

Playing the Rebus Game: DR.DENKER Measures Visual Lateral Reasoning in Language Models

Michiel Pronk, Rik van Noord, Malvina Nissim
CLCG & GroNLP, University of Groningen, The Netherlands
michiel.t.pronk@gmail.com, r.i.k.van.noord@rug.nl, m.nissim@rug.nl

Abstract—Rebus puzzles seem deceptively simple: combine a small set of images, symbols, and letters, and human solvers can usually infer the hidden word with moderate effort. Yet, this kind of lateral inference poses a substantial challenge for AI systems. We introduce DR.DENKER, a benchmark of 1,196 hand-crafted Dutch rebus puzzles. Each rebus requires a solver to combine visual interpretation, abstract linguistic reasoning, and Dutch cultural knowledge to arrive at a single word or short phrase. We evaluate four state-of-the-art multimodal models and two flagship models across a structured set of prompting conditions. Without any contextual guidance, all models solve at most 2% of rebuses. Providing solution-constraining context (theme and answer length) raises the best small-model accuracy to 34.5% and flagship performance to 65.5%. Error analysis reveals that the primary bottleneck is the set of lateral thinking steps required for abstract, indirect, and phonetic interpretation: current models approximate lateral thinking under constrained conditions but lack open-ended creative reasoning.

Index Terms—rebus puzzles, visual reasoning, lateral thinking, multimodal language models, benchmark, Dutch NLP

I. INTRODUCTION

Artificial Intelligence systems have been extensively tested in their game-playing abilities, and have eventually mastered classic strategic games like chess and Go, consistently surpassing top human players [1], [2]. A forty-year-old Dutch newspaper puzzle based on rebuses, though, might be harder to crack (examples in Figures 1–4). While AI systems excel at *vertical thinking* [3], i.e., sequential, rule-governed reasoning which is core to games such as chess and Go that emphasize sequential search and constrained optimization, *lateral thinking* involves something different. Mastering lateral thinking requires relying more heavily on associative interpretation, ambiguity resolution, and creativity, and in most cases the ability to drop the obvious, literal interpretation in favor of an unconventional one. Rather than focusing on an exhaustive search, a solver must first discover how the puzzle should even be interpreted.

Thanks to their combination of visual and language clues, *rebus puzzles* offer a unique, principled vehicle for assessing visual lateral thinking in AI systems, and more specifically in Large Language Models (LLMs). In the kind of rebuses we consider, the hidden solution for each rebus is unique and constrained by a given number of letters and a pre-specified general theme, and must be reached through the combination of visual interpretation, linguistic creativity, and cultural



Fig. 1. Rebus from the 2018 puzzle, with *Time* as the theme. It is solved by recognizing that HAVO is a type of *Middelbaar Onderwijs* or *MO* (secondary education in Dutch), the short person is a *ventje* and the *vis* (fish) followed by the M should be read as *VisM* (*V is M*), i.e. $V=M$, hence *ventje* $\rightarrow$ *mentje*. The combination gives you the answer *momentje* (short moment).

knowledge. This combination of precision and creativity makes these rebuses an exciting test-bed for game-playing AI systems: precisely because success requires creative reinterpretation rather than pattern matching, rebuses can expose a gap in the models' reasoning abilities that more structured game benchmarks do not.

We probe state-of-the-art multimodal models on DR.DENKER, a set of 1,196 rebuses from the long-running Dutch puzzle series *Dr. Denker*, published annually in the regional newspaper *Dagblad van het Noorden*.¹ Each rebus demands Dutch linguistic knowledge, cultural familiarity, and the ability to apply abstract visual transformations and letter-modification rules. Figure 1 shows that solving such a rebus requires at minimum: (i) correct visual identification, (ii) generation of Dutch synonyms, (iii) application of positional or letter-modification rules, and (iv) verification against theme and solution length. In our experiments, without guidance, models solve fewer than 2% of rebuses, defaulting to literal visual interpretation. Providing solution-constraining context improves accuracy to 34% for the best small models and 65.5% for flagship models, but substantial failures remain in the core reasoning steps the task requires.

Our main contributions are: (i) DR.DENKER, a culturally grounded rebus puzzle benchmark for game-playing AI; (ii) a structured evaluation of how different types of game context affect AI solver performance; and (iii) an error analysis that identifies the lateral thinking steps where tested models fail.

¹*Dr. Denker* translates as “Dr. Thinker”. DR.DENKER benchmark: Dataset Regarding Dutch Evaluation for Non-linear, Kreative & Envisioned Reasoning.

II. RELATED WORK

A. Vertical vs. lateral thinking evaluation

Existing LLM benchmarks predominantly evaluate forms of vertical reasoning, emphasizing sequential inference toward a single correct solution. Representative examples include SWE-Bench for code repair [4], ZebraLogic for constraint satisfaction [5], and PlanBench for sequential planning [6]. In each case, performance depends on structured decomposition and convergent reasoning. Benchmarks targeting lateral reasoning remain comparatively limited. BrainTeaser evaluates lateral puzzle solving in a multiple-choice format and finds that larger models rely primarily on commonsense associations rather than genuinely creative inference [7]. LatEval extends this paradigm to interactive yes/no puzzles requiring exploratory questioning strategies [8], while MacGyver evaluates unconventional tool use in open-ended physical problem-solving scenarios [9]. In game-solving contexts, Todd et al. used the *New York Times* Connections game to assess abstract semantic grouping in LLMs [10]. However, these benchmarks generally do not involve open-ended visual interpretation and instead constrain reasoning to textual or predefined representational spaces.

B. Rebus benchmarks

Some previous work exists on testing LLMs’ abilities on rebuses, though in languages other than Dutch, and with puzzles of slightly different sorts than the ones we are dealing with in our experiments. Sarti et al. tested LLMs on Italian rebuses, but in this case, the visual interpretation step is not included since the model is provided with *verbalized* images, i.e., with all relevant portions of the picture represented by textual descriptions [11]. Even with this kind of help, the tested models struggled substantially with the task. REBUS by Gritsevskiy et al. [12] and COLUMBUS by Kraaijveld et al. [13] both define constrained puzzle environments with explicitly structured representations, where REBUS operates in an English-only setting with rule-based visual-verbal transformations, and COLUMBUS uses synthetic icon-based items paired with multiple-choice response formats that restrict outputs to predefined answer sets. In contrast, Lee et al. [14] introduce a collection of human-authored English puzzles with open-ended responses and annotations over cognitive skill categories, enabling less constrained interpretation. Reported results indicate that current reasoning models achieve only partial success, particularly on items requiring abstraction beyond surface-level cues or shifts in representational form.

DR.DENKER contributes to this research line by integrating features absent from prior work: open-ended responses, highly diverse visual data, and a Dutch-language and cultural context.

III. THE DR.DENKER BENCHMARK

Dr. Denker is a yearly rebus puzzle published in the *Dagblad van het Noorden*. Each annual edition contains approximately 40 rebuses, all sharing a single theme (e.g. *Time*, *Farm life*, *Furniture*). Figure 2 shows the complete puzzle from 2025, where the theme is *Games & Boardgames*. Each rebus encodes one Dutch word (for which the total number of letters is

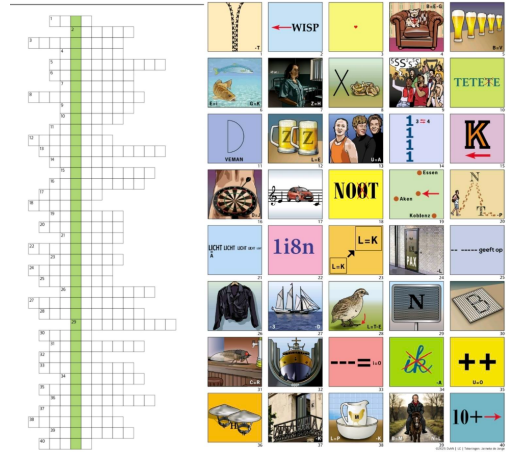


Fig. 2. Complete 2025 Dr. Denker puzzle. The theme is *Games & Boardgames*.

known), through a combination of imagery, embedded text, and letter-modification rules. Each solution word eventually contributes to yield a final saying (in green in the grid) which is conceptually associated with the general theme of the puzzle. Each full puzzle is also accompanied by a fun and engaging broad description of the theme of approximately 500 words.

We digitized the official collection *Helemaal Dr. Denker 1980–2020* [15] using a Python image-processing pipeline (grayscale conversion, adaptive thresholding, contour detection) to extract individual rebuses as PNG images. Puzzle descriptions (“descriptive text” below) were extracted with GPT-5 and manually corrected; answers were transcribed by hand. Puzzles before 1994 were excluded as they are black-and-white with a visually inconsistent style, leaving 1,196 valid rebuses across 30 puzzles (1994–2020; a few years had multiple puzzles).

Each dataset instance carries: the rebus image, solution, release year, puzzle theme, descriptive text, solution length, and a unique identifier at both rebus and puzzle level. The dataset is split at puzzle level to avoid thematic leakage: 20 puzzles (796 rebuses) for training/prompt development, 5 for development (200 rebuses), and 5 held-out for testing (200 rebuses). The 2025 edition (theme *Games & Boardgames*, 40 rebuses, Figure 2) was collected separately and used exclusively as an unseen generalization test (see Section V-D), as the models potentially saw the pre-2025 puzzles during their training. Approximately 63% of rebuses contain visual imagery; the remaining 37% are text-only and rely entirely on abstraction and wordplay. The dataset and code are publicly available.²

IV. EXPERIMENTAL SETUP

A. Models

We evaluated four multimodal models spanning a range of capabilities and cost profiles:

- **GPT-5 mini** (gpt-5-mini-2025-08-07)
- **Gemini 2.5 Flash** (gemini-2.5-flash)

²DR.DENKER dataset: <https://huggingface.co/datasets/MichiPrnk/dr-denker-helemaal>; code: https://github.com/MichiPrnk/master_thesis.

- **Claude Sonnet 4** (claude-sonnet-4-20250514)
- **Qwen3-VL-30B-A3B**, open-weight and run locally.

As an upper bound, and to get an idea of what players might experience when using state-of-the-art commercial models, we also evaluated **GPT-5.2** (gpt-5.2-2025-12-11) and **Gemini 3 Pro** (gemini-3-pro-preview).

All closed models were queried in Dutch via their APIs. Each rebus was presented in isolation. Reasoning was toggled: for the no-reasoning baseline we used minimal/budget-zero settings; for the reasoning condition we used medium/budget settings (5,000 tokens for Claude Sonnet 4; dynamic for Gemini 2.5 Flash; medium for GPT-5 mini; the thinking variant for Qwen3-VL). The models delivered their answer in JSON format. To measure performance, we calculate *accuracy* as the percentage of exact-match correct solutions (lowercased, whitespace-stripped).

B. Experimental Conditions

We evaluated four conditions per model: 1. **Baseline**: rebus image only, no reasoning; 2. **Reasoning baseline**: rebus image only, reasoning enabled; 3. **Puzzle-context**: adds solution length, theme, year, and the puzzle’s descriptive text to the prompt; 4. **Best configuration (cfg)**: best combination of extra strategies. For the latter, we experimented with several other prompting options, all of which supplied the model with a different type of information or strategy: (i) **Few-shot examples**: supplying the model with specific puzzles + solutions to learn from; (ii) **Candidate generation**: the model must first generate possible solutions fitting the constraints (theme and number of characters) before solving; (iii) **Human-expert tips**: supplying the model with 15 crowdsourced solving tips³; (iv) **Decomposition**: the model must explicitly decompose the puzzle’s visual element before solving.

Through a number of ablation studies, we found that the **best configuration (cfg)** combines the puzzle-context, decomposition of images and a one-shot example with solution and explanation. Interestingly, the human-expert tips did not help the models, nor did candidate generation.⁴

V. RESULTS AND ANALYSIS

A. Main model comparison

Table I shows accuracy across the four conditions on development and test sets. Two clear patterns emerge. First, without reasoning all models fail catastrophically. They correctly identify visual elements but interpret them literally, producing semantically plausible but wrong answers. Second, adding reasoning yields modest but consistent gains; the context condition then produces the largest single jump.

GPT-5 mini consistently outperforms the other small models. The puzzle context alone raises its accuracy from 9.5% to 29.0%. An ablation of the context components showed that *solution length* accounts for the largest drop when removed

³Tips were gathered using an open form shared on a Facebook page of Dr. Denker fans, the editorial office of the *Dagblad van het Noorden* and the authors’ personal network. About 10 people responded.

⁴For the latter setting GPT-5 mini chose an answer from a candidate list in 98.5% of cases, while the correct answer was only present in 32.5% of them.

TABLE I
MODEL ACCURACY (%) ACROSS EVALUATION STAGES.

Model	Baseline		Reasoning		+ Context		Best cfg.	
	D	T	D	T	D	T	D	T
GPT-5 mini	0.5	1.0	9.5	9.0	29.0	27.5	34.0	34.5
Gemini 2.5 Flash	2.0	2.5	6.0	5.5	13.5	16.0	15.5	16.5
Claude Sonnet 4	0.5	0.5	3.0	2.5	9.5	11.0	8.5	9.0
Qwen3-VL	0.0	1.0	1.5	1.0	3.5	1.5	2.5	1.5

TABLE II
FLAGSHIP MODEL ACCURACY (%) ON TEST SET.

Model	Baseline	Reasoning	+ Context	Best cfg.
GPT-5.2 (medium)	7.5	8.5	33.0	34.0
GPT-5.2 (xhigh)	7.5	15.5	52.5	54.0
Gemini 3 Pro	4.0	18.0	65.5	61.5

(29.0% → 16.0%), followed by theme (29.0% → 23.0%), confirming that hard constraints on the answer space are the primary driver of improvement. Qwen3-VL lags far behind all closed models, indicating that open-weight models are not yet competitive on this task.

B. Flagship model performance

Table II presents results for GPT-5.2 and Gemini 3 Pro on the test set. Gemini 3 Pro with puzzle-context achieves **65.5%** accuracy, far above any smaller model. This demonstrates that sufficient reasoning capacity can unlock substantial lateral thinking. In fact, the best-performing prompt configuration found in our previous experiments actually *reduces* Gemini 3’s accuracy by 4 points, suggesting that flagship models already possess the knowledge encoded in decomposition instructions and one-shot examples, and that additional guidance can be counterproductive. The level of reasoning investment (medium vs. extended) appears to matter as much as model size: GPT-5.2 with extended reasoning (xhigh) substantially outperforms the same model at medium reasoning.

C. Rebus type and error analysis

We annotated the dev and test rebuses on whether they are *visual* (containing imagery) or *text-only*. This revealed a consistent pattern: all models solve visual rebuses more reliably than text-only ones (e.g. 41.9% vs. 21.1% for GPT-5 mini on the development set). This is likely due to text-only rebuses relying entirely on wordplay, phonetics, and positional letter manipulation, demanding higher levels of abstract lateral reasoning with no visual anchor. An oracle predictor picking from all model predictions achieves 43–46.5% accuracy on dev/test, substantially above any individual model. Of the rebuses unsolved by any model, 46.6% are text-only, substantially exceeding their 37% share of the benchmark, implying that not vision but abstraction is the binding constraint. Figures 3 and 4 contrast the easiest and hardest items. As mentioned, the most-solved rebuses have clear, unambiguous

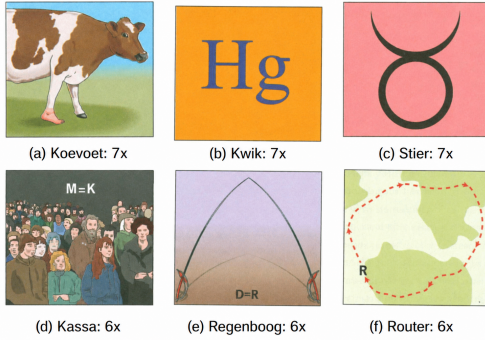


Fig. 3. Most-solved rebuses. Items have clear, unambiguous visual referents and short lateral steps. *Koevoet* is literally *cowfoot*, *kwik* is simply the Dutch name for the chemical element mercury (Hg), and *Stier* is literally bull or taurus. The three below are a bit harder but still relatively easy: a *massa* (mass) of people with $M=K$ leads to *kassa* (register), a *boog* (arch) of *degens* (epees) with $D=R$ leads to *regenboog* (rainbow) and a route with an R leads to *router*.

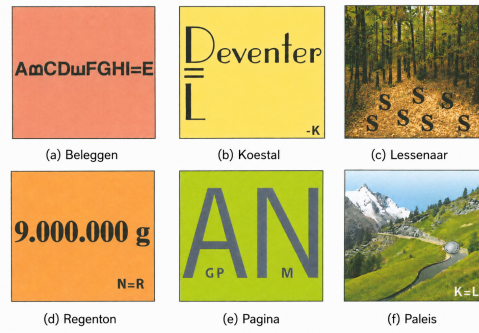


Fig. 4. Examples of rebuses unsolved by any model.

- The B and E are lying down (*liggen*) and I=E leads to *beleggen* (investing)
- Deventer is a Dutch city known as *koekstad* (cake city), after which $D=L$ and $-K$ lead you to *koestal* (cowshed), a relatively obscure Dutch word
- Multiple esses (*essen*) are *inside* a *laar* (open space in the woods), leading to *l-essen-aar* or *lessenaar* (lectern)
- 9 million grams is 9 tons (*nege ton*), $N=R$ leads to *regenton* (rain barrel)
- The m in N (MinN) simplifies to minus N. Then P after G (*P na G*) is in A, giving PnaGinA $-N$ and therefore *pagina* (page)
- The large rock (*kei*) is *inside* the (*berg*)*pas* (mountain pass), applying $K=L$ leads to Pa-lei-s and thus *paleis* (palace).

visual referents while the hardest ones are predominantly text-only and require multiple steps that are not straightforward, often relying on Dutch cultural knowledge.

D. Dr. Denker 2025

Since (some of) the models might have leveraged memorization from training data to arrive at the correct answer, we also evaluated the models, without browsing capabilities and in the days immediately following the release, on the puzzle which came out in December 2025 (*Games & Boardgames*). Under the best-performing configuration, accuracies closely match the held-out test set (GPT-5 mini: 37.5%; Gemini 2.5 Flash: 15.0%; Claude Sonnet 4: 12.5%; Qwen3-VL: 5.0%). GPT-5.2 (xhigh) reaches 67.5% and Gemini 3 Pro 62.5%, consistent with Table II. More than three quarters (77.5%) of all rebuses were solved by at least one model, suggesting the benchmark ceiling is approachable collectively but not individually. A small error

analysis on unsolved rebuses identified three failure modes: (i) *cross-rebus dependencies* (one rebus refers to another’s solution which is undoable without puzzle-level solving); (ii) *missing lateral steps* (correct visual identification but failure to make the abstract or phonetic leap); and (iii) *visual misrecognition* that derails subsequent reasoning. Mode (ii) accounts for 7 of 9 unsolved cases.

VI. CONCLUSION

Our evaluation on DR.DENKER, a benchmark of 1,196 Dutch rebus puzzles, demonstrates that (i) literal visual interpretation is insufficient for this task; (ii) solution-constraining context (answer length and theme) is by far the most effective intervention; (iii) flagship models with extended reasoning approach but do not reach consistent lateral thinking, solving up to 65.5% of rebuses; and (iv) the remaining errors are concentrated in exactly the steps that define lateral thinking, namely abstract reinterpretation, phonetic manipulation and indirect association rather than visual recognition.

Future work should include human baselines, multilingual extensions, and annotations of individual lateral thinking strategies to enable more fine-grained capability assessment.

ACKNOWLEDGMENTS

Computational resources were provided by the CLICK-NL PAGINA project (www.paginaproject.com, TKI2006-CI23018) and the CIT of the University of Groningen through the Hábrók HPC cluster.

REFERENCES

- D. Silver *et al.*, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- , “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play,” *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.
- E. de Bono, *The Mechanism of Mind*. Simon and Schuster, 1969.
- C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan, “SWE-bench: Can language models resolve real-world GitHub issues?” in *ICLR*, 2024.
- B. Y. Lin *et al.*, “ZebraLogic: On the scaling limits of LLMs for logical reasoning,” in *Proc AAAI*, 2025.
- K. Valmeekam, K. Stechly, and S. Kambhampati, “LLMs still can’t plan; can LRMs? a preliminary evaluation of OpenAI’s o1 on PlanBench,” in *NeurIPS 2024 Workshop on Open-World Agents*, 2024.
- Y. Jiang, F. Ilievski, K. Ma, and Z. Sourati, “BrainTeaser: Lateral thinking puzzles for large language models,” in *Proceedings of EMNLP*, 2023.
- S. Huang *et al.*, “LatEval: An interactive LLMs evaluation benchmark with incomplete information from lateral thinking puzzles,” in *Proc of LREC-COLING 2024*, 2024, pp. 10 186–10 197.
- Y. Tian *et al.*, “MacGyver: Are large language models creative problem solvers?” in *Proc of NAACL*, 2024, pp. 5303–5324.
- G. Todd, T. Merino, S. Earle, and J. Togelius, “Missed connections: Lateral thinking puzzles for large language models,” in *2024 IEEE Conference on Games (CoG)*, 2024, pp. 1–8.
- G. Sarti, T. Caselli, A. Bisazza, and M. Nissim, “EurekaRebus: Verbalized rebus solving with LLMs – a CALAMITA challenge,” in *Proceedings of CLiC-it 2024*, 2024, pp. 1202–1208.
- A. Gritsevskiy *et al.*, “REBUS: A robust evaluation benchmark of understanding symbols,” in *arXiv preprint arXiv:2401.05765*, 2024.
- K. Kraaijveld, Y. Jiang, K. Ma, and F. Ilievski, “COLUMBUS: Evaluating cognitive lateral understanding through multiple-choice rebuses,” in *Proceedings AAAI*, vol. 39, no. 4, 2025, pp. 4410–4418.
- H. Lee *et al.*, “Puzzled by puzzles: When vision-language models can’t take a hint,” in *Proc of EMNLP*. ACL, 2025, pp. 21 724–21 737.
- Dr. Denker and J. de Jong, *Helemaal Dr. Denker 1980–2020: alle kerst-, paas- en extra puzzels*. Dagblad van het Noorden, 2020.