporal atrophy score of the right side	missing	284	33
		0	132	10
		0-1	19	2
		1	228	25
		1-2	43	1
		2	166	20
		2-3	21	2
		3	47	7
		3-4	3	0
		4	4	0
Normal hippocampus	Binary indicator set to 1 if the report only mentions normal hippocampal volume without providing a specific MTA score.	0	856	89
		1	91	11
GCA frontal	Frontal lobe global cortical atrophy (GCA) score. If a frontal lobe-specific score is unavailable, the whole-brain GCA score may be used instead.	missing	282	35
		0	81	8
		0-1	29	3
		1	258	20
		1-2	52	4
		2	197	24
		2-3	22	1
		3	26	5
GCA temporal	Global cortical atrophy score of the temporal lobe. If a temporal lobe-specific score is unavailable, the whole-brain GCA score may be used instead.	missing	305	36
		0	98	7
		0-1	31	3
		1	269	23
		1-2	46	4
		2	145	21
		2-3	19	2
		3	34	4
GCA occipital	Global cortical atrophy score of the occipital lobe. If an occipital lobe-specific score is unavailable, the whole-brain GCA score may be used instead.	missing	361	41
		0	134	8
		0-1	34	4

		1	275	23
		1-2	35	3
		2	89	17
		2-3	7	1
		3	12	3
GCA parietal	Global cortical atrophy score of the parietal lobe. If a parietal lobe-specific score is unavailable, the whole-brain GCA score may be used instead.	missing	256	34
		0	58	7
		0-1	28	3
		1	208	15
		1-2	47	6
		2	284	29
		2-3	30	1
		3	36	5
GCA overall	Global cortical atrophy score of the whole brain. If the score is given separately for different regions of the brain without specifying a global score, the largest of the partial values is reported.	missing	239	31
		0	48	7
		0-1	29	3
		1	208	22
		1-2	39	2
		2	285	26
		2-3	33	3
		3	66	6
Absence of atrophy	Binary indicator set to 1 if the report only mentions no atrophy (or similar) without providing a specific GCA score	0	901	93
		1	46	7
Fazekas	Fazekas score as mentioned in the report. If the score is given separately for the left and right side, the largest of the two values is reported.	missing	270	31
		0	163	9
		0-1	12	2
		1	274	30
		1-2	24	2
		2	120	12
		2-3	14	4
		3	70	10
Absence of WMH	Binary indicator set to 1 if the report mentions the absence of white matter hyperintensity without providing a specific Fazekas score.	0	913	95
		1	34	5
Normal signal intensity	Binary indicator set to 1 if the report mentions a normal signal intensity of the white and gray matter (or similar).	0	876	92
		1	71	8
Presence of cortical infarcts	Whether the report mentions the presence or absence of cortical infarcts in the brain. If cortical infarcts are not mentioned the variable is considered missing.	missing	447	64
		absent	465	32
		present	35	4
Presence of lacunes	Whether the report mentions the presence or absence of lacunes in the brain. If cortical infarcts are not mentioned the variable is considered missing.	missing	389	51
		absent	423	32
		present	135	17
Presence of cerebellar infarcts	Whether the report mentions the presence or absence of cerebellar infarcts in the brain. If cerebellar infarcts are not mentioned the variable is considered missing.	missing	504	62
		absent	396	28
		present	47	10
		missing	529	69

Presence of splinter infarcts	Whether the report mentions the presence or absence of splinter infarcts (small oblique cerebellar infarcts) in the brain. If splinter infarcts are not mentioned the variable is considered missing.	absent	402	28
		present	16	3
Presence of any infarcts	Whether the report mentions the presence or absence of any type of infarcts in the brain. If infarcts are not mentioned the variable is considered missing.	missing	412	48
		absent	364	29
		present	171	23
Presence of microbleeds	Whether the report mentions the presence or absence of microbleeds in the brain. If microbleeds are not mentioned the variable is considered missing.	missing	369	50
		absent	430	31
		present	148	19
SWI abnormalities	Whether the report mentions the presence or absence of susceptibility imaging abnormalities. If the SWI is not mentioned the variable is considered missing.	missing	848	80
		absent	68	11
		present	31	9
DWI abnormalities	Whether the report mentions the presence or absence of diffusion restriction on the DWI sequence. If the DWI is not mentioned the variable is considered missing.	missing	387	52
		absent	527	44
		present	33	4
Siderosis	Whether the report mentions the presence or absence of superficial siderosis. If superficial siderosis is not mentioned it is considered absent.	absent	930	99
		present	17	1
Quantitative evaluation	Indicator set to 1 if the report mentions the use of quantitative evaluation using the Quantib software.	0	686	98
		1	261	2
Artifacts	Indicator set to 1 if the report mentions the presence of imaging artifacts (e.g. motion) hindering interpretation.	0	911	91
		1	36	9
Numerical Variables		Statistic	Total dataset	Double-annotated subset
Cortical infarcts	Number of cortical infarcts. If not mentioned the value is set to -1.	Number missing	459	64
		Number = 0	465	32
		Mean of non-missing	0.06	0.14
		STD of non-missing	0.32	0.42
Lacunes	Number of lacunes. If not mentioned the value is set to -1.	Number missing	429	57
		Number = 0	422	32
		Mean of non-missing	0.47	0.60
		STD of non-missing	1.24	1.37
Cerebellar infarcts	Number of cerebellar infarcts. If not mentioned the value is set to -1.	Number missing	520	65
		Number = 0	398	28
		Mean of non-missing	0.11	0.31
		STD of non-missing	0.49	0.90
All infarcts	Number of all infarcts. If not mentioned the value is set to -1.	Number missing	495	55
		Number = 0	368	29
		Mean of non-missing	0.43	0.78
		STD of non-missing	1.23	1.41
Microbleeds	Number of microbleeds. If not mentioned the value is set to -1.	Number missing	417	59
		Number = 0	430	31
		Mean of non-missing	0.68	1.85
		STD of non-missing	3.61	7.26
Location variables (text fields)		Statistic	Total dataset	Double-annotated subset

Infarcts	Text description of the locations of all infarcts.	N non-missing	165	23
Microbleeds	Text description of the locations of all microbleeds.	N non-missing	146	19

N: number of samples

C. Results for quality control variables

Table S2: Quantitative performance of the LLaMA 3.1 against two human raters and the inter-rater reliability of the two raters on the QC variables. A consensus rating was used as a reference standard.

	LLaMA 3.1	**Annotator 1**	**Annotator 2**	**Inter-rater reliability**
Categorical Variables	**Balanced Accuracy (%)**			**Cohen's κ**
QC Variables				
Normal hippocampus	0.86	0.99*	0.91*	0.84
Absence of atrophy	0.68	0.92	1.00	0.78
Absence of WMH	0.77	0.99	0.70	0.43
Normal signal intensity	0.82	0.81⁑	0.92*	0.43
Quantitative evaluation	0.75	0.75	1.00	0.66
Presence of artifacts	0.97	0.89	0.78	0.47

ICC: Intra-class correlation; (*) The rater significantly outperforms the model, and (⁑) The model significantly outperforms the rater, both based on the McNemar test ($p < 0.05$).

Table S3: Zero-shot information extraction performance on original Dutch reports vs. English translations by LLaMA 3.1 and OPUS-MT on the QC variables.

	Dutch report	**LLaMA 3.1 Translation**	**OPUS-MT Translation**
Categorical Variables (%)	**Balanced Accuracy (%)**		
QC Variables			
Normal hippocampus	0.84 (0.81, 0.87)	0.82 (0.80, 0.85)	0.78 (0.74, 0.83)
Absence of atrophy	0.74 (0.65, 0.82)	0.76 (0.68, 0.84)	0.72 (0.65, 0.79)
Absence of WMH	0.85 (0.77, 0.94)	0.82 (0.74, 0.91)	0.77 (0.69, 0.84)
Normal signal intensity	0.77 (0.75, 0.79)	0.78 (0.75, 0.80)	0.74 (0.69, 0.79)
Quantitative evaluation	0.88 (0.86, 0.91)	**0.96 (0.95, 0.97)**	<u>0.70 (0.65, 0.75)</u>
Presence of artifacts	0.97 (0.96, 0.98)	0.97 (0.96, 0.97)	0.91 (0.84, 0.99)

Scores are reported as: value (95%-CI). **Bold** means translation provides significantly better performance. <u>Underlined</u> means translation provides significantly worse performance.

Table S4: Few-shot information extraction performance using fixed examples, randomly selected examples, structural (Jaccard) similarity-based examples and semantic similarity-based examples.

Selection method	**Zero-shot**	**Random**	**Fixed**	**Structural Similarity**	**Semantic Similarity**
Categorical Variables	**Balanced Accuracy (%)**				
QC Variables					
Normal hippocampus	0.84 (0.81, 0.87)	0.82 (0.79, 0.86)	0.84 (0.74, 0.94)	0.84 (0.79, 0.89)	0.85 (0.82, 0.88)
Absence of atrophy	0.74 (0.65, 0.82)	0.77 (0.69, 0.86)	0.78 (0.69, 0.86)	0.79 (0.69, 0.90)	0.79 (0.71, 0.86)
Absence of WMH	0.85 (0.77, 0.94)	0.79 (0.71, 0.86)	0.80 (0.63, 0.98)	0.84 (0.74, 0.95)	0.82 (0.68, 0.96)
Normal signal intensity	0.77 (0.75, 0.79)	0.77 (0.69, 0.86)	0.81 (0.77, 0.84)	0.77 (0.70, 0.85)	0.79 (0.72, 0.87)
Quantitative evaluation	0.88 (0.86, 0.91)	0.92 (0.89, 0.94)	0.86 (0.70, 1.00)	0.95 (0.90, 1.00)	0.94 (0.90, 0.97)
Presence of artifacts	0.97 (0.96, 0.98)	0.97 (0.94, 1.00)	0.96 (0.92, 1.00)	0.97 (0.97, 0.98)	0.96 (0.92, 0.99)

D. Comparison between LLaMA 3.1 70B and LLaMA 3.1 8B

Table S5: Zero-shot performance comparison between the 70B and 8B parameters versions of LLaMA 3.1

	LLaMA 3.1 70B	**LLaMA 3.1 8B**
Categorical Variables	**Balanced Accuracy**	
Atrophy Measures		
MTA left	0.90 (0.77, 1.00)	0.81 (0.70, 0.93)
MTA right	**0.96 (0.94, 0.99)**	0.86 (0.80, 0.93)
GCA overall	**0.87 (0.83, 0.91)**	0.71 (0.65, 0.76)
GCA frontal	**0.80 (0.73, 0.87)**	0.65 (0.59, 0.71)
GCA temporal	**0.73 (0.69, 0.78)**	0.65 (0.62, 0.68)
GCA occipital	0.58 (0.47, 0.69)	0.41 (0.34, 0.48)
GCA parietal	**0.77 (0.70, 0.84)**	0.59 (0.53, 0.64)
Vascular Markers		
Fazekas score	**0.94 (0.93, 0.96)**	0.82 (0.75, 0.89)
Any brain infarcts	0.82 (0.80, 0.84)	0.85 (0.82, 0.87)
Cortical infarcts	0.75 (0.69, 0.81)	0.78 (0.74, 0.83)
Lacunes	0.83 (0.80, 0.86)	0.89 (0.86, 0.91)
Cerebellar infarcts	0.69 (0.66, 0.71)	**0.80 (0.74, 0.86)**
Splinter infarcts	0.66 (0.64, 0.67)	0.77 (0.69, 0.85)
Microbleeds	0.93 (0.92, 0.95)	0.91 (0.90, 0.92)
SWI abnormalities	0.70 (0.61, 0.79)	0.59 (0.48, 0.70)
Superficial siderosis	0.97 (0.85, 1.00)	0.97 (0.85, 1.00)
DWI abnormalities	0.89 (0.81, 0.96)	0.86 (0.79, 0.93)
Variables for Quality Control S		
Normal hippocampus	**0.84 (0.81, 0.87)**	0.77 (0.76, 0.79)
Absence of atrophy	0.74 (0.65, 0.82)	0.78 (0.69, 0.86)
Absence of WMH	**0.85 (0.77, 0.94)**	0.70 (0.65, 0.75)
Normal signal intensity	**0.77 (0.75, 0.79)**	0.61 (0.60, 0.62)
Quantitative evaluation	0.88 (0.86, 0.91)	0.93 (0.90, 0.97)
Presence of artifacts	0.97 (0.96, 0.98)	0.94 (0.88, 1.00)
Numerical Variables	**Accuracy [Mean Absolute Error]**	
Any brain infract	**0.66 (0.63, 0.68)** [0.04]	0.54 (0.50, 0.59) [0.48]
Cortical infarcts	**0.61 (0.58, 0.64)** [0.01]	0.51 (0.47, 0.56) [0.11]
Lacunes	**0.69 (0.66, 0.72)** [0.10]	0.52 (0.47, 0.57) [0.32]
Cerebellar infarcts	0.61 (0.58, 0.64) [0.03]	0.55 (0.52, 0.59) [0.41]
Microbleeds	**0.80 (0.78, 0.82)** [0.00]	0.54 (0.51, 0.57) [0.03]
Text Variables	**Text Similarity**	
Location of infarcts	**0.95 (0.95, 0.95)**	0.94 (0.93, 0.94)
Location of microbleeds	0.95 (0.95, 0.96)	0.94 (0.94, 0.95)

E. Impact of inter-rater disagreement on the reliability of single-rater annotations

E.1 Bootstrapping Analysis

We assessed the reliability of using individual rater annotations versus consensus annotations as a reference for evaluating model performance. To estimate variability and generate confidence intervals, we performed a bootstrap experiment with 500 iterations. In each iteration, reports were sampled with replacement to create a bootstrap sample of the same size as the original dataset. Model performance was computed in two ways for each sample:

- Consensus reference: the model predictions were compared to the consensus annotations of the bootstrap-sampled reports.
- Single-rater reference: for each bootstrap-sampled report, one of the two raters (R1 or R2) was selected randomly and independently, and the model prediction was compared to that rater's annotation.

This process was repeated across all iterations, producing distributions of performance metrics for each variable. The mean performance and 95% confidence intervals were then calculated across bootstrap iterations. The results are summarized in Fig. S1.

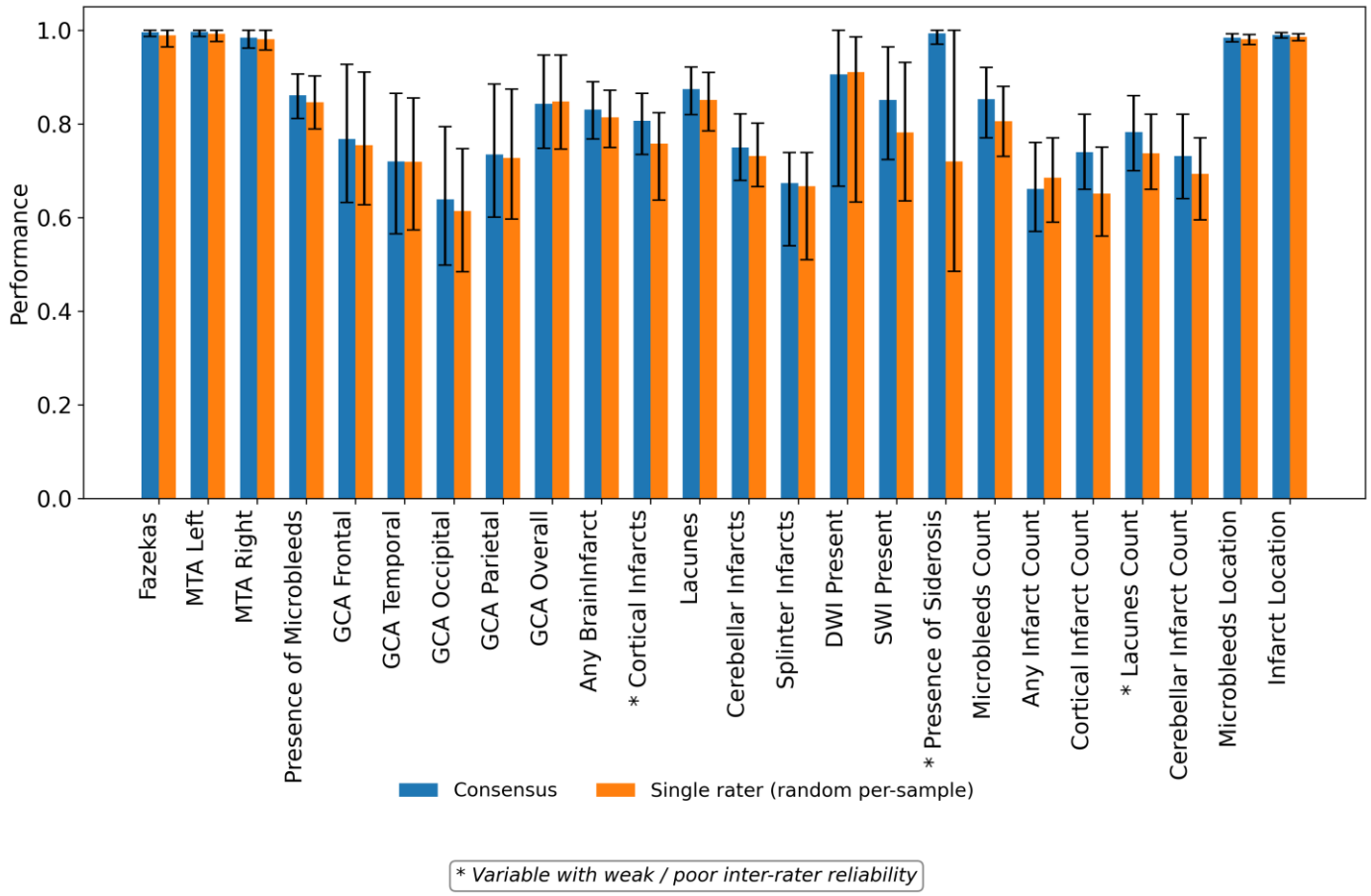

Figure S1: Comparison of performance measures obtained using the consensus annotations and single-rater annotations on the 100 double-annotated reports. Error bars show the 95% confidence intervals computed over 500 bootstrapped iterations.

Model performance was generally higher when evaluated against the consensus annotations than when evaluated against a single-rater reference chosen randomly per sample. For most variables, the mean performance differences were modest, and the 95% confidence intervals largely overlapped, indicating that the two evaluation approaches are broadly comparable.

For variables with weaker inter-rater reliability, the mean difference in performance between the consensus rating and the single-rater reference was similar to or larger than for other variables, reflecting the influence of rater variability. Categorical variables such as Fazekas scores and MTA ratings exhibited near-perfect performance in both conditions. Variables with lower prevalence or more subjective assessment (e.g., cortical infarcts, splinter infarcts, presence of superficial siderosis) showed wider confidence intervals with single-rater evaluation. Overall, these results indicate that while using consensus annotations provides a slightly more stable reference, evaluation with randomly chosen single-rater annotations yields comparable performance estimates across most variables.

E.2 Sensitivity Analysis by Agreement Level Stratification

We assessed how inter-rater agreement relates to model performance across the full evaluation set (n = 947). Variables were stratified into three agreement bins. Within each bin, we report the number of variables, mean IRR, and the mean and standard deviation of model performance (Table S6). To quantify the relationship between IRR and model performance, we computed Spearman's rank correlation between the IRR value and the performance metric across variables. Because variable types use different performance and IRR metrics, the primary correlation was restricted to categorical variables (n = 17), which share a single IRR metric (Cohen's κ) and a single performance metric (balanced accuracy). A 95% confidence interval for Spearman's ρ was obtained via percentile bootstrap with 1000 resamples.

Numerical variables (n = 5) are reported descriptively given the small sample size, and text variables (n = 2) are reported for completeness. The relationship across all 24 variables is visualized in Fig. S2.

Table S6: Model performance stratified by inter-rater agreement level. Bins are defined by field-standard thresholds for each IRR metric (Cohen's κ, ICC, or text similarity). Performance refers to balanced accuracy for categorical variables, accuracy for numerical variables, and semantic similarity for text variables.

Variable type	Bin	N variables	Mean IRR	Performance (mean ± std)
Categorical	High (κ ≥ 0.80)	10	0.93	0.82 ± 0.11
	Moderate (0.60 ≤ κ < 0.80)	5	0.72	0.77 ± 0.12
	Low (κ < 0.60)	2	0.49	0.86 ± 0.15
Numerical	High (ICC ≥ 0.75)	2	0.90	0.70 ± 0.13
	Moderate (0.50 ≤ ICC < 0.75)	2	0.68	0.63 ± 0.03
	Low (ICC < 0.50)	1	0.45	0.69
Text field	High (similarity ≥ 0.90)	2	0.94	0.95 ± 0.00

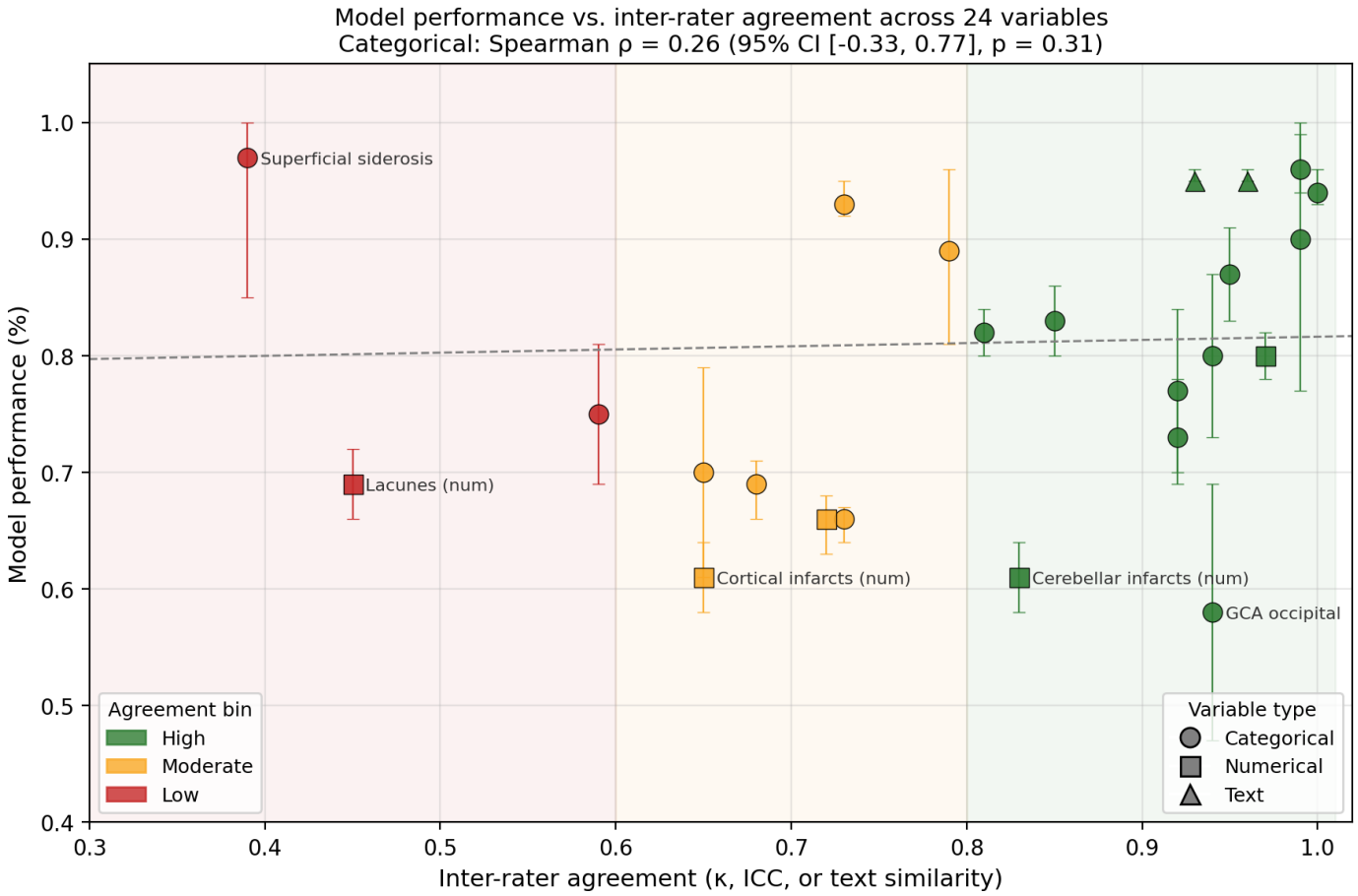

Fig S2: Model performance versus inter-rater agreement across all 24 variables. Marker shape indicates variable type (categorical, numerical, text), and color indicates agreement bin (High, Moderate, Low) based on the categorical variables. Error bars show 95% confidence intervals on performance. The dashed line is a linear fit to the categorical variables, shown for visual reference.

Across categorical variables, we found no statistically significant correlation between Cohen's κ and balanced accuracy (Spearman ρ = 0.26, 95% CI [−0.34, 0.77], p = 0.31). Mean balanced accuracy was comparable across the three agreement bins, and performance in the Low-agreement bin was not lower than in the High-agreement bin. This pattern reflects the sensitivity

of κ to label skew: for rare findings such as superficial siderosis (only one of the 100 double-annotated reports labeled positive), κ is negatively affected by class imbalance despite high raw agreement, and the model achieves high balanced accuracy. For numerical variables, the Spearman correlation between ICC and accuracy was similarly non-significant (ρ = 0.21, p = 0.74, n = 5), although this estimate should be interpreted as indicative given the small number of variables. Both text variables fell in the High-agreement bin and achieved high similarity scores. Overall, these results indicate that model performance is not systematically degraded on variables with weaker inter-rater agreement, though this conclusion is tempered by the small number of variables in the Low-agreement bin. Together with the bootstrap analysis above, this suggests that annotation uncertainty, while a legitimate concern for individual low-agreement variables, does not broadly compromise the reported performance estimates.

F. Results for spliter infarcts-specific translation

Table S7: Extraction performance obtained using the general LLaMA 3.1 translation and the LLaMA 3.1 translation with specific instructions for splinter infarcts.

	General translation	**Splinter infarcts-specific translation**
Categorical Variables	**Balanced Accuracy**	
Atrophy Measures		
MTA left	0.90 (0.77, 1.00)	0.91 (0.79, 1.00)
MTA right	0.96 (0.94, 0.99)	0.95 (0.88, 1.00)
GCA overall	0.88 (0.84, 0.91)	0.88 (0.84, 0.91)
GCA frontal	0.81 (0.75, 0.87)	0.81 (0.75, 0.87)
GCA temporal	0.78 (0.73, 0.82)	0.76 (0.71, 0.80)
GCA occipital	0.61 (0.51, 0.72)	0.60 (0.51, 0.69)
GCA parietal	0.78 (0.73, 0.83)	0.79 (0.74, 0.83)
Vascular Markers		
Fazekas score	0.94 (0.93, 0.96)	0.94 (0.92, 0.96)
Any brain infarcts	0.81 (0.79, 0.84)	0.82 (0.80, 0.85)
Cortical infarcts	0.74 (0.67, 0.81)	0.75 (0.69, 0.81)
Lacunes	0.85 (0.82, 0.88)	0.84 (0.79, 0.88)
Cerebellar infarcts	0.68 (0.65, 0.71)	0.69 (0.65, 0.73)
Splinter infarcts	*0.37 (0.31, 0.44)*	*0.65 (0.59, 0.70)*
Microbleeds	0.94 (0.92, 0.95)	0.94 (0.93, 0.96)
SWI abnormalities	0.75 (0.68, 0.82)	0.73 (0.64, 0.82)
Presence of siderosis	0.94 (0.80, 1.00)	0.94 (0.80, 1.00)
DWI abnormalities	0.85 (0.77, 0.94)	0.85 (0.76, 0.94)
Numerical Variables	**Accuracy [Mean Absolute Error]**	
Cortical infarcts	0.62 (0.59, 0.66) [0.01]	0.62 (0.59, 0.65) [0.05]
Lacunes	0.75 (0.71, 0.78) [0.08]	0.70 (0.67, 0.74) [0.07]
Cerebellar infarcts	0.62 (0.58, 0.67) [0.03]	0.61 (0.56, 0.65) [0.05]
All infarcts	0.67 (0.63, 0.70) [0.04]	0.65 (0.62, 0.69) [0.07]
Microbleeds	0.81 (0.78, 0.84) [0.00]	0.77 (0.75, 0.80) [0.00]
Text Variables	**Text Similarity**	
Location of infarcts	0.93 (0.92, 0.93)	0.93 (0.92, 0.93)
Location of microbleeds	0.91 (0.90, 0.91)	0.91 (0.90, 0.92)

G. Evaluation using additional metrics

Table S8: Performance of the LLaMA 3.1 and two human raters against the consensus annotations measured using F1-Score for categorical and Spearman's correlation for numerical variables

	LLaMA 3.1 vs Consensus	Rater 1 vs Consensus	Rater 2 vs Consensus
Categorical Variables	**F1-Score**		
Atrophy Measures			
MTA left	1.00	0.99	1.00
MTA right	0.98	0.99	1.00
GCA overall	0.76	0.94	1.00
GCA frontal	0.76	0.96	0.92
GCA temporal	0.79	0.96	0.94
GCA occipital	0.67	0.96	0.90
GCA parietal	0.74	0.97	0.92
Vascular Markers			
Fazekas score	0.97	0.99	1.00
Any brain infarct	0.80	0.92	0.91
Cortical infarcts	0.71	0.94	0.81
Lacunes	0.84	0.98	0.94
Cerebellar infarcts	0.74	0.88	0.91
Splinter infarcts	0.68	0.88	0.94
Microbleeds	0.82	0.90	0.95
SWI abnormalities	0.87	0.92	0.82
Superficial siderosis	0.83	0.69	1.00
DWI abnormalities	0.94	0.78	0.95
Numerical Variables	**Spearman's Correlation [number of extracted entries]**		
Any brain infarct	1.00 [13]	0.94 [34]	0.93 [42]
Cortical infarcts	NA [11]	0.82 [35]	1.00 [35]
Lacunes	0.93 [25]	1.00 [41]	1.00 [42]
Cerebellar infarcts	1.00 [11]	0.89 [32]	1.00 [34]
Microbleeds	1.00 [27]	0.95 [37]	1.00 [39]

Table S9: Zero-shot information extraction performance on original Dutch reports vs. English translations by LLaMA 3.1 and OPUS-MT measured using F1-score for categorical and Spearman's correlation for numerical variables.

	Dutch report	**LLaMA 3.1 Translation**	**OPUS-MT Translation**
Categorical Variables (%)	**F1-Score**		
Atrophy Measures			
MTA left	0.83 (0.66, 1.00)	0.81 (0.70, 0.93)	0.73 (0.56, 0.90)
MTA right	0.91 (0.80, 1.00)	0.88 (0.86, 0.90)	0.79 (0.67, 0.91)
GCA overall	0.79 (0.76, 0.82)	0.79 (0.77, 0.81)	0.73 (0.71, 0.75)
GCA frontal	0.75 (0.64, 0.86)	0.77 (0.66, 0.87)	0.69 (0.57, 0.80)
GCA temporal	0.71 (0.67, 0.75)	0.76 (0.65, 0.88)	0.64 (0.61, 0.68)
GCA occipital	0.58 (0.47, 0.68)	0.67 (0.55, 0.79)	0.53 (0.46, 0.61)
GCA parietal	0.72 (0.67, 0.77)	0.73 (0.69, 0.77)	0.65 (0.55, 0.76)
Vascular Markers			
Fazekas score	0.90 (0.86, 0.93)	0.91 (0.87, 0.95)	0.77 (0.63, 0.90)
Any brain infarct	0.80 (0.78, 0.82)	0.80 (0.78, 0.82)	0.67 (0.65, 0.70)
Cortical infarcts	0.71 (0.65, 0.78)	0.70 (0.65, 0.75)	0.60 (0.55, 0.65)
Lacunes	0.83 (0.80, 0.86)	0.84 (0.81, 0.88)	0.66 (0.61, 0.70)
Cerebellar infarcts	0.65 (0.61, 0.69)	0.64 (0.60, 0.67)	0.59 (0.54, 0.64)
Splinter infarcts	0.61 (0.56, 0.65)	0.36 (0.26, 0.47)	0.53 (0.43, 0.63)
Microbleeds	0.93 (0.91, 0.95)	0.94 (0.92, 0.96)	0.65 (0.61, 0.69)
SWI abnormalities	0.73 (0.64, 0.81)	0.78 (0.72, 0.84)	0.61 (0.52, 0.70)
Presence of siderosis	0.95 (0.87, 1.00)	0.97 (0.89, 1.00)	0.79 (0.60, 0.98)
DWI abnormalities	0.91 (0.86, 0.96)	0.88 (0.82, 0.93)	0.67 (0.60, 0.74)
Numerical Variables	**Spearman's Correlation [number of entries]**		
Any brain infarct	0.94 (0.86, 1.00) [79]	0.95 (0.82, 1.00) [80]	1.00 (1.00, 1.00) [42]
Cortical infarcts	1.00 (1.00, 1.00) [67]	0.93 (0.86, 1.00) [71]	0.96 (0.90, 1.00) [38]
Lacunes	0.93 (0.84, 1.00) [132]	0.91 (0.79, 1.00) [155]	0.82 (0.66, 0.98) [55]
Cerebellar infarcts	0.94 (0.85, 1.00) [45]	0.92 (0.83, 1.00) [51]	0.86 (0.71, 1.00) [23]
Microbleeds	1.00 (1.00, 1.00) [174]	1.00 (1.00, 1.00) [180]	0.97 (0.93, 1.00) [68]

Table S10: Few-shot information extraction performance using fixed examples, randomly selected examples, structural (Jaccard) similarity-based examples and semantic similarity-based examples measured using F1-Score for categorical and Spearman's correlation for numerical variables.

Selection method	**Zero-shot**	**Random**	**Fixed**	**Structural Similarity**	**Semantic Similarity**
Categorical Variables	**F1-Score**				
Atrophy Measures					
MTA left	0.83 (0.66, 1.00)	0.86 (0.79, 0.92)	0.86 (0.71, 1.00)	0.85 (0.69, 1.00)	0.87 (0.70, 1.00)
MTA right	0.91 (0.80, 1.00)	0.94 (0.85, 1.04)	0.96 (0.84, 1.00)	0.98 (0.96, 0.99)	0.97 (0.89, 1.00)
GCA overall	0.79 (0.76, 0.82)	0.80 (0.67, 0.94)	0.85 (0.72, 0.99)	0.88 (0.81, 0.95)	0.86 (0.75, 0.97)
GCA frontal	0.75 (0.64, 0.86)	0.75 (0.63, 0.87)	0.86 (0.75, 0.96)	0.87 (0.83, 0.91)	0.84 (0.73, 0.95)
GCA temporal	0.71 (0.67, 0.75)	0.73 (0.60, 0.85)	0.84 (0.72, 0.95)	0.83 (0.79, 0.88)	0.82 (0.75, 0.88)
GCA occipital	0.58 (0.47, 0.68)	0.62 (0.49, 0.75)	0.73 (0.54, 0.91)	0.72 (0.63, 0.80)	0.71 (0.61, 0.81)
GCA parietal	0.72 (0.67, 0.77)	0.74 (0.62, 0.85)	0.80 (0.69, 0.92)	0.82 (0.72, 0.92)	0.82 (0.70, 0.94)
Vascular Markers					
Fazekas score	0.90 (0.86, 0.93)	0.90 (0.77, 1.02)	0.90 (0.81, 1.00)	0.92 (0.86, 0.97)	0.91 (0.86, 0.96)
Any brain infarct	0.80 (0.78, 0.82)	0.81 (0.76, 0.86)	0.83 (0.74, 0.92)	0.86 (0.81, 0.91)	0.84 (0.80, 0.89)
Cortical infarcts	0.71 (0.65, 0.78)	0.74 (0.68, 0.81)	0.78 (0.69, 0.86)	0.79 (0.75, 0.84)	0.78 (0.73, 0.84)
Lacunes	0.83 (0.80, 0.86)	0.84 (0.80, 0.88)	0.85 (0.77, 0.92)	0.89 (0.85, 0.93)	0.87 (0.84, 0.91)
Cerebellar infarcts	0.65 (0.61, 0.69)	0.68 (0.61, 0.76)	0.74 (0.68, 0.81)	0.79 (0.73, 0.86)	0.75 (0.71, 0.78)
Splinter infarcts	0.61 (0.56, 0.65)	0.73 (0.66, 0.80)	0.77 (0.72, 0.82)	0.83 (0.81, 0.86)	0.80 (0.74, 0.85)
Microbleeds	0.93 (0.91, 0.95)	0.95 (0.93, 0.96)	0.95 (0.93, 0.97)	0.94 (0.93, 0.96)	0.94 (0.92, 0.96)
SWI abnormalities	0.73 (0.64, 0.81)	0.72 (0.66, 0.77)	0.66 (0.45, 0.87)	0.75 (0.69, 0.80)	0.71 (0.61, 0.82)
Superficial Siderosis	0.95 (0.87, 1.00)	0.83 (0.52, 1.15)	0.93 (0.81, 1.00)	0.87 (0.58, 1.00)	0.88 (0.64, 1.00)
DWI abnormalities	0.91 (0.86, 0.96)	0.90 (0.85, 0.96)	0.87 (0.60, 1.00)	0.93 (0.87, 0.98)	0.93 (0.89, 0.97)
Numerical Variables	**Spearman's Correlation [number of entries]**				
Any brain infarct	0.94 (0.86, 1.00) [79]	0.86 (0.71, 1.00) [182]	0.88 (0.79, 0.97) [203]	0.80 (0.57, 1.00) [198]	0.85 (0.59, 1.00) [196]
Cortical infarcts	1.00 (1.00, 1.00) [67]	0.93 (0.86, 1.00) [181]	0.94 (0.86, 1.00) [207]	0.93 (0.85, 1.00) [204]	0.94 (0.87, 1.00) [202]
Lacunes	0.93 (0.84, 1.00) [132]	0.88 (0.77, 0.99) [225]	0.92 (0.80, 1.00) [234]	0.86 (0.74, 0.97) [237]	0.85 (0.70, 0.99) [232]
Cerebellar infarcts	0.94 (0.85, 1.00) [45]	0.92 (0.87, 0.97) [132]	0.94 (0.88, 1.00) [167]	0.91 (0.85, 0.97) [180]	0.90 (0.82, 0.98) [160]
Microbleeds	1.00 (1.00, 1.00) [174]	1.00 (0.98, 1.00) [249]	1.00 (0.98, 1.00) [252]	1.00 (0.98, 1.00) [246]	1.00 (0.99, 1.00) [248]

H. Qualitative examples

Table S11: Illustrative examples with reports, their manual annotations and the output of LLaMA 3.1 obtained using few-shot prompting with structural similarity-based example selection.

Report (Translated with LLaMA 3.1)	Variable	Manual extraction	Model output
Example 1			
Medical History: Acute depression in [date X] with cognitive complaints, followed by gradual recovery. Elsewhere interpreted as possible Alzheimer's. **Question:** Kindly re-evaluation of the MRI brain scan of [date Y], degree of atrophy. **Description:** Re-evaluation of MRI examination performed elsewhere on [date Y], without contrast. Infratentorial normal volume of the brainstem and cerebellar brain parenchyma. Supratentorial thin ventricular system. GCA 0-1 in all brain lobes. MTA 0 bilaterally. Punctiform subcortical white matter lesions, Fazekas 1. No cortical or lacunar infarcts. No diffusion restrictions. No microbleeds. **Conclusion:** No notable atrophy. No pronounced vascular lesions. **Original report:** Klinische Gegevens: vrij acute depressie met cognitieve klachten, daarna geleidelijk herstel. Elders geduid als mogelijke Alzheimer. Vraagstelling: Gaarne herbeoordeling MRI-Cerebrum d.d. [date X], mate van atrofie? Beschrijving: Herbeoordeling van MRI-onderzoek verricht elders op [date Y]. Betreft een onderzoek zonder contrast. Infratentorieel normaal volume van de hersenstam en cerebellaire hersenparenchym. Supratentorieel slank ventrikelsysteem. GCA 0-1 in alle hersenkwabben. MTA 0 beiderzijds. Punctiforme subcorticale witte stoflaesies, Fazekas 1. Geen corticale of lacunaire infarcten. Geen diffusie restricties. Geen microbloedingen. Conclusie: Geen opvallende atrofie. Geen uitgesproken vasculaire laesies.	MTA left	0	0
	MTA right	0	0
	GCA overall	0-1	0-1
	GCA frontal	0-1	0-1
	GCA temporal	0-1	0-1
	GCA occipital	0-1	0-1
	GCA parietal	0-1	0-1
	Fazekas score	1	1
	Any brain infarcts	Absent	Absent
	Cortical infarcts	Absent	Absent
	Lacunes	Absent	Absent
	Cerebellar infarcts	Absent	Absent
	Splinter infarcts	Absent	Absent
	Microbleeds	Absent	Absent
	SWI abnormalities	Missing	Missing
	Presence of siderosis	No	No
	DWI abnormalities	Absent	Absent
	# any brain infarct	0	0
	# cortical infarcts	0	0
	# lacunes	0	0
	# cerebellar infarcts	0	0
	# microbleeds	0	0
	Infarct location	Missing	Missing
	Microbleed location	Missing	Missing
Example 2			
CLINICAL FINDINGS: Second opinion due to cognitive complaints. No strong suspicion of dementia syndrome anamnestically. **QUESTION:** Re-evaluation of external imaging MRI-Cerebrum [date X] and PET-CT [date Y]. Atrophy? **DESCRIPTION**: Re-evaluation of MRI brain performed elsewhere on [date X]. There is a normal volume of infratentorial and supratentorial brain parenchyma, GCA 0-1. Slender ventricular system. There are no white matter lesions. There are no indications of past infarcts. MTA 0 bilaterally. No indications of precuneus atrophy. No diffusion restrictions. No microbleedings. **CONCLUSION:** No pattern of atrophy. No vascular lesions. No structural abnormalities. **Original report**: KLINISCHE GEGEVENS: 2nd opinion vanwege cognitieve klachten. Anamnestisch geen sterke verdenking op dementieel syndroom. VRAAGSTELLING: Herbeoordeling externe beeldvorming MRI-Cerebrum [date X] en PET-CT	MTA left	0	0
	MTA right	0	0
	GCA overall	0-1	0-1
	GCA frontal	**0**	**0-1**
	GCA temporal	**0**	**0-1**
	GCA occipital	**0**	**0-1**
	GCA parietal	**0**	**0-1**
	Fazekas score	**0**	**Missing**
	Any brain infarcts	**Missing**	**Absent**
	Cortical infarcts	**Missing**	**Absent**
	Lacunes	**Missing**	**Absent**
	Cerebellar infarcts	**Missing**	**Absent**
	Splinter infarcts	**Missing**	**Absent**
	Microbleeds	Absent	Absent
	DWI abnormalities	Absent	Absent
	Presence of siderosis	No	No

[date Y]. Atrofie? BESCHRIJVING: Herbeoordeling van MRI hersenen verricht elders op [date X]. Er is een normaal volume van infratentorieel en supratentorieel hersenparenchym, GCA 0-1. Slank ventrikelsysteem. Er zijn geen witte stoflaesies. Er zijn geen aanwijzingen voor doorgemaakte infarcten. MTA 0 beiderzijds. Geen aanwijzingen voor precuneus atrofie. Geen diffusie restricties. Geen microbloedingen. CONCLUSIE: Geen patroon van atrofie. Geen vasculaire laesies. Geen structurele afwijkingen	SWI abnormalities	Missing	Missing
	# any brain infarcts	**-1**	**0**
	# cortical infarcts	**-1**	**0**
	# lacunes	**-1**	**0**
	# cerebellar infarcts	**-1**	**0**
	# microbleeds	0	0
	Infarct location	Missing	Missing
	Microbleed location	Missing	Missing
Example 3			
Clinical information: Suspicion of FTD. Question of whether there was MS in the past. Request for re-evaluation of MRI [date X and Y]. Do the abnormalities match FTD or MS? **Report:** Re-evaluation of MRI according to dementia protocol compared to the [date Y] study. Minimal periventricular white matter abnormalities. No evident white matter abnormalities elsewhere. Left high frontal and several hypointensities on the gradient echo consistent with small hemorrhagic components (image 23 series 5). No areas with diffusion restriction. GCA score: 1, especially the wide fissure of Silvii. No evident asymmetry. MTA score 4 Fazekas score 0. No evident MS lesions. **Conclusion:** The image is more suggestive of Alzheimer's disease due to the significant hippocampus atrophy. MTA 4 No evident signs of FTD or MS. **Original report:** Klinische informatie. Verdenking FTD. Zou in voorgeschiedenis MS hebben. Vraagstelling Graag herbeoordeling MRI [date X and Y]. Afwijkingen passen bij FTD wel bij MS? Verslag. Herbeoordeling MRI volgens dementie protocol vergeleken met onderzoek uit [date Z]. Geringe periventriculaire wittestof afwijkingen. Elders geen evidente witte stofafwijkingen. Links hoog frontaal enkele hypo-intensiteiten op de gradient echo passend bij kleine bloedingscomponenten (beeld 23 serie 5). Geen gebieden met diffusierestrictie. GCA score: 1 Met name wat wijde fissura van Silvii. Geen evidente asymmetrie. MTA score 4 Fazekas score 0. Geen evidente MS laesies. Conclusie: Beeld eerder passend bij Alzheimer disease gezien de forse hippocampus atrofie. MTA 4 Geen evidente tekenen van FTD of MS.	MTA left	4	4
	MTA right	4	4
	GCA overall	1	1
	GCA frontal	**1**	**Missing**
	GCA temporal	**1**	**Missing**
	GCA occipital	**1**	**Missing**
	GCA parietal	**1**	**Missing**
	Fazekas score	0	0
	Any brain infarcts	**Missing**	**Absent**
	Cortical infarcts	**Missing**	**Absent**
	Lacunes	**Missing**	**Absent**
	Cerebellar infarcts	**Missing**	**Absent**
	Splinter infarcts	**Missing**	**Absent**
	Microbleeds	**Missing**	**Present**
	SWI abnormalities	Missing	Missing
	Presence of siderosis	No	No
	# any brain infarcts	**-1**	**0**
	# cortical infarcts	**-1**	**0**
	# lacunes	**-1**	**0**
	# cerebellar infarcts	**-1**	**0**
	# microbleeds	**-1**	**0**
	Infarct location	missing	Missing
	Microbleed location	**missing**	**left high frontal**
Example 4			
Clinical Data: [sensitive] **Question**: atrophy, hippocampal atrophy, infarct? RIP? metastases? subdural hematoma? **Description**: MRI brain according to metastasis protocol. Evaluation is complicated by movement artifacts. Cortical atrophy: Left: Frontal GCA: 1 Parietal GCA: 2 Occipital GCA: 1 Temporal GCA: 2-3 Right: Frontal GCA: 1 Parietal GCA: 2 Occipital GCA: 1 Temporal GCA: 2-3 **Medial temporal lobe atrophy (MTA): Not formally assessable due to lack of blank T1 weighted series in the correct plane. Based on the available images, there is moderate mesotemporal atrophy on the right and mild mesotemporal atrophy on the left**. Aspect cerebellum: No evident atrophy. Aspect brainstem: No evident atrophy. Ventricular system: Central liquor system is widened, this is in proportion to the peripheral liquor spaces and the white matter abnormalities. White matter abnormalities: Bilateral extensive periventricular and	MTA left	**Missing**	**2**
	MTA right	**Missing**	**1**
	GCA overall	2.5	2.5
	GCA frontal	1	1
	GCA temporal	2.5	2.5
	GCA occipital	1	1
	GCA parietal	2	2
	Fazekas score	**2**	**2**
	Any brain infarcts	Absent	Absent
	Cortical infarcts	Absent	Absent
	Lacunes	Absent	Absent
	Cerebellar infarcts	Absent	Absent
	Splinter infarcts	**Absent**	**Missing**

to a lesser extent subcortical, Fazekas 2. Less than 25% of the white matter is involved. Lacunar or strategic infarcts: None DWI diffusion restriction: None No positive indications for bleeding, however not assessable for microbleeding due to lack of gradient echo or Swan sequence. Quantitative / Quantib: Not possible due to lack of 3D T1 series without contrast. Other: No adjacent lesions, no suspicion of metastases (due to movement artifacts a very small pointiform adjacent sign cannot be ruled out). **Conclusion:** [...] **Original report:** Klinische Gegevens: [sensitive] Vraagstelling: atrofie, hippocampusatrofie, infarct? rip? metastasen? subduraal hematoom? Beschrijving: MRI hersenen volgens metastase protocol. Beoordeling wordt bemoeilijkt door bewegingsartefacten. Corticale atrofie: Links: Frontaal GCA: 1 Parietaal GCA: 2 Occipitaal GCA: 1 Temporaal GCA: 2-3 Rechts: Frontaal GCA: 1 Parietaal GCA: 2 Occipitaal GCA: 1 Temporaal GCA: 2-3 Medial temporal lobe atrofie (MTA): Niet formeel te beoordelen door ontbreken van blanco T1 gewogen serie in het juiste vlak. Op basis van de beschikbare beelden is er rechts lichte en links een matige mesotemporale atrofie. Aspect cerebellum: Geen evidente atrofie. Aspect hersenstam: Geen evidente atrofie. Ventrikelsysteem: Centrale liquorsysteem is verwijd, dit is in verhouding tot de perifere liquorruimten en de witte stof afwijkingen. Witte stofafwijkingen: Beiderzijds uitgebreid periventriculair en in mindere mate subcorticaal, Fazekas 2. Net minder dan 25% van de witte stof is betrokken. Lacunaire of strategische infarcten: Geen DWI diffusierestrictie: Geen Geen positieve aanwijzingen voor bloedingen, echter niet te beoordelen op microbloedingen bij ontbreken van gradiëntecho of Swan sequentie . Kwantitatief / Quantib: Niet mogelijk bij ontbreken 3D T1 serie zonder contrast. Overig: Geen aankleurende laesies, geen verdenking op metastasen (door de bewegingsartefacten kan een zeer kleine puntiforme aankleuring niet worden uitgesloten). Conclusie: - [...]	Microbleeds	Missing	Missing
	SWI abnormalities	Missing	Missing
	Presence of siderosis	No	No
	DWI abnormalities	Absent	Absent
	# any brain infarcts	**0**	**-1**
	# cortical infarcts	**0**	**-1**
	# lacunes	**0**	**-1**
	# cerebellar infarcts	**0**	**-1**
	# microbleeds	-1	-1
	Infarct location	Missing	Missing
	Microbleed location	Missing	Missing

I. Confusion matrices

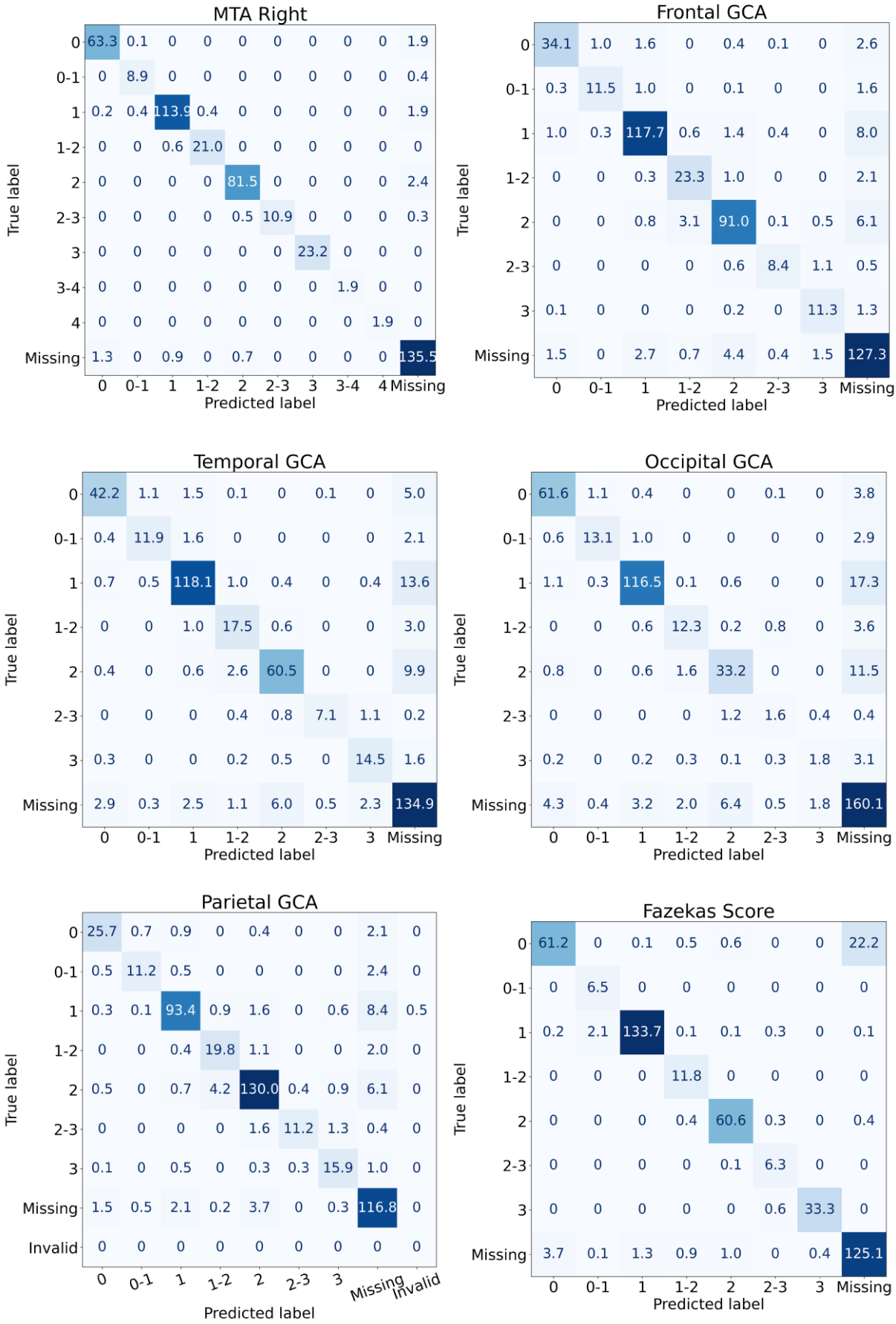

Figure S3: Confusion matrices of atrophy visual rating scores (MTA, GCA, and Fazekas) obtained using few-shot prompting with structural similarity-based example selection.

Figure S4: Confusion matrices of atrophy measures obtained using few-shot prompting with structural similarity-based example selection.

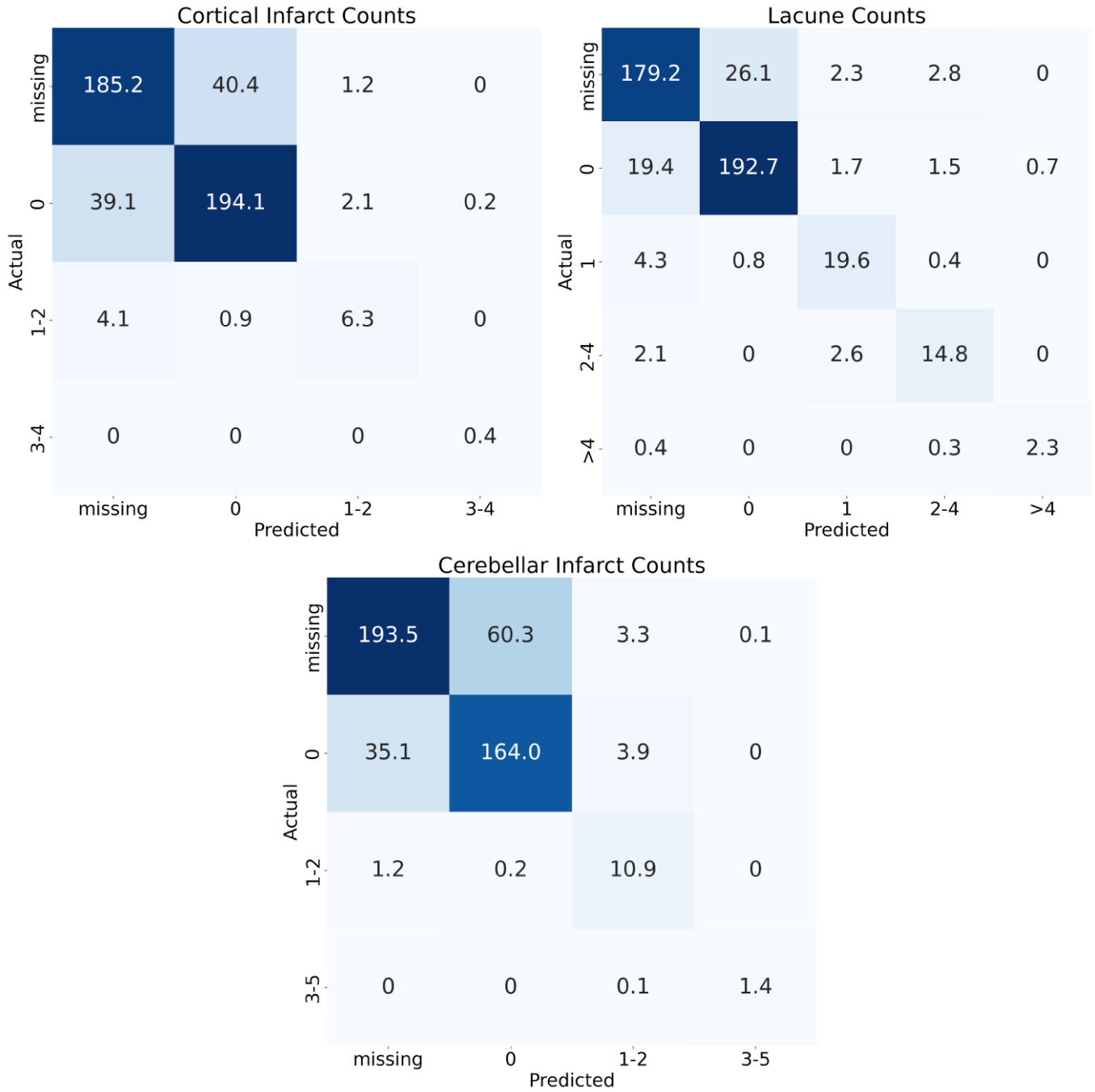

Figure S5: Confusion matrices of lesion counts obtained using few-shot prompting with structural similarity-based example selection.

J. Radar plots

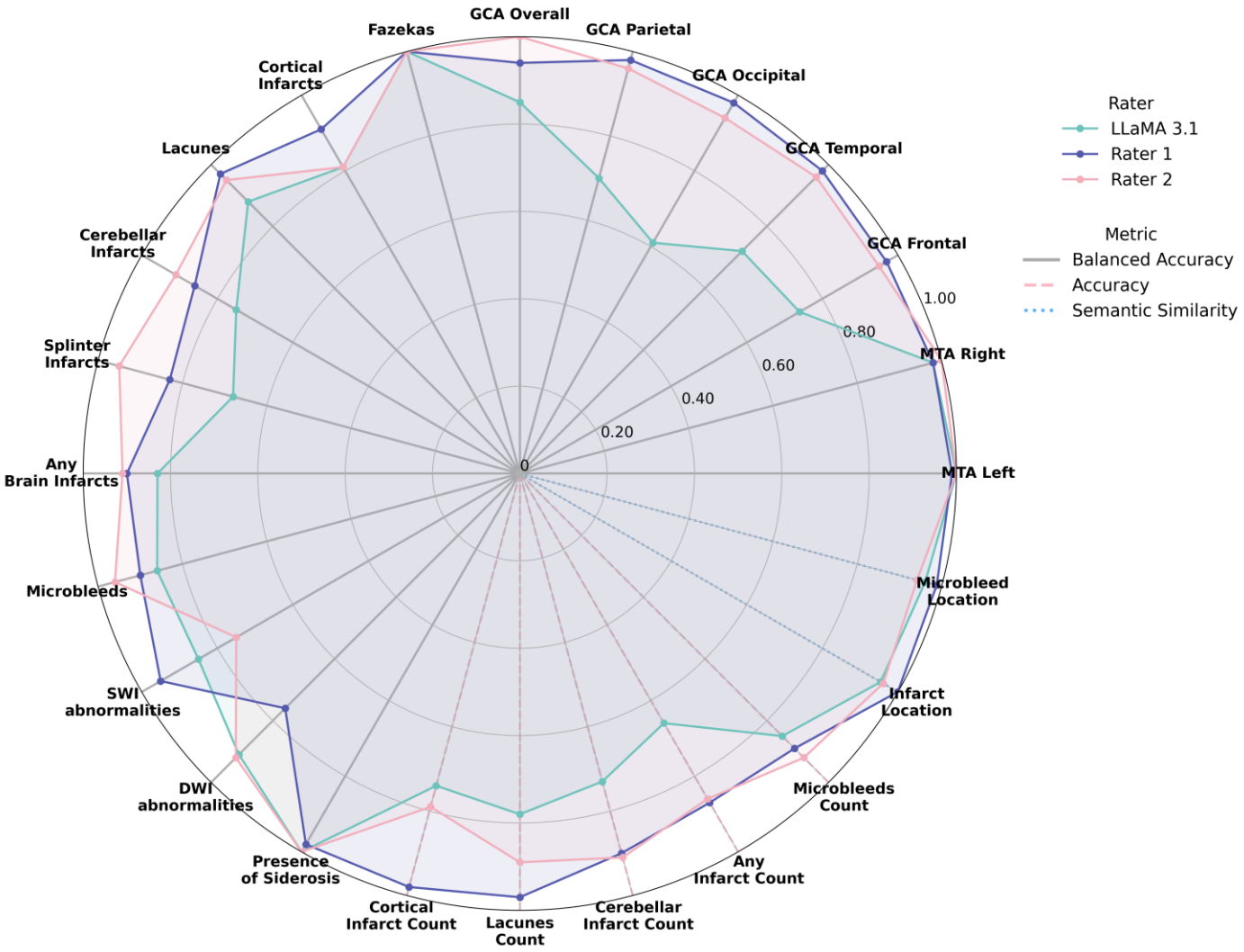

Figure S6: Radar plot of extraction performance per variable achieved by LLaMA 3.1 and each of two human raters as evaluated using the consensus annotations as reference standard. Performance is evaluated on a subset of 100 reports.

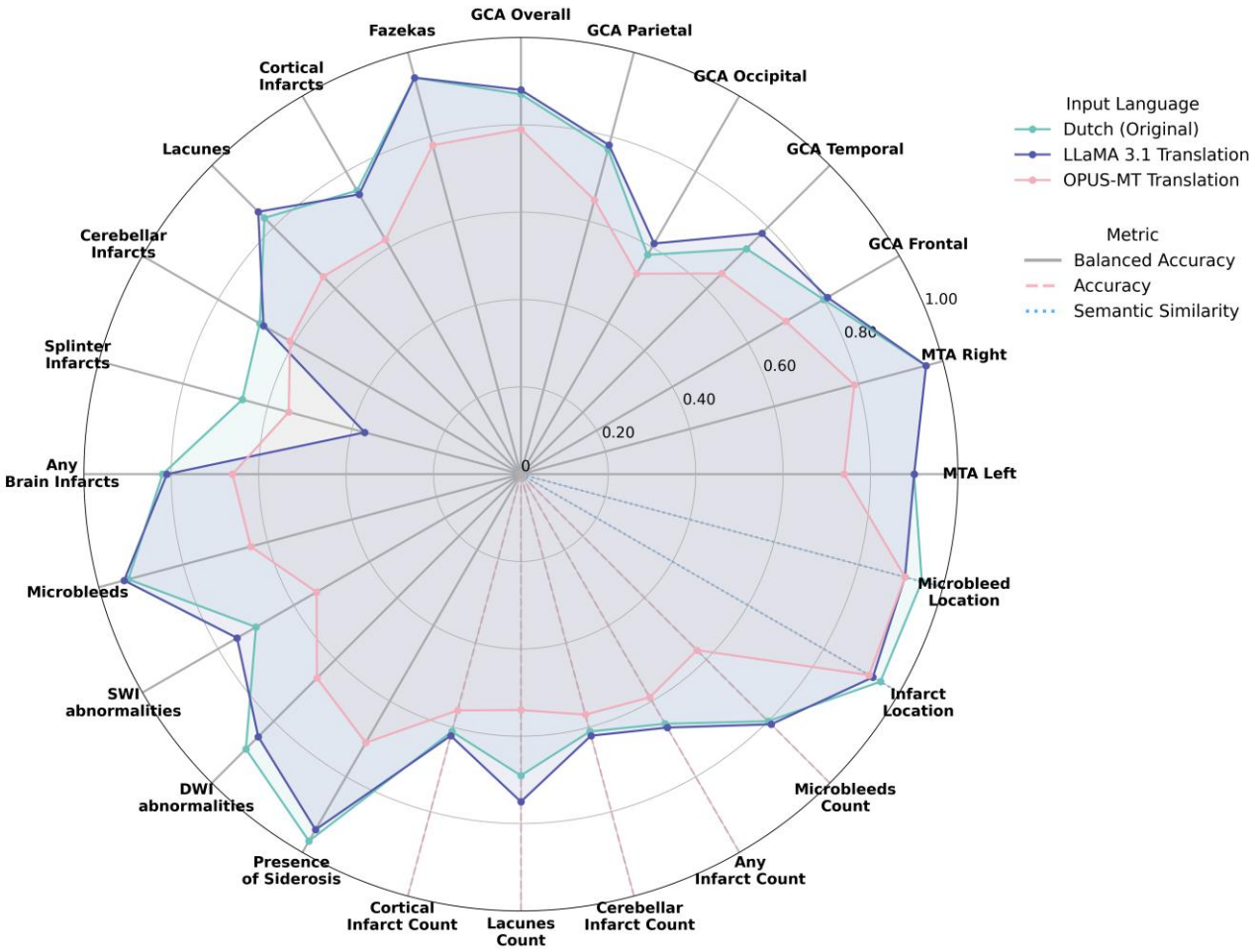

Figure S7: Radar plot of extraction performance per variable achieved using the original Dutch reports, the translation obtained using LLaMA 3.1, and the translation obtained using OPUS-MT as input. Performance is evaluated using the manual annotations as reference standard. Performance is averaged over 10 random splits of 50% of the 947 reports.

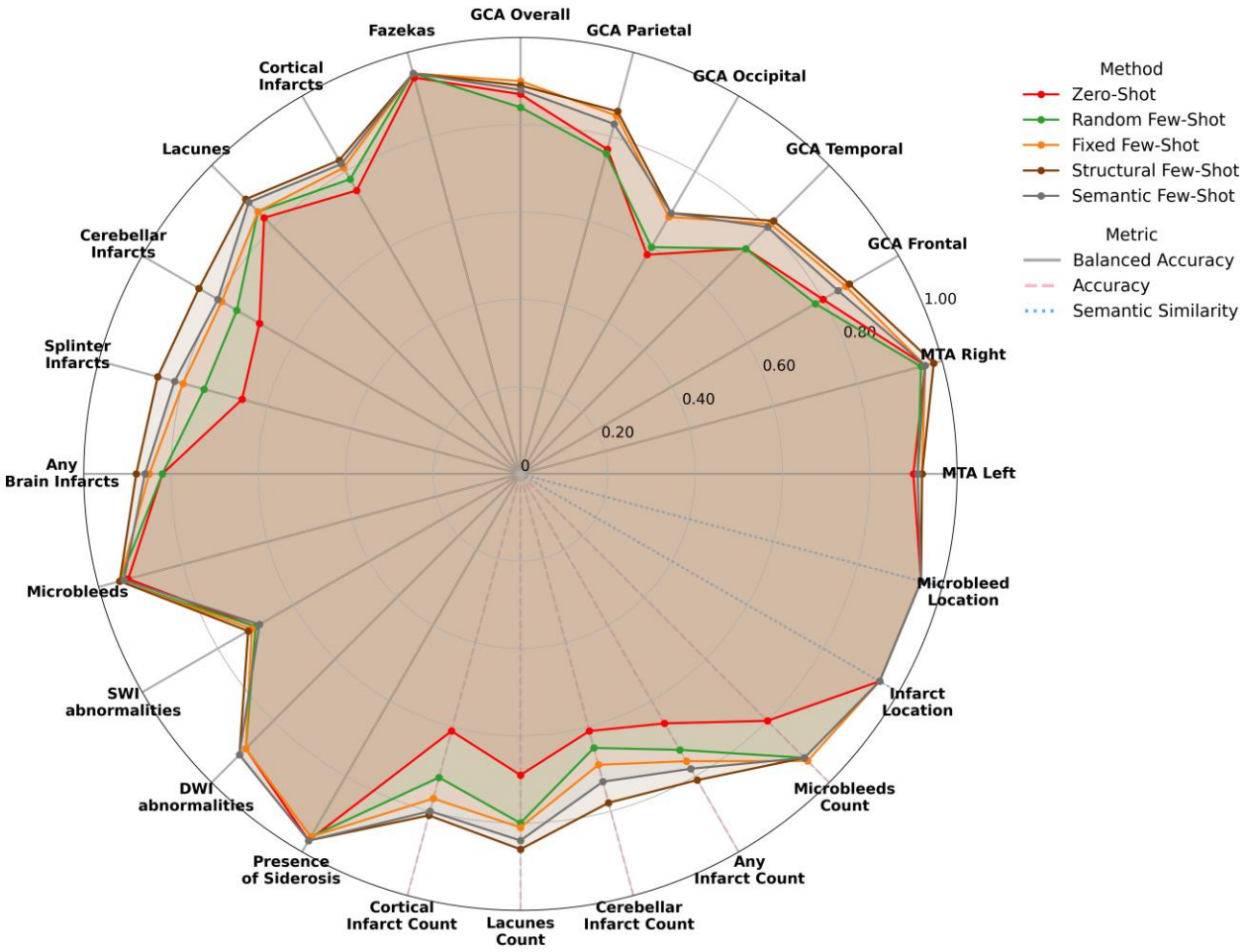

Figure S8: Radar plot of extraction performance per variable achieved using zero-shot prompting and few-shot prompting with random selection, fixed selection, structural similarity-based selection, and semantic semilarity based selection. Performance is evaluated using the manual annotations as reference standard. Performance is averaged over 10 random splits of 50% of the 947 reports.

K. Comparison between structural and semantic similarity-based selection

To compare the behavior of the two similarity-based selection methods, we examined the overlap in nearest neighbours selected for each test report. For each of the ten random data splits, we identified the three closest training reports using the structural-similarity measure and the semantic-similarity measure. For every test report, we then quantified how many of the three retrieved neighbours were shared by both methods. This overlap score provides insight into the extent to which structural and semantic similarity identify different few-shot examples. The results are shown in Fig. S7.

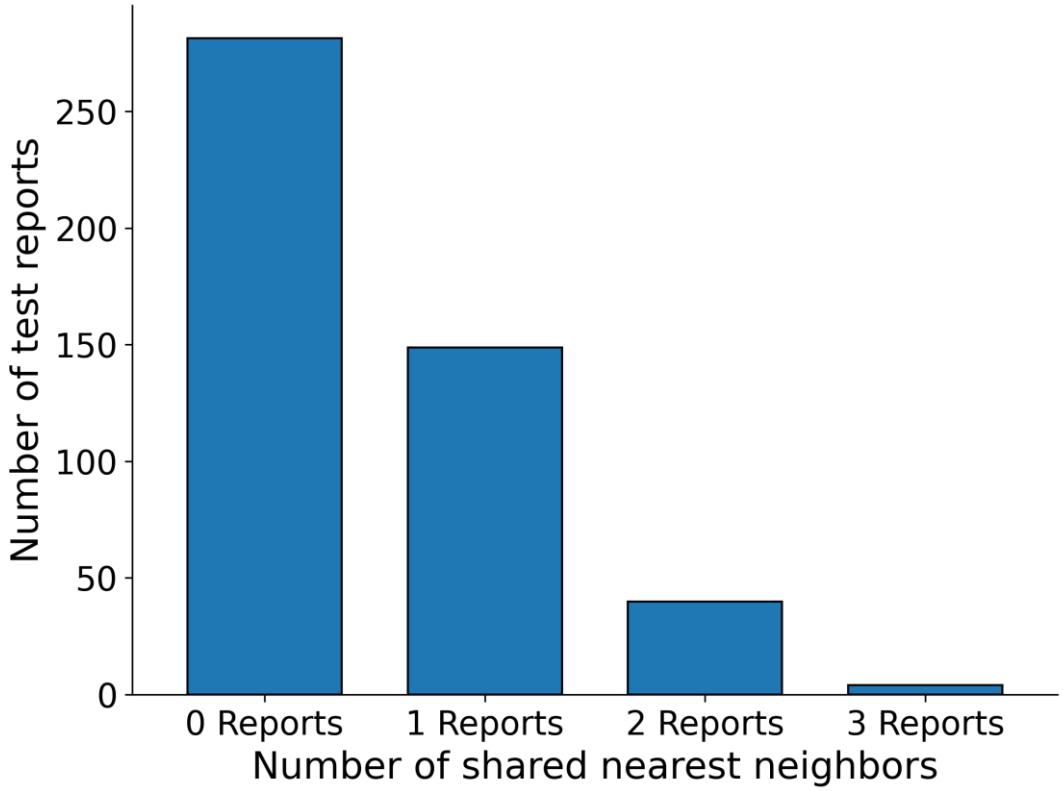

Figure S9: Overlap in nearest-neighbor selection between structural and semantic similarity-based selection. The numbers are averaged over the 10 random splits.

The overlap between structural and semantic similarity is limited. Most test reports share no nearest neighbors across the two methods, indicating that the criteria identify largely different examples in the training set. A substantial minority of reports share one neighbor, while cases with two or three shared neighbors are rare. This pattern suggests that structural and semantic similarity capture distinct aspects of report similarity and therefore retrieve complementary examples rather than converging on the same neighborhood structure.

L. Sensitivity Analysis: Treatment of the "Missing" Label in Categorical Variables

For categorical variables, the "missing" label (indicating that a finding was not reported) was treated as a distinct class in the primary evaluation, on the grounds that absence of mention is semantically different from an explicitly reported negative finding. To empirically assess the robustness of the reported performance to this choice, we recomputed balanced accuracy under two alternative evaluation schemes:

1. **Excluding missing:** reports with a "missing" reference label were removed, and balanced accuracy was computed on the remaining samples.
2. **Merging missing with negative:** labels coded as "missing" were recoded to the negative finding class ("absent" or “0” depending on the variable, see table S1) in both the reference and predicted labels, and balanced accuracy was computed on the full dataset.

Both schemes were applied for the structured similarity-based few shot prompting method per fold of the 10-fold Monte Carlo cross-validation and averaged across folds, consistent with the primary evaluation protocol. For each variable, we report the mean and standard deviation of balanced accuracy under each scheme, along with the mean per-fold difference (Δ) relative to the original scheme (Table S12).

Table S12: Balanced accuracy (BA) of categorical variables under alternative treatments of the 'missing' label. BA is reported under three evaluation schemes: the original scheme (missing treated as a separate class), excluding reports with a missing reference label, and merging missing into the negative class. Values are mean ± SD across 10 MCCV folds. Δ denotes the mean difference relative to the original scheme.

Variable	Missing (%)	BA Original	BA Excluding Missing	Δ vs. Original (Excl.)	BA Merging Missing → Negative	Δ vs. Original (Merge)
Atrophy Measures						
MTA left	29	0.92 ± 0.06	0.91 ± 0.06	-0.01	0.91 ± 0.06	0.00
MTA right	29	0.98 ± 0.01	0.98 ± 0.01	0.00	0.98 ± 0.01	0.00
GCA overall	25	0.89 ± 0.02	0.88 ± 0.02	-0.01	0.88 ± 0.02	-0.01
GCA frontal	29	0.87 ± 0.02	0.86 ± 0.03	-0.01	0.87 ± 0.02	0.00
GCA temporal	32	0.82 ± 0.02	0.81 ± 0.02	-0.01	0.82 ± 0.02	0.00
GCA occipital	38	0.69 ± 0.04	0.66 ± 0.05	-0.03	0.67 ± 0.05	-0.03
GCA parietal	26	0.86 ± 0.03	0.85 ± 0.03	-0.01	0.86 ± 0.03	0.00
Vascular Markers						
Fazekas score	28	0.95 ± 0.01	0.95 ± 0.01	0.00	0.99 ± 0.01	+0.04
Any brain infarct	43	0.88 ± 0.02	0.92 ± 0.01	+0.04	0.97 ± 0.01	+0.09
Cortical infarcts	47	0.83 ± 0.02	0.85 ± 0.04	+0.01	0.92 ± 0.04	+0.09
Lacunes	40	0.89 ± 0.01	0.91 ± 0.02	+0.02	0.96 ± 0.01	+0.06
Cerebellar infarcts	53	0.85 ± 0.02	0.90 ± 0.02	+0.05	0.99 ± 0.01	+0.14
Splinter infarcts	55	0.86 ± 0.02	0.90 ± 0.01	+0.04	1.00 ± 0.00	+0.14
Microbleeds	39	0.94 ± 0.01	0.96 ± 0.01	+0.02	0.99 ± 0.01	+0.05
SWI abnormalities	90	0.72 ± 0.04	0.59 ± 0.05	-0.13	0.75 ± 0.04	+0.03
DWI abnormalities	41	0.91 ± 0.03	0.89 ± 0.05	-0.02	0.90 ± 0.05	0.00

Superficial siderosis was excluded since it has no “missing” class.

For variables with a low to moderate prevalence of missing labels, including all atrophy measures (MTA, GCA), Fazekas score, and DWI abnormalities, balanced accuracy under both sensitivity schemes remained within 3% of the primary evaluation, indicating that the reported performance is robust to the handling of missing labels for these variables. For variables with a higher prevalence of missing labels (notably the presence of infarcts with their different subtypes and the presence of microbleeds) merging "missing" with "absent" produced substantially higher balanced accuracy, with increases of up to 14% (cerebellar and splinter infarcts). This confirms that treating "missing" as a separate class represents the more conservative performance estimate, and that merging the two would yield optimistic values by collapsing a genuine model error mode (confusing "missing" with "negative", as characterized in the section "Analysis of common mistakes") into correct predictions by construction. The exclusion scheme produced smaller increases for the same variables, reflecting that removing missing-labeled samples also eliminates a portion of the harder boundary cases. One variable, SWI abnormalities, showed a 13% decrease under the exclude scheme. This reflects the very high prevalence of missing labels for this variable (90%), which leaves a small and heavily imbalanced residual sample after exclusion and produces an unstable estimate; the merge scheme, which retains the full sample, was close to the original (+3%).