Eur. Phys. J. Spec. Top. (2026) 235:8691–8702
https://doi.org/10.1140/epjs/s11734-026-02534-1

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Can an LLM reproduce an expert knowledge base in neuroeducation?

Tatyana Bukina[1,a], Marina Khramova[2,b], Nikita Smirnov[1,c], Semyon Kurkin[1,d], and Alexander Hramov[1,e]

[1] Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia
[2] Higher School of Cybertechnology, Mathematics and Statistics, Plekhanov Russian University of Economics, Moscow, Russia

Received 28 May 2026 / Accepted 7 July 2026 / Published online 16 July 2026
© The Author(s), under exclusive licence to EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract Knowledge-based educational recommenders depend on a manually curated mapping from interventions to the cognitive functions they develop—a costly bottleneck that limits coverage and updating. We ask whether a large language model (LLM) can reproduce such an expert knowledge base (KB) directly from the primary literature. Because every KB entry is already traceable to its source paper, the KB itself serves as ground truth: we seed a corpus of 102 articles with the 32 source papers of the KB and 70 topically matched distractors, run a single-pass LLM extraction (Gemini 2.5 Flash over native PDFs), and measure how much of the KB is recovered. The LLM re-identifies the correct activity in every seed paper (activity recall 32/32), but reproduces the expert's four cognitive-function labels unevenly (micro-function recall 64%): faithfully for working memory, yet below a random-tagging baseline for mental arithmetic, and over-applied for combined functions. From the distractors, the model also surfaces school-relevant activities absent from the KB, demonstrating a viable path to semi-automated KB expansion. We conclude that LLM extraction is reliable for discovering interventions, while function labelling remains uneven and best kept under expert control.

1 Introduction

Translating a learner's cognitive profile into a concrete development plan requires a structured link between cognitive functions and the activities that develop them. Knowledge-based recommenders mediate between abstract learning goals and concrete resources through such an intermediate, structured representation: Beutling and Spahic-Bogdanovic [1] link career goals to courses via competency ontologies extracted by LLMs, and Gao et al. [2] build tutoring systems on knowledge graphs. Activity-recommendation decision-support expert system (ADES) [3, 4], a neuro-symbolic recommender that suggests extracurricular activities to schoolchildren from their EEG-derived cognitive profiles, extends the open-loop neuroadaptive pipeline of Grubov et al. [5] and relies on a hand-curated knowledge base (KB) that maps each activity to the cognitive functions it is believed to develop. This KB is the system's pedagogical backbone and its principal bottleneck. Every entry was assembled by experts who located an intervention study, judged which cognitive function it targets, and recorded the schedule, age range, and resources. Manual curation is slow, hard to keep current with a fast-growing literature, and difficult to audit.

A natural remedy is to extract this knowledge automatically. Large language models (LLMs) can read scientific articles and emit structured descriptions of interventions and outcomes: they are strong few-shot extractors of clinical information [6], recover structured records from complex materials science text [7], and are competitive on relation extraction [8]. When the target is a fixed schema, the task becomes schema-guided or zero-shot information

[a] e-mail: bukinatatyanav@gmail.com (corresponding author)
[b] e-mail: hramova.mv@rea.ru
[c] e-mail: nikita.smirnov.m@gmail.com
[d] e-mail: kurkinsa@gmail.com
[e] e-mail: hramovae@gmail.com

extraction [9], the regime our closed four-function list imposes. Treated as an annotator, an LLM can rival or exceed crowd workers on labelling tasks [10], and the curation cost of hand-built structures has motivated growing interest in populating them automatically [11]. But before trusting an LLM to expand a knowledge base, we must ask whether it can reproduce one that experts have already built. If the model cannot recover the activities and function mappings the experts derived from the very same papers, automatic expansion is unsafe.

The functions targeted by our neuroeducational system [3, 4] rest on well-established frameworks: working memory on Baddeley's model [12], visual search on Treisman's feature-integration theory [13], and cognitive flexibility within the executive-function framework of Miyake et al. [14]. These constructs are individually well grounded but not mutually orthogonal, and mapping a given intervention onto them is itself an expert judgement; this assignment step, rather than the underlying constructs, is where an LLM and the expert KB most often disagree when reading the same source papers.

We make this question measurable by reusing the experts' own labelling instead of collecting fresh annotations. The expert KB is already a labelled dataset: each (cognitive function, activity) entry carries a citation to the source paper from which the experts drew it. We exploit this to build a *seeded recall benchmark*, which requires only a single manual step: checking whether an extracted activity refers to the same intervention as the ground-truth one. We assemble a corpus that mixes the KB's source papers (*seeds*, with known correct extractions) with a larger set of topically similar *distractors*, run LLM extraction blind to this distinction, and measure (i) how many KB activities are recovered from their seed papers (recall), (ii) how faithfully the expert's cognitive-function labels are reproduced, and (iii) what additional, KB-absent activities the distractors surface as expansion candidates.

2 Methods

Figure 1 summarises the end-to-end design: the expert KB and a set of topically matched distractors are merged into a single corpus under shuffled identifiers, a single-pass LLM extraction is run blind over all PDFs, and the held-out seed/distractor key splits the output into a seed branch (activity recall and function agreement against the ground truth) and a distractor branch (age-filtered expansion candidates).

2.1 The expert knowledge base as ground truth

Our reference KB underlies ADES [3, 4]. It maps extracurricular activities to four cognitive functions assessed in the EEG battery: *visual search* (selective and sustained attention), *mental arithmetic*, *working memory*, and *combined functions* (cognitive flexibility/executive control). Each row pairs one cognitive function with one activity and cites the intervention study that justifies the pairing; the same activity may appear under several functions, and the same paper may support several functions [15]. Throughout, by *intervention*, we mean a trainable extracurricular activity delivered over time and evaluated against a measured cognitive outcome.

We parse the curated reference table into a paper-level ground truth. Each source paper is keyed by DOI (or, for a few venues, by reference identifier) and associated with the set of (function, activity) pairs it supports, together with the age range reported. The ground truth comprises 32 source papers, 25 distinct activities, and 76 (paper, function) pairs across the four functions.

2.2 Corpus construction

The benchmark corpus mixes seeds and distractors (Table 1). The 32 seed PDFs are the source papers above. The 70 distractors are topically matched controls harvested from Semantic Scholar across 12 queries spanning the KB's domains (executive-function and working-memory training, physical activity and cognition, music training, martial arts, mindfulness, video games, etc.; the full query list is in the Appendix 2. We kept only open-access primary studies (excluding reviews and meta-analyses) with an abstract in a relevant field, removed any paper matching a seed DOI or title, and balanced the pool across query domains by round-robin selection. The resulting 102-PDF corpus is copied under opaque, shuffled identifiers so that extraction is blind to seed/distractor status; the seed/distractor key is held out as the answer key.

2.3 Extraction

Each PDF is sent natively (not as pre-extracted text) to Gemini 2.5 Flash via the OpenRouter API at temperature 0, with a structured-JSON prompt. For every trainable extracurricular activity studied as a cognitive intervention, the model returns: a short activity name, the cognitive outcomes the study measured in its own words (open vocabulary), a mapping of the activity onto the closed list of four KB functions (each accompanied by its definition in the prompt), the participant age range, schedule, resource requirements, a description, and a

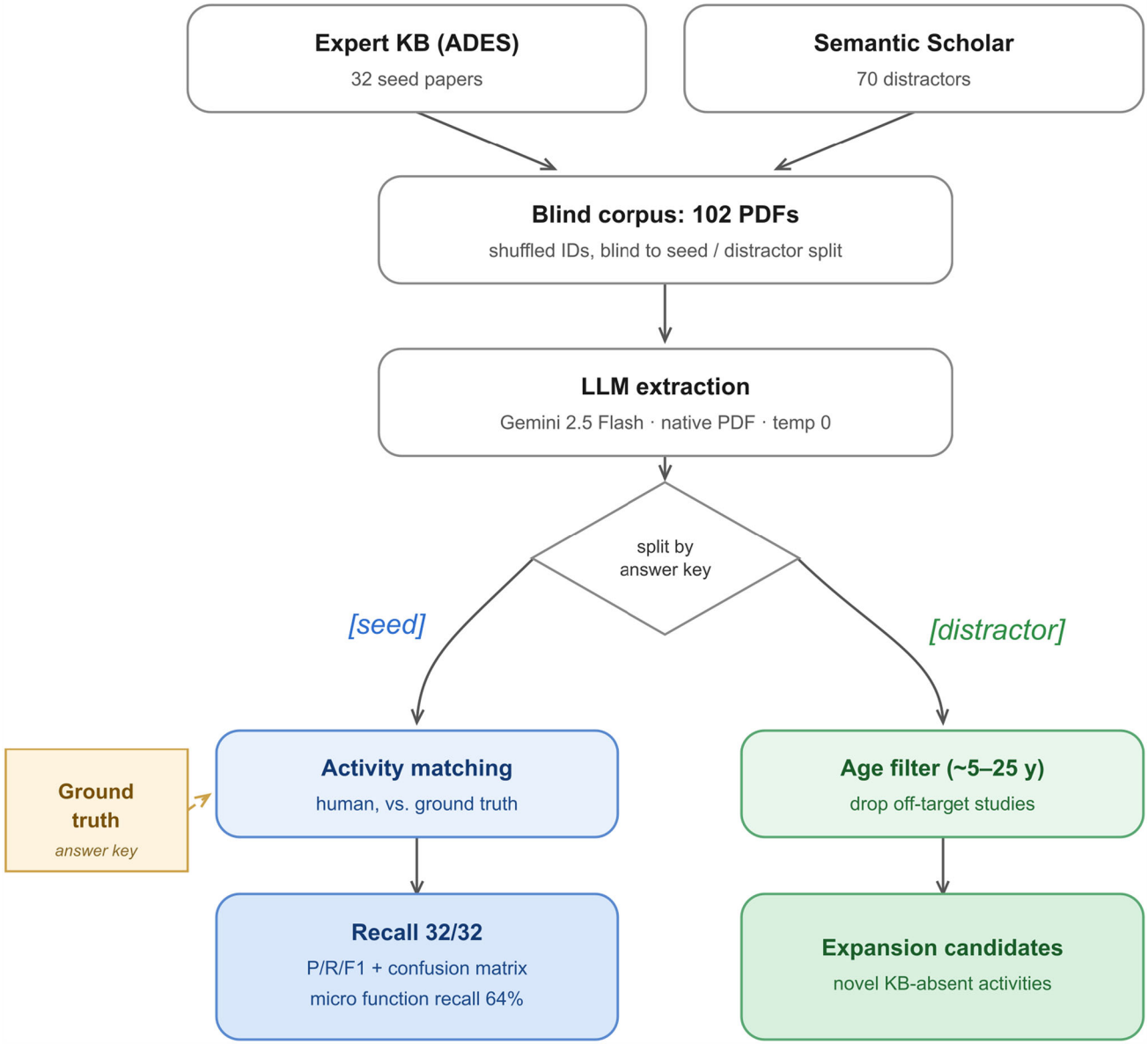


Fig. 1 End-to-end benchmark pipeline. The expert KB (32 seed papers) and 70 topically matched Semantic Scholar distractors are merged into a blind 102-PDF corpus under shuffled identifiers. A single-pass LLM extraction (Gemini 2.5 Flash, native PDF, temperature 0) is run over the whole corpus; the seed/distractor key, held out as ground truth, then splits the output into a seed branch (human activity matching, recall, and the function confusion matrix) and a distractor branch (age filtering to the target band, yielding KB-absent expansion candidates)

Table 1 Benchmark corpus composition

Subset	Papers	Role	Ground truth
Seed	32	Known KB sources	76 (paper, function) pairs, 25 activities
Distractor	70	Topically matched controls	None (expansion candidates)
Total	102	Blind extraction input	

supporting verbatim quote. Papers that are not intervention studies are expected to return an empty activity list. The full prompt is given in the Appendix 1.

2.4 Matching and metrics

Activity matching For each seed paper, the ground-truth activity is known, so the only question is whether any extracted activity denotes the same intervention (e.g. "Football" vs. "Football training"). We adjudicate activity identity by human review, recording for each paper whether the ground-truth activity was extracted and which candidate it corresponds to.

Function metrics Once the activity is matched, function agreement is a set comparison between the matched activity's predicted functions and the expert's functions for that paper. We report *activity recall* (fraction of seed papers whose ground-truth activity was extracted), *function recall* (fraction of the 76 ground-truth (paper, function) pairs reproduced), and a confusion matrix $M[g][p]$ counting, over matched papers, the co-occurrence of an expert function g with a predicted function p.

Novel-activity analysis For distractors, we resolve each extracted activity's participant age (back-filling a focused age-only extraction pass when the first pass left it null) and retain activities overlapping the KB's target band ($\sim$ 5–25 years), discarding preschool, adult, elderly, and non-human studies. The upper bound of $\sim$ 25 years reflects the KB's own coverage: intervention studies targeting schoolchildren specifically are scarce, and ADES was planned in the future to cover schoolchildren and university students, so the KB admits young-adult evidence; this matches the extraction prompt, which targets "children, schoolchildren, or students", with final age-appropriateness left to the expert-validation step that precedes any KB entry. Retained activities not present in the KB are expansion candidates.

3 Results

3.1 Activity recall

The model flagged all 32 seed papers as intervention studies and, by human adjudication, extracted the correct activity in every case (activity recall = 32/32 = 100%). Paraphrase was the norm—"Abacus" → "Abacus course", "Karate" → "Karate", "Horse training" → "Activities with horses"—and several papers yielded finer-grained arms (e.g. intervention vs. control), but the target intervention was always present.

3.2 Function recall and confusion

Reproduction of the expert's cognitive-function labels is far less uniform (micro-recall = 49/76 = 64%; bootstrap 95% CI 55–73% over the 32 papers). Recall alone is hard to read, because the model assigns on average 1.94 of the four labels per paper, so a random tagger of that cardinality would already reach $\sim$48% recall and a "predict-all-four" tagger 100%. We therefore report precision, recall, and $F1$ per function (Table 2) alongside the confusion matrix (Table 3).

Precision reshapes the naive recall ranking. *Working memory* is both precise and complete ($P = R = 95\%$): a genuine point of agreement. *Combined functions* reaches 80% recall but only 59% precision (11 false positives)—the

Table 2 Per-function precision, recall, and $F1$ over the 32 seed papers (closed four-label set); n = papers the expert assigned to that function

Expert function	P (%)	R (%)	$F1$	n
Working memory	95	95	**0.95**	21
Visual search	100	56	0.71	18
Combined functions	59	80	0.68	20
Mental arithmetic	75	18	**0.29**	17
Macro-avg	82	62	0.66	–
Micro-avg	79	64	0.71	–

A random tagger matching the model's mean output cardinality (1.94 labels/paper) would score $\sim$48% recall; a "predict-all-four" tagger scores 100% recall at precision equal to each label's prevalence (53–66%)
The best and worst $F1$ scores in four cognitive functions are highlighted in bold

Table 3 Label co-occurrence matrix M[expert][predicted] over matched seed papers

Expert\pred.	VS	MA	WM	CF	n
Visual search	10	3	11	15	18
Mental arithmetic	1	3	13	14	17
Working memory	4	4	20	16	21
Combined functions	5	4	18	16	20

Rows are the expert's function label; columns are the function the LLM assigned. Because the model assigns 1.94 labels per paper on average, this is not a single-label confusion matrix: an off-diagonal cell means the LLM also applied that label, not that it substituted it, and rows need not sum to n. This also explains the dense WM and CF columns—on multi-label papers, WM is frequently in the gold set too, which is why WM precision stays at 95% despite being predicted widely.
VS visual search, *MA* mental arithmetic, *WM* working memory, *CF* combined functions. The last column is the number of seed papers the expert assigned to that function

model applies this executive-function label liberally, so its high recall is partly over-tagging rather than agreement: it is predicted for the majority of papers in every expert category (14–16 of each, Table 3). *Visual search* is the opposite: precise but incomplete ($P = 100\%$, $R = 56\%$)—when predicted it is always correct, but 8 of 18 papers are instead labelled combined functions or working memory. *Mental arithmetic* is recovered in only 3/17 papers (18%, below the $\sim 48\%$ a random tagger would reach): on these papers, the model also tags combined functions (14) and working memory (13), reading abacus and mental-calculation training as working-memory/executive-function interventions rather than assigning a distinct "mental arithmetic" label.

3.3 Distractor behaviour and expansion candidates

Of the 70 distractors, the model correctly returned no activity for 7 non-intervention papers and extracted 120 activities from the remainder. Age resolution (including a back-fill pass for papers whose first-pass age was null) removed 13 off-target extractions—4 adult/elderly, 5 preschool, and 4 flagged by population (Alzheimer's patients, an animal study, adult neurological patients)—leaving 107 school-relevant activities. The majority duplicate KB categories (physical activity, music, mindfulness, working-memory/cognitive training, video games), re-discovering the KB on unseen papers. Beyond these, the model surfaced activity types absent from the current KB—among them neurofeedback, virtual-reality and exercise game training, sport stacking, board games other than chess, sensorimotor-integration training, and gamified mathematics—each a candidate for expert-validated KB expansion.

4 Discussion

Extraction discovers activities reliably Perfect activity recall on the seeds, plus correct rejection of non-intervention distractors and re-discovery of KB categories on unseen papers, indicates that single-pass LLM extraction is dependable at the level of what intervention a paper studies. For the practical goal of growing the KB, this is the operative competence, and it requires no fine-tuning or text pre-processing.

There are two taxonomy weak points Two function labels are reproduced poorly. *Mental arithmetic* falls below a random-tagging baseline (recovered in only 3/17 papers): the model rarely treats mental calculation as a distinct outcome, instead reading abacus and calculation training as working-memory or executive-function interventions. *Combined functions* is recovered often (80%) but at low precision (59%), because the model applies this broad executive-function label too liberally. We cannot say from a single extraction run why this happens. Three explanations are plausible and not mutually exclusive: the source studies may genuinely report visuospatial or executive outcomes rather than "mental arithmetic"; the model may favour broad, high-frequency labels; or the expert functions may simply overlap, leaving several labels defensible for one activity. Separating these causes would require multiple extractors and expert relabelling. What the benchmark does establish is where the taxonomy and a competent automatic reader diverge: mental arithmetic (in recall) and combined functions (in precision) are the cells most in need of expert review.

Implications for semi-automated curation The asymmetry suggests a division of labour: let the LLM propose activities and their measured outcomes (high reliability), but keep the mapping onto a system-specific function taxonomy under expert control, especially for the cells where extraction and the taxonomy diverge (here mental arithmetic and combined functions). The confusion matrix can guide taxonomy redesign so that the symbolic function labels and the neural extraction converge.

Limitations The ground truth inherits any errors of the original curation, which we did not re-audit. Activity matching is human-adjudicated on 32 papers—tractable and transparent at this scale, but not automated. Crucially, function mapping is guided: the model receives the four target functions and their definitions in the prompt, so it performs constrained classification rather than open-vocabulary discovery; the resulting label distribution—including the over-application of combined functions—may be sensitive to how those definitions are worded, and a prompt-sensitivity analysis is left for future work. We evaluate a single extractor (Gemini 2.5 Flash); other models may differ, and a model comparison is left for future work. Finally, the distractor pool is open-access-biased, and the novel-activity list is a screening output that still requires expert validation and evidence grading before entering the KB.

5 Conclusion

We turned an existing expert knowledge base into a recall benchmark for scientific information extraction—reusing the experts' own labelling rather than collecting new annotations—by seeding the KB's source papers among topically matched distractors and measuring how much an LLM recovers. A single-pass extractor reproduced the correct activity in all 32 seed papers and re-discovered KB categories on unseen papers, while also surfacing school-relevant activities the KB lacks—supporting LLM-assisted KB expansion. Yet, function-label agreement was uneven and label-dependent: mental arithmetic fell below a random-tagging baseline and combined functions was reproduced only at low precision, marking both as the cells where automatic extraction and the taxonomy most diverge—a single run cannot separate the literature's framing from the model's labelling bias. The broader lesson is methodological: a curated KB that records its sources is itself a low-cost benchmark for the extraction step that might one day maintain it.

Author contributions

T. Bukina conceived the study, curated the expert knowledge base used as ground truth, performed the activity adjudication, and wrote the first draft. M. Khramova and S. Kurkin contributed to the methodology and validation. N. Smirnov implemented the extraction and evaluation pipeline and performed the data analysis. A. Hramov supervised the work and contributed to its conception. All the authors reviewed and approved the final manuscript.

Funding The article was prepared as part of research project No. FSSW-2026-0008 "Development of fundamental principles for constructing digital twins of infrastructure systems in the data economy" using funds from a subsidy from the federal budget within the framework of the state assignment of the Ministry of Education and Science of Russia.

Data availability The 102-PDF corpus is identified by DOI: a machine-readable list of the 32 seed and 70 distractor DOIs, together with the paper-level ground truth derived from the expert reference table and the human activity-adjudication verdicts, permits full reconstruction of the corpus (copyrighted PDFs are not redistributed). The full extraction prompt and the complete corpus listing (all 102 papers with DOIs) are provided in the appendix.

Materials availability Not applicable.

Code availability The extraction and evaluation code is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval Not applicable. This study analyses previously published scientific articles and an existing expert knowledge base; it does not involve any new experiments on human participants or animals and therefore did not require ethics approval or informed consent.

Consent for publication Not applicable.

Appendix 1: Extraction prompt

Each PDF was sent natively to Gemini 2.5 Flash (temperature 0, JSON response format). The prompt was run in Russian (the language of the knowledge base); we give a faithful English rendering here, and the verbatim original ships with the code.

System message. "You are an expert in cognitive science and the analysis of scientific articles. You extract, from studies, data on trainable activities and their effect on cognitive functions. You answer strictly in JSON."

User message. "Below is a scientific article. Determine whether it studies any TRAINABLE EXTRACURRICULAR ACTIVITY (a class, sport, game, course, or training programme) as an intervention to improve the cognitive functions of children, schoolchildren, or students. If so, extract EVERY such activity. For each return:

- `activity_name_ru`, `activity_name_en`—a short name of the activity;
- `measured_outcomes`—the cognitive functions/indicators the study actually measured or improved, in the article's own terms (open vocabulary, e.g. "inhibition", "visuospatial working memory", "selective attention");
- `target_functions`—a subset of the four functions below that the activity is meant to improve according to the article, using only these exact labels;
- `age_min`, `age_max`—participant age range (integer years; null if unspecified);
- `schedule`—frequency and duration of sessions;
- `requirements`—required resources/equipment, if any;
- `description`—1–2 sentences on the content of the activity;
- `evidence_quote`—a short verbatim quote linking the activity to the cognitive effect.

The four target functions are:

1. *Visual Search*—visual and selective attention: searching for target stimuli among distractors, speed and accuracy of visual search, as well as sustained attention, attention control, and resistance to distraction.
2. *Mental Arithmetic*—mental calculation and numerical operations.
3. *Working Memory*—holding and manipulating information in mind (digit span, n-back, Sternberg paradigm, visuospatial working memory).
4. *Combined Functions*—cognitive flexibility and executive functions in combination: switching, inhibition, planning, simultaneous operation of several processes.

If the article does NOT study a trainable activity (e.g. a review, methodological, or irrelevant paper), return an empty `activities` list. Return strictly the JSON: {`\is_intervention_study": true/false, \activities": [{...}, ...]`}."

Appendix 2: Distractor harvest queries

The 70 distractors were harvested from Semantic Scholar using the following 12 domain queries. Results were interleaved by query (round-robin) so that the downloaded pool spans all domains rather than the largest query buckets:

1. executive function training intervention children randomized
2. cognitive training children working memory intervention
3. physical activity cognition children executive function
4. music training cognitive development children
5. martial arts children attention self-regulation
6. sport training cognitive function children adolescents
7. school-based intervention attention children
8. aerobic exercise executive function adolescents
9. video game cognitive training attention
10. mindfulness children attention executive function
11. dance movement cognition children
12. academic achievement physical education cognition

Appendix 3: Corpus listing

Table 4 lists the 32 seed papers with their ground-truth activity and target functions; Table 5 lists the 70 distractors. Identifiers (`P###`) are the opaque, shuffled corpus IDs under which extraction was run blind to seed/distractor status.

Table 4 Seed papers (the 32 source articles of the expert KB) with their ground-truth activity and target functions (VS = visual search, MA = mental arithmetic, WM = working memory, CF = combined functions)

ID	DOI/source	Activity	Functions
P001	https://cyberleninka.ru/article/n/vliyanie-zanyatiy-sportivnym-orientirovaniem-na-umstvennuyu-rabotosposobnost-bakalavrov-fizicheskoy-kultury	Orienteering	VS, CF, WM
P005	10.1111/sms.12093	Intensive physical exercise	MA
P006	10.1007/s10826-018-1073-9	Mindfulness training	VS
P008	10.1111/desc.12866	Computer training	VS, CF, WM
P009	10.1016/j.learninstruc.2021.101442	Music (instrument)	CF, MA, WM
P010	https://escholarship.org/uc/item/55g8446v	Chess	VS, CF, MA, WM
P016	10.3389/fpsyg.2022.982704	Music (string instruments)	CF, MA, WM
P018	10.1371/journal.pone.0212482	Active-play breaks	MA
P019	10.3389/fpsyg.2017.02303	Music (instrument)	WM
P021	10.1080/09297049.2012.736483	Cognitive training	CF, MA, WM
P023	10.1123/pes.2015-0084	Physical education/sport	VS, CF, WM
P027	10.1002/brb3.3148	Equine activities	CF, MA, WM
P028	10.3389/fpsyg.2015.01627	Football	VS
P031	10.34835/issn.2308-1961.2021.7.p52-56	Breathing practices	VS, CF, WM
P032	10.3389/fnins.2020.00567	Music (string instruments)	VS, CF, MA, WM
P037	10.52563/2618-7604_2023_1_5	Table tennis	VS
P039	10.1016/j.cub.2013.01.044	Action video games	VS
P040	https://www.elibrary.ru/download/elibrary_39386284_35174844.pdf	Basketball	VS
P045	https://pmc.ncbi.nlm.nih.gov/articles/PMC4187589/	Karate	VS, CF, WM
P051	10.3389/fnhum.2022.924809	Tennis	WM
P055	10.1249/mss.0000000000001399	Gymnastics	CF, MA, WM
P057	10.37005/2687-0231-2020-0-5-20-31	Calligraphy	CF, MA, WM
P059	10.1080/01411920701609299	Increased physical activity	MA
P060	10.1038/s41598-023-32470-2	Basketball	CF, MA, WM
P063	10.3389/fnins.2018.00103	Art	VS, CF, MA, WM
P076	10.1177/1087054710379735	Physical education/sport	VS, CF
P082	10.1002/pits.22441	Abacus	VS, CF, MA, WM
P084	10.1016/j.appdev.2004.04.002	Taekwondo	VS, CF, WM
P086	10.1523/jneurosci.3195-18.2019	Abacus	VS, CF, MA, WM
P087	10.1016/j.neuroscience.2020.02.033	Abacus	CF, MA, WM
P092	10.3390/ijerph192316238	Swimming	MA
P102	10.1038/nature01647	Action video games	VS

Table 5 Distractor papers (70 topically matched controls harvested from Semantic Scholar)

ID	DOI	Title
P002	10.1186/s12889-016-3330-4	Effectiveness evaluation of a health promotion programme in primary schools: a cluster randomised controlled trial
P003	10.38159/ehass.202341222	Rural Early Childhood Educators' Perception of Music-Based Pedagogy in Teaching Communication Skills to Children
P004	10.1186/s12889-016-2971-7	Rationale and design of a randomized controlled trial examining the effect of classroom-based physical activity on math achievement
P007	10.1186/s12888-017-1433-9	ESCAschool study: trial protocol of an adaptive treatment approach for school-age children with ADHD including two randomised trials
P011	10.4172/2471-271x.1000159	Variability on the Spectrum: A Self-Monitoring Single-Case Design Study for Students with Autism Spectrum Disorders and Attentional Deficits
P012	10.1186/s41182-017-0074-5	The burden of dengue, source reduction measures, and serotype patterns in Myanmar, 2011 to 2015–R2
P013	10.1007/s41465-023-00271-0	How Do Executive Functions Influence Children's Reasoning About Counterintuitive Concepts in Mathematics and Science?
P014	10.1038/s41514-022-00093-y	Integrated cognitive and physical fitness training enhances attention abilities in older adults
P015	10.17267/2965-3738bis.2023.e4954	Transcranial direct current stimulation and cognitive stimulation therapy in children with autism spectrum disorder: randomized, sham-contro
P017	10.2196/40438	The Effects of Exergaming on Attention in Children With Attention Deficit/Hyperactivity Disorder: Randomized Controlled Trial
P020	10.58489/2836-8630/004	The Impact of A 12-Week Exercise Intervention on The Cognitive Functioning of Young People in Ghana
P022	10.5812/jmcl-148666	The Effect of Online Training of Competitive Domino and Cup Stacking Games on Mindfulness, Emotion Regulation and Executive Functions of Chi
P024	10.1186/s40359-018-0239-y	A protocol for a three-arm cluster randomized controlled superiority trial investigating the effects of two pedagogical methodologies in Swe
P025	10.26443/mjm.v14i2.199	Taxing Unhealthy Food: Is it an Effective Health Promotion Policy?
P026	10.1186/s12889-022-13886-3	Associations of physical activity with academic achievement and academic burden in Chinese children and adolescents: do gender and school gr
P029	10.1371/journal.pone.0082007	Two Randomized Trials Provide No Consistent Evidence for Nonmusical Cognitive Benefits of Brief Preschool Music Enrichment
P030	10.1017/s0142716419000535	Music effects on phonological awareness development in 3-year-old children
P033	10.1017/s1355617723002941	90 School-based Implementation of Educational and Neurocognitive Interventions in Children with Neurodevelopmental Disorders
P034	10.1186/s12887-019-1639-8	Study protocol and rationale of the "Cogni-action project" a cross-sectional and randomized controlled trial about physical activity, brain
P035	10.1016/j.psychsport.2019.101583	The Daily Mile^{TM} initiative: Exploring physical activity and the acute effects on executive function and academic performance in primary scho
P036	10.1038/srep05854	A Longitudinal Study on Children's Music Training Experience and Academic Development
P038	10.12779/dnd.2017.16.1.7	Improvement of Cognitive Function after Computer-Based Cognitive Training in Early Stage of Alzheimer's Dementia
P041	10.1073/pnas.1808412115	Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children
P042	10.1111/desc.13081	Individual Differences in Musical Ability are Stable Over Time in Childhood
P043	10.1037/dev0001307	Does Taekwondo improve children's self-regulation? If so, how? A randomized field experiment.
P044	10.1073/iti2818115	In This Issue

Table 5 (continued)

ID	DOI	Title
P046	10.30574/wjbphs.2023.15.1.0312	Working Memory (WM) training to improve ADHD in children, theory and the role of digital technologies
P047	10.4236/health.2014.616259	Investigating Efficacy of "Working Memory Training Software" on Students Working Memory
P048	10.1044/2022_jslhr-22-00253	Embedding Executive Function Training Into Early Literacy Instruction for Dual Language Learners: A Pilot Study
P049	10.53730/ijhs.v6ns2.8704	influence of music on the cognitive development of primary school children
P050	10.1002/aur.2735	Beyond group differences: Exploring the preliminary signals of target engagement of an executive function training for autistic children
P052	10.1371/journal.pone.0051474	Creativity in the Wild: Improving Creative Reasoning through Immersion in Natural Settings
P053	10.5507/ag.2016.021	The effect of internal and external focus of attention on game performance in tennis
P054	10.1007/s10802-023-01137-x	Callous-unemotional Traits and Child Response to Teacher Rewards, Discipline, and Instructional Methods in Chinese Preschools: A Classroom O
P056	10.21009/jpud.172.08	Introduction to Counting Symbols in Early Childhood with Stick Math (STIKMA) Educational Tool Games
P058	10.1186/s12889-019-6635-2	Integrating physical activity into the primary school curriculum: rationale and study protocol for the "Thinking while Moving in English" cl
P061	10.1186/s12887-017-0846-4	The Preschool Activity, Technology, Health, Adiposity, Behaviour and Cognition (PATH-ABC) cohort study: rationale and design
P062	10.2196/preprints.34915	Effects of Aerobic Exercise and High-Intensity Interval Training on the Mental Health of Adolescents Living in Poverty (Preprint)
P064	10.1038/s41598-022-18371-w	Author Correction: Effect of 5-weeks participation in The Daily Mile on cognitive function, physical fitness, and body composition in childr
P065	10.1186/s13102-015-0016-7	Sport-2-Stay-Fit study: Health effects of after-school sport participation in children and adolescents with a chronic disease or physical di
P066	10.1007/s12671-021-01802-6	The Role of Executive Function in Children's Mindfulness Experience
P067	10.1007/s13158-021-00297-5	Learning to Play the Piano Whilst Reading Music: Short-Term School-Based Piano Instruction Improves Memory and Word Recognition in Children
P068	10.1186/s12888-018-1811-y	Mindfulness for children with ADHD and Mindful Parenting (MindChamp): Protocol of a randomised controlled trial comparing a family Mindfulne
P069	10.5812/jmcl-147293	Effect of Sensory-Motor Integration Trainings on Executive Functions and Social Interactions of Children with High Functioning Autism Disord
P070	10.1038/s41539-017-0015-4	Dynamic, continuous multitasking training leads to task-specific improvements but does not transfer across action selection tasks
P071	10.1371/journal.pone.0058546	Enhancing Cognition with Video Games: A Multiple Game Training Study
P072	10.6007/ijarped/v8-i3/6424	Improving Executive Functioning Skills in Children with Autism through Cognitive Training Program
P073	10.1371/journal.pone.0029676	Brain Training Game Improves Executive Functions and Processing Speed in the Elderly: A Randomized Controlled Trial
P074	10.1038/s41598-022-14752-3	Reduced and delayed myelination and volume of corpus callosum in an animal model of Fetal Alcohol Spectrum Disorders partially benefit from
P075	10.32629/jher.v3i1.639	Research on the Intervention Program of Dance Therapy on the Joint Attention of Children with Autism
P077	10.1186/s12889-019-6742-0	Rationale and methods of the MOVI-da10! Study–a cluster-randomized controlled trial of the impact of classroom-based physical activity prog

Table 5 (continued)

ID	DOI	Title
P078	10.1186/s12883-015-0381-6	Mitii™ ABI: study protocol of a randomised controlled trial of a web-based multi-modal training program for children and adolescents with an
P079	10.31083/j.jin2204089	The Effect of Equine-Assisted Activities in Children Aged 7–8 Years Inhibitory Control: An fNIRS Study.
P080	10.2196/34915	Effects of Aerobic Exercise and High-Intensity Interval Training on the Mental Health of Adolescents Living in Poverty: Protocol for a Rando
P081	10.1007/s12671-024-02373-y	Does Virtual Reality Training Increase Mindfulness in Aboriginal Out-of-Home Care Children?
P083	10.1371/journal.pone.0253733	Breaking up classroom sitting time with cognitively engaging physical activity: Behavioural and brain responses
P085	10.54097/ehss.v8i.4608	The Importance and Training of Executive Functions among Children and Children with Autism Spectrum Disorder
P088	10.1186/s13063-018-2700-x	The impact of early-years provision in Children's Centres (EPICC) on child cognitive and socio-emotional development: study protocol for a r
P089	10.7759/cureus.46919	Efficacy of Mandala Coloring Intervention on Executive Functioning and Emotional & Motivational Self-Regulation Among Children With Symptoms
P090	10.24193/subbpsyped.2022.2.07	Teaching Music Education at Primary School in Romania: A Qualitative Analysis on Teachers Training
P091	10.1371/journal.pone.0072403	Enhanced Activation of Motor Execution Networks Using Action Observation Combined with Imagination of Lower Limb Movements
P093	10.1186/s12889-018-5514-6	High intensity intermittent games-based activity and adolescents' cognition: moderating effect of physical fitness
P094	10.1017/s135561772300293x	89 A Pilot Study of a Parent-Delivered Game-Based Cognitive Intervention in Children Born Preterm
P095	10.2196/46177	Tablet-Based Puzzle Game Intervention for Cognitive Function and Well-Being in Healthy Adults: Pilot Feasibility Randomized Controlled Trial
P096	10.1155/2018/2539748	Impact of Coordinated-Bilateral Physical Activities on Attention and Concentration in School-Aged Children
P097	10.1186/s13063-015-0992-7	Physical activity intervention (Movi-Kids) on improving academic achievement and adiposity in preschoolers with or without attention deficit
P098	10.1186/s40814-019-0468-8	A classroom-based intervention targeting working memory, attention and language skills in 4–5 year olds (RECALL): study protocol for a clust
P099	10.31031/rism.2020.06.000645	Physical Activity Interventions in School and their impact on Scholastic Performance
P100	10.5539/jedp.v12n2p1	Effect of Mindfulness Intervention on Inattentive Behaviors in Children at Risk for ADHD: A Single-Subject Study
P101	10.25726/e2988-5301-6762-b	Features of the organization of psychological and pedagogical work on the prevention of school maladaptation of high school students

References

1. N. Beutling, M. Spahic-Bogdanovic, Personalised course recommender: linking learning objectives and career goals through competencies. In: *Proceedings of the AAAI Symposium Series*, vol. 3. AAAI, Stanford, CA, USA (2024), pp. 72–81. https://doi.org/10.1609/aaaiss.v3i1.31185
2. F. Gao, S. Xu, W. Hao, T. Lu, KA-RAG: integrating knowledge graphs and agentic retrieval-augmented generation for an intelligent educational question-answering model. Appl. Sci. **15**(23), 12547 (2025). https://doi.org/10.3390/app152312547
3. T. Bukina, M. Khramova, N. Smirnov, S. Kurkin, A. Hramov, A neuro-symbolic approach to extracurricular activity recommendation based on cognitive profiling, in *Artificial Intelligence in Education. AIED 2026*, ed. by E.G. Blanchard,

G. Chen, M. Chi, S. Isotani. Lecture Notes in Computer Science, vol 16581 (Springer, Cham, 2026). pp 309–324. https://doi.org/10.1007/978-3-032-29744-0_21
4. T.V. Bukina, M.V. Khramova, S.A. Kurkin, D.A. Andrikov, S.S. Goman, A.E. Dedkov, A.E. Hramov, Neuroeducational software recommendation service as a tool for personalizing the educational process. Inform. Educ. **39**(5), 50–62 (2024). https://doi.org/10.32517/0234-0453-2024-39-5-50-62
5. V.V. Grubov, M.V. Khramova, S. Goman, A.A. Badarin, S.A. Kurkin, D.A. Andrikov, E. Pitsik, V. Antipov, E. Petushok, N. Brusinskii, T. Bukina, A.A. Fedorov, A.E. Hramov, Open-loop neuroadaptive system for enhancing student's cognitive abilities in learning. IEEE Access **12**, 49034–49049 (2024). https://doi.org/10.1109/ACCESS.2024.3383847
6. M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are few-shot clinical information extractors. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2022), pp. 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130
7. A. Dunn, J. Dagdelen, N. Walker, S. Lee, A.S. Rosen, G. Ceder, K. Persson, A. Jain, Structured information extraction from complex scientific text with fine-tuned large language models. arXiv preprint. arXiv:2212.05238 (2022). https://doi.org/10.48550/arXiv.2212.05238
8. S. Wadhwa, S. Amir, B.C. Wallace, Revisiting relation extraction in the era of large language models, In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)* (2023), pp. 15566–15589. https://doi.org/10.18653/v1/2023.acl-long.868
9. X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, W. Han, ChatIE: zero-shot information extraction via chatting with ChatGPT. In: arXiv Preprint. arXiv:2302.10205 (2023). https://doi.org/10.48550/arXiv.2302.10205
10. F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks. Proc. Natl. Acad. Sci. **120**(30), 2305016120 (2023). https://doi.org/10.1073/pnas.2305016120
11. S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: a roadmap. IEEE Trans. Knowl. Data Eng. **36**(7), 3580–3599 (2024). https://doi.org/10.1109/TKDE.2024.3352100
12. A. Baddeley, The episodic buffer: a new component of working memory? Trends Cogn. Sci. **4**(11), 417–423 (2000). https://doi.org/10.1016/s1364-6613(00)01538-2
13. A.M. Treisman, G. Gelade, A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980). https://doi.org/10.1016/0010-0285(80)90005-5
14. A. Miyake, N.P. Friedman, M.J. Emerson, A.H. Witzki, A. Howerter, T.D. Wager, The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cogn. Psychol. **41**(1), 49–100 (2000). https://doi.org/10.1006/cogp.1999.0734
15. M. Khramova, A. Hramov, A. Fedorov, Current trends in the development of neuroscientific research in education. Educ. Stud. Mosc. **4**, 275–316 (2023). https://doi.org/10.17323/vo-2023-16701

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.