The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume L-4/W3-2026
10th International Conference on Smart Data and Smart Cities (SDSC), 29 September–1 October 2026, Sofia, Bulgaria

Urban Heat-Aware Walkability Index Based on LLMs Analysis: A Case Study of Sofia

Denis Dimitrov[1], Lidia L. Vitanova[1*], Dessislava Petrova-Antonova[1]

[1]GATE Institute, Sofia University "St. Kliment Ohridski", 1164 Sofia, Bulgaria

Keywords: Walkability Index, Wet Bulb Globe Temperature, Large Language Models, Urban heat, Spatial Reasoning.

Abstract

A novel Artificial Intelligence (AI)-driven urban heat-aware walkability index (UHAWI) is proposed to assess pedestrian wellbeing and heat exposure. The approach integrates a conventional walkability index with the Wet Bulb Globe Temperature (WBGT) to assess how thermal stress affects the suitability of urban streets and areas for walking. AI methods are used to analyse spatial patterns and identify thermally comfortable and low-risk pedestrian zones. A Large Language Model (LLM) is used to analyse the calculated index and suggest an intervention. The pipeline is applied to street segments in Sofia, Bulgaria, across 5 consecutive July dates, producing UHAWI scores ranging from 23.1 to 66.5 (mean 43.9), which are then normalised for visualisation, indicating that summer thermal stress binds pedestrian conditions city-wide. Street segments with critical heat-walkability deficits are identified and targeted interventions are suggested, each grouped by empirically documented cooling and walkability effects, with street tree planting yielding the largest mean improvement (+13.3 points). Rubric-based evaluation of the LLM recommendations by a human (overall mean 3.25/5) and an independent LLM-as-judge (2.75/5) found strong spatial coherence and actionability but limited numerical faithfulness, identifying output grounding as the principal limitation of small locally deployed models. The proposed approach offers municipalities a reproducible, privacy-preserving decision support tool for prioritising climate adaptation investment at the street level, and the compute-then-reason architecture generalises to other urban analytics domains where AI assistance must remain locally executable and fully auditable.

1. Introduction

Rapid urbanisation, combined with the acceleration of climate change, has intensified urban heat stress, making it a critical public health concern. The Urban Heat Island (UHI) effect contributes to a consistent increase in street-level and inner-city temperatures, typically 2.0 to 5.0°C compared to surrounding rural areas, with the most pronounced impacts occurring in densely built districts during summer (Vitanova et al., 2021). Wet Bulb Globe Temperature (WBGT) is a widely used heat stress index that combines temperature, humidity, wind, and solar radiation to reflect how hot conditions affect the human body (Vitanova et al., 2026). Prolonged exposure to elevated WBGT levels imposes significant physiological strain on individuals, limiting outdoor activity, reducing walkability, and disproportionately affecting vulnerable groups, including older adults and outdoor workers (Kumakura et al., 2024). In response to these growing challenges, advances in geospatial data science and Artificial Intelligence (AI) are enabling more sophisticated approaches to environmental assessment. Multi-source data fusion pipelines can now integrate thermal datasets, pedestrian network data, and land-use information to provide a more comprehensive understanding of urban heat dynamics (Biljecki et al., 2023). Walkability indices have evolved into practical tools for urban planning and policy. However, existing approaches, such as those based on Shannon land-use or pedestrian diversity metrics, do not account for thermal stress (Petrova-Antonova et al., 2025a).

Such large-scale AI systems support the development of platforms and geospatial foundation models, enabling more advanced geographic information extraction and urban analysis (Manvi et al., 2023). However, important gaps remain, particularly in the application of AI-driven spatial climate interventions, especially in Eastern European cities with constrained open-data infrastructure. While Large Language Model (LLM)-based spatial reasoning has primarily been explored in high-resource environments with abundant data, its deployment in more resource-limited local contexts raises both technical and governance challenges. A Nature Computational Science article argues that LLMs have substantial untapped potential for urban planning, serving as intelligent assistants to human planners to synthesise design concepts, generate urban designs, and evaluate planning outcomes (Zheng et al., 2025). Concrete systems have begun to optimise this potential: UrbanLLM adapts spatio-temporal LLMs for city-scale prediction tasks, CityGPT empowers spatial cognition for urban reasoning tasks, and GeoLLM demonstrates that LLMs encode latent geospatial knowledge that can be extracted from downstream geographic inference (Feng et al., 2024) (Li et al., 2024). Parallel advances in small, locally deployable language models have made on-device inference increasingly feasible, raising the prospect of privacy-preserving, infrastructure-independent AI tools for municipal planning contexts that lack the data governance frameworks or computational budgets required for cloud-based AI services (Lu et al., 2024). Despite this progress, the use of LLMs as active spatial reasoning agents, rather than passive knowledge extraction tools, still remains largely unexplored.

Sofia city has undergone rapid urban densification, accompanied by increasingly extreme summer conditions. Its pedestrian environment is highly heterogeneous, shaped by a mixture of high-density residential buildings, historic street networks, and newer commercial developments, resulting in uneven walking conditions throughout the city (Petrova-Antonova et al., 2025b). Despite this exposure, no studies provide Sofia's municipal planners with a street segment resolution evidence base that jointly characterises thermal stress and walkability, leaving climate adaptation investment decisions without the spatial evidence needed for effective prioritisation.

To address the challenges and gaps mentioned above, this study introduces a new Urban Heat-Aware Walkability Index (UHAWI), supported by an LLM pipeline, for street-level thermal walkability assessment. It incorporates accessibility and WBGT as equal components, scoring them jointly to determine street-level liveability. UHAWI addresses the following research questions:

- *(RQ1) How can UHAWI be developed by integrating walkability metrics and WBGT to assess pedestrian heat exposure?*
- *(RQ2) Which street segments or spatial zones within Sofia exhibit the most critical thermal stress and walkability deficits, and which intervention scenario is the greatest improvement?*
- *(RQ3) To what extent can the LLM models produce intervention plans, and what are the limitations?*

The proposed UHAWI supports planning for tourism and recreational activities by identifying areas that combine high walkability with low WBGT-based thermal stress. It highlights pedestrian routes and public spaces that are more comfortable for walking, sightseeing, and outdoor recreation in hot conditions. This facilitates better route planning, improves visitor experience, and contributes to healthier urban environments.

The rest of the paper is structured as follows. Section 2 describes the data and methodologies used in this study. Section 3 provides the results with visualisation maps and model outputs. Section 4 concludes the study and future work.

2. Data and Methodology

Figure 1 illustrates the workflow, including data collection, preparation, index calculations, LLM reasoning and output reporting.

2.1 Study Area and Data Collection

Three primary data sources are used for the development of the proposed approach: 1) WBGT data simulated by a regional climate model (for 15-19 July 2024), 2) the walkability index data built upon previously developed street-level metrics (Petrova-Antonova et al., 2025a), and 3) pedestrian network data, including streets and roads related to urban walkability and preprocessed into street segments.

WBGT: The WBGT grids at 250 m spatial resolution were obtained for five consecutive summer days (15-19 July), each providing hourly WBGT values (00 to 23LST) per grid cell, yielding approximately 6,900 thermal observations per date. Figure 2 refers to the heat-health mapping in Sofia related to WBGT (Vitanova et al., 2026).

Walkability index: The building-level walkability scores were sourced from a pre-computed walkability score and two additional metrics: the Shannon Index (SHDI) and Simpson's Diversity Index (SDI), both expressed on a 0-100 scale. The pedestrian street network was derived from OpenStreetMap-sourced vector data filtered to retain pedestrian-relevant infrastructure, footways, paths, alleys, pedestrian crossings, underpasses, and pedestrian bridges, while excluding vehicular carriageways, parking areas, and administrative border outliers. Following filtering by feature type and a minimum segment length threshold of 50 m to remove fragmentary connector segments, the network comprised approximately 25,000 analysable street segments. Figure 3 refers to the walkability and diversity scores calculated for buildings in Sofia.

The study applies weighting factors and extreme-risk thresholds to construct a unified dataset of thermal stress values for each cell unit by aggregating building-level walkability scores. The LLM is used to generate the UHAWI for each grid cell. The stability of the index is then evaluated in different scenarios and methodological approaches. Finally, the analysis identifies the streets and areas with the highest thermal risk within Sofia's mixed urban environment.

2.2 Data Preparation

WBGT: WBGT values vary at the 250 m grid resolution and require finer spatial treatment: rather than sampling N evenly spaced points along its length (including both endpoints), and the nearest WBGT grid cell was assigned to each sample point within a distance cut-off, ensuring complete spatial coverage even for cells located near the periphery of the thermal grid extent. The N sampled WBGT values were averaged to produce a single representative thermal value per segment per date, improving accuracy for longer segments that may span multiple thermal grid cells.

Walkability index: For each street segment, building-derived scores (Walkability, PDI, SHDI) were assigned via an inverse-distance-weighted spatial join over all buildings within a 200 m radius of the segment midpoint, under the assumption that these scores represent smooth, slowly varying fields at the block scale.

2.3 Four-Class Classification and Index calculation

WBGT: Raw WBGT values were converted to a thermal comfort score ranging 0-100 using a four tier piecewise linear mapping grounded in the (of Biometeorology, 2023) heat stress classification: Caution (WBGT score $\leq$ 25 °C, walkability score 76-100), Warning (25 °C $<$ WBGT $\leq$ 28 °C, comfort score 51-75), Severe Warning (28 °C $<$ WBGT $\leq$ 31 °C, comfort score 26-50), and Danger (WBGT $\geq$ 31 °C, comfort score 0-25). Within each tier, comfort scores were linearly interpolated to avoid artificial discontinuities at tier boundaries, such that segments at 27.9 °C and 28.1 °C receive proximate rather than abruptly diverging scores. This approach presents a physiologically meaningful tier structure established in the heat stress literature while ensuring that the resulting final comfort score is continuous and suitable for downstream weighted aggregation.

Walkability index: The tier-mapped thermal comfort score is described in 2.1, Walk is the segment-level walkability score, and SHI and PDI are the corresponding land-use diversity scores. Thermal comfort is weighted at 0.50 to reflect the status of a co-equal primary determinant of street-level livability alongside the combined walkability and diversity dimensions, which together also sum to 0.50, split evenly across the 3 components. Because the raw walkability scores in the input data did not span the full theoretical 0-100 range (observed maximum around 68), walkability was min-max normalised across the full building dataset prior to weighting, ensuring the component contributed proportionally to its intended weight rather than being implicitly compressed.

https://doi.org/10.5194/isprs-archives-L-4-W3-2026-33-2026

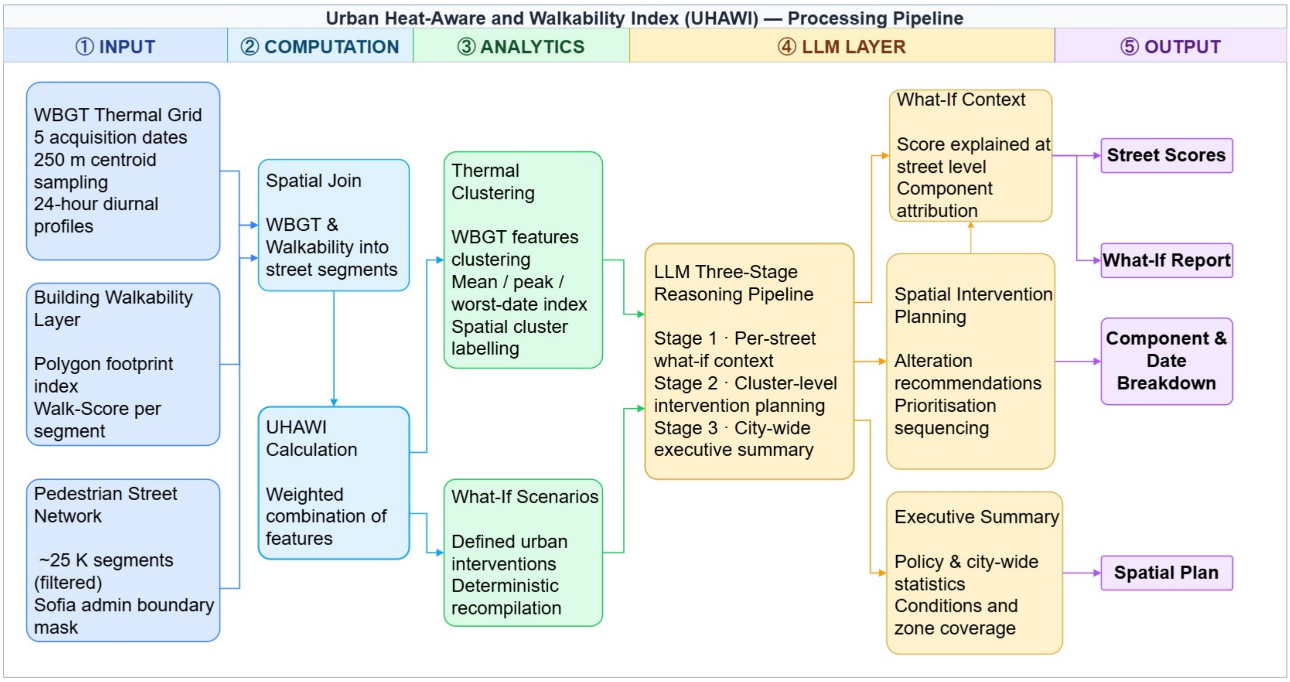

Figure 1. Study Workflow.

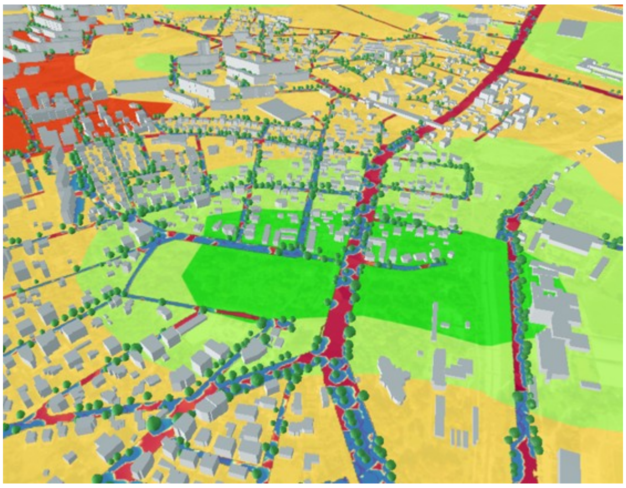

Figure 2. Segment of the visualisation of the Wet Bulb Globe Temperature (WBGT) Data.

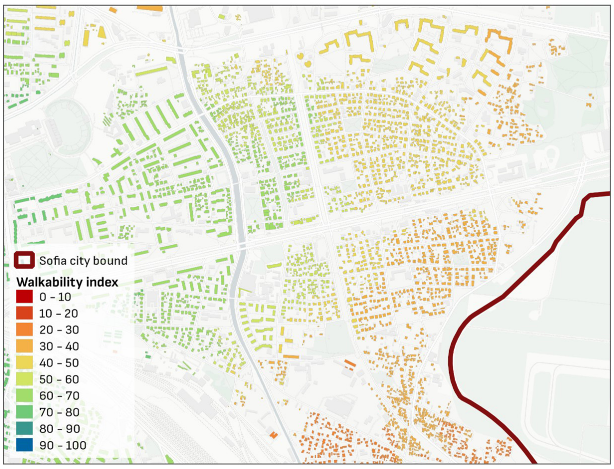

Figure 3. Segment from the visualisation of the Walkability index.

The UHAWI is composed as a weighted linear combination of four normalised components, all in the same 0-100 scale:

$$UHAWI = 0.5WBGT_c + \frac{0.5}{3}Walk + \frac{0.5}{3}SHDI + \frac{0.5}{3}SDI, \quad (1)$$

where c = principle distance
$WBGT_c$ = thermal comfort score
$Walk$ = segment level walkability score
$SHDI, PDI$ = diversity scores

Next, to support the LLM reasoning and interpretation layer described in Figure 1, three parallel classification schemes were defined. WBGT values are classified by physiological tier, as mentioned in 2.1. Walkability scores are classified in four tiers – Car Dependent (9-25), Some Walkable Locations (26-50), Somewhat Walkable (50-70), and Very Walkable (70-100) – adapted from conventional walking schemes (U.S, 2021). Finally, a heat walkability index is classified in four categories – Critical ($\leq$25), Poor (25-25); Moderate (50-75); and Excellent (75-100) – providing a single, interpretable label per street segment that synthesises the thermal and walkability dimensions for both visualisation and LLM analysis (Table 1).

2.4 Thermal clustering

The street segments were partitioned into spatially and thermally coherent zones using a K-means clustering algorithm (k=5) with 10 random restarts and a pure NumPy implementation, operating on three features per segment: mean WBGT across all observations, peak WBGT across all dates, and an encoded representation of the worst-performing observation date.

WBGT Score	WBGT Class	Walkability Score	Walkability Class	UHWAI Score	UHWAI Class
over 31 °C	Danger	0-25	Car Dependant	0-25	Caution
28 °C to 31 °C	Severe Warning	25-50	Some walkable places	25-50	Moderate
25 °C to 28 °C	Warning	50-70	Somewhat Walkable	50-72.5	Good
under 25 °C	Caution	70-100	Very walkable	72.2-100	Excellent

Table 1. UHAWI classification.

This clustering serves two purposes: 1) it identifies the worst-performing thermal zone to prioritise intervention planning, and 2) it enables the spatial reasoning stage of the LLM pipeline to reason over coherent thermal groups rather than the full network.

2.5 Deterministic What-If scenario Analysis

Four intervention scenarios were evaluated for every street segment using fixed, literature-driven parameter deltas applied deterministically: street tree planting (-1.5°C), high-albedo pavement materials (-1.0°C), and blue infrastructure such as water features (-1.5°C) (Bowler et al., 2010) (Giles-Corti et al., 2022) (Akbari et al., 2009) (Manteghi et al., 2015). For each scenario, the corresponding deltas were applied to the segment's baseline WBGT and walkability values. The UHAWI formula was re-evaluated, and the resulting gain, tier transition, and post-intervention classification were recorded. This deterministic design ensures that every intervention is replicable and directly traceable to a cited empirical source, in contrast to those that rely entirely on LLM-generated numerical estimates.

2.6 LLM-based reasoning

A locally deployed LLM, Qwen2.5, served via Ollama on consumer CPU hardware, operates as a three-stage reasoning agent over the pre-computed analytics described above in Figure 1. The LLM receives only the aggregated street-level scores and classification and never the underlying raw WBGT grid (approximately 34,000 cells), making local inference tractable and computationally viable.

In Stage 1, the LLM is prompted for each of the N lowest-scoring street segments, along with the segment's WBGT class, walkability class, and the precomputed scenario deltas. Based on this information, it generates a natural language explanation that contextualises why the computed deltas are appropriate for the segment's specific thermal conditions and walkability characteristics (Figure 4).

"Executive summary"
"You are an urban heat policy analyst. Write the executive summary. Cover: (1) overall heat-walkability state across thermal clusters, (2) which cluster demands immediate action, (3) top 2 interventions with specific impact numbers, (4) phased rollout strategy tied to lead times."

Figure 4. Prompt for segment analysis, given the calculated data.

In Stage 2, the LLM is provided with the complete street-level dataset for the high-risk thermal cluster, augmented with spatial proximity groupings identifying segments located within 300 m of one another and a tiered intervention budget. Based on this information, it generates a structured spatial intervention plan comprising priority street selection, justification for the formation of cool corridors, phased implementation recommendations, and an analysis of the trade-offs between competing intervention strategies (Figure 5).

In Stage 3, the LLM generates a city-wide executive summary from aggregated cluster-level statistics, intended for non-technical municipal stakeholders.

"PRIORITY STREETS:"
"List exactly the streets to intervene on first. For each, state which intervention and why this street over others. Identify the best contiguous corridor that can be formed from the proximity groups above. Explain what thermal and walkability benefit. Provide a phased plan: immediate, mid-term, long-term. Justify the sequence. Note the key trade-offs in this cluster: e.g..."

Figure 5. Prompt for intervention planning.

For what-if scenarios, deterministic interventions are applied to street segments, and later the spatial planning agent reasons for them, as described in Section 2.5. In the next stage, the worst-scoring segments are evaluated, and an explanation is given in the context of the urban and thermal domains. The locally deployed LLM serves as a spatial reasoning agent, computing street-level analysis to contextualise interventions, prioritising generation in the highest-risk zones, and producing a city-wide policy narrative.

3. Results and Discussion

This section presents the outcomes of applying the UHAWI pipeline to Sofia's pedestrian network, structured to answer the three research questions in turn: the spatial distribution and validity of the index (RQ1), the identification of priority zones and effective interventions (RQ2), and the evaluation of the LLM reasoning layer (RQ3), defined in Section 1.

3.1 Spatial index

The pipeline scored around 25,000 pedestrian network segments at 15:00 LST (peak temperature for Bulgaria), aggregated across five consecutive July dates.

Figure 6 reveals a strong contrast within residential areas. The streets in green colour consistently fall within the "Good" class, reflecting more comfortable thermal conditions and better walkability. In contrast, many streets in the residential areas score a moderate index, highlighting compounded challenges where high heat exposure coincides with less favourable walking conditions, an issue characteristic of dense inner urban fabric (orange colour).

High-scoring segments concentrate in the south-eastern part of the study area, correlating to the "Kanala" and the "Druzhba" corridor, where extensive three sections, lower building density, and proximity to green infrastructure and water bodies jointly suppress WBGT values while sustaining pedestrian infrastructure quality. Likewise, small Moderate zones appear near the city centre and other parks.

In contrast, the central and north-western urban fabric displays a predominantly Poor-to-Moderate signature, driven by elevated thermal stress in areas with impervious surface density and

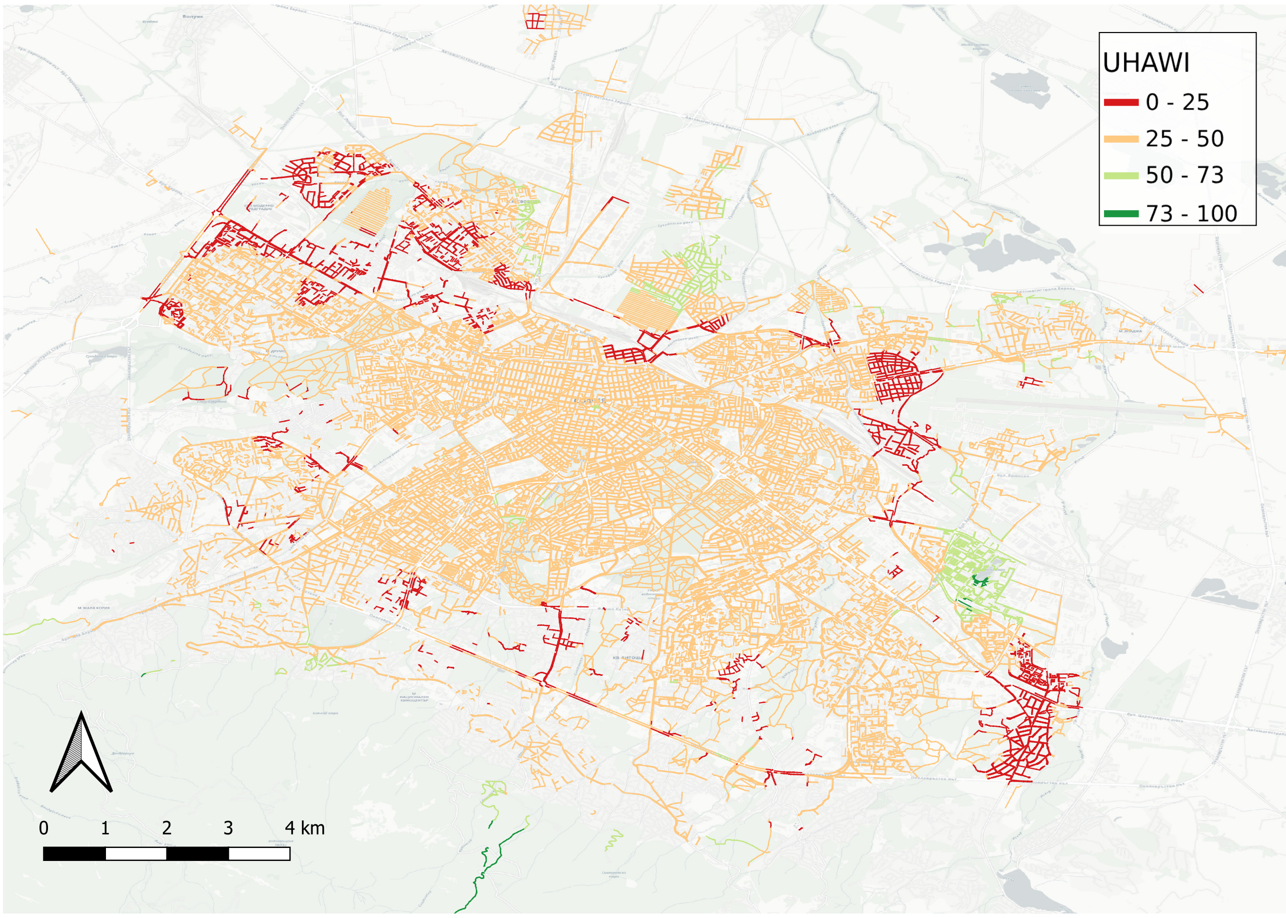

Figure 6. UHAWI score for all of the administrative areas of Sofia.

limited vegetation. Several isolated lower-scoring segments appear within otherwise moderate areas, suggesting localised heat traps associated with specific street-canyon configurations.

This pattern is consistent with the UHI structure in continental Eastern European cities and directly corresponds to the street-segment evidence base, the absence of which motivated this study (Santamouris, 2015).

For additional analysis, the UHAWI scores are discussed at the neighbourhood level, as shown in Figure 7. The results highlight clear spatial patterns in the relationship between heat exposure and walkability across the study area. High-scoring segments are concentrated in locations with lower building density and increased vegetation. For comparison, the same neighbourhood shows lower values for the purely walkability input metric, as shown in Figure 8, which illustrates the neighbourhood with street segments' walkability scores based on the building values in the input data.

The usefulness of the novice index is validated by examining whether it differs significantly from the baseline input metrics on which it is based (Figure 8) The comparison illustrates that the baseline walkability score, which is predominantly in the lower-walkability class, especially in the urban core, follows the same trend as the UHAWI (Figure 7) score, which, however, has "Moderate" score segments for the streets on the outside of the residential area, which are influenced by better thermal conditions.

A weight sensitivity analysis re-computed the UHAWI with different variations of the weight distribution: 4x0.25, 0.5 WBGT + 0.5 Walkability, 0.7 WBGT + 3x0.1 Walkability. The heavier thermal influence scenario was also calculated. Rankings were more stable but too close to the input WBGT distribution and the thermal clustering, and did not yield significant changes or meaningful data for later analysis. The substantive conclusions are that the identity of the priority zones and the ranking of interventions are therefore reasonable variations in the weighting prior to, and are not artefacts of the specific scheme.

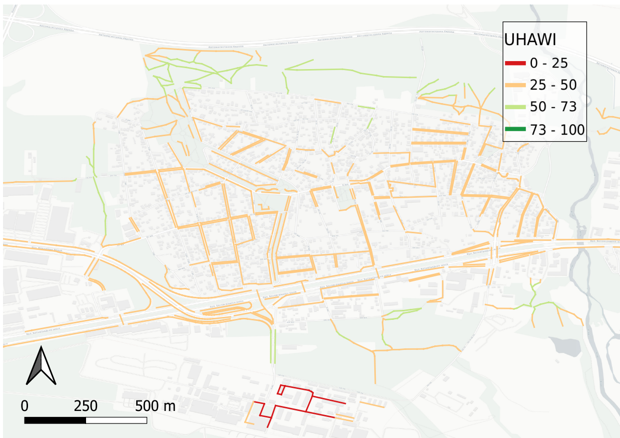

Figure 7. UHAWI street score at neighbourhood level at "Vrajdebna". The UHAWI scores are introduced in Table 1.

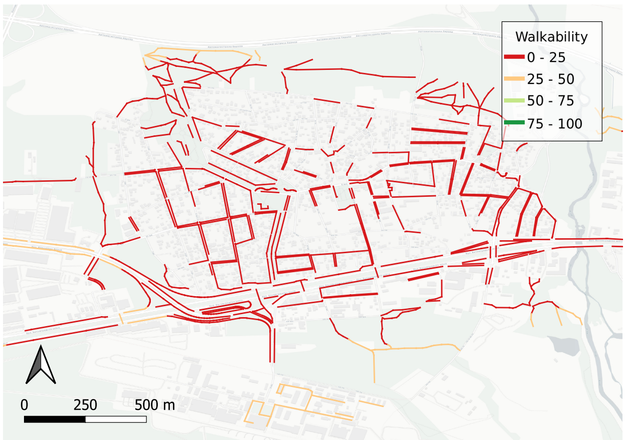

Figure 8. Walkability Index street score at neighbourhood level at ”Vrajdebna”.

3.2 Intervention analysis

The following section presents the results of the model-reasoning component of the pipeline and the analysis of the what-if scenarios. Figure 9 illustrates an example of a selected street segment with the worst thermal UHAWI score that requires immediate intervention. The potential for intervention is analysed, and a preselected method is suggested to lower the delta. In this specific case, tree planting is recommended to alleviate heat stress.

[PRIORITY STREETS]
seg$_9$1648
Intervention: Plant Street Trees
Reasoning: This street is part of Group 1, which has a high adjacency with other streets in the group. The UHAWI delta for this street is +13.3, indicating significant potential for improvement. Given that it falls within the immediate budget and has a medium lead time, planting trees will provide quick relief from heat while enhancing walkability.

Figure 9. LLM intervention output.

Figure 10 shows the LLM suggestion for a group of street segments forming a cool corridor. Tree planting is again the optimal intervention for the near segments at a street level.

[COOL CORRIDOR]
Group 1: seg_91648, seg_91649, seg_92171
Thermal Benefit: Forming a contiguous corridor of these streets will create a significant reduction in WBGT due to the combined effect of tree planting.

Figure 10. Cool corridor analysis.

Figure 11 illustrates the output for the highest-priority segments with their UHAWI and WBGT values and the date with the recorded highest temperature and class. The segments make up ”Malak rezan” street (Figure 12), which is located near the ”Malinova dolina” district and is in the ”Kambanite” park. It is near greenery, and the lower UHAWI may be skewed due to the lower walkability score.

The ”Malak rezan” street is in poor condition and requires renovation, but planting more trees is not optimal for the region, and the suggestion is not practical in a real-life scenario (Figure 12). Here, the bias of using predetermined intervention scenarios is evident, because the model does not account

HIGHEST-PRIORITY SEGMENTS

- {seg_43149; (Local alley) UHAWI= 23.1; WBGT=29.5°C; [Severe Warning] worst=18 july;}
- {seg_43445; (Malak rezan street) UHAWI= 23.8; WBGT=29.1°C; [Severe Warning] worst=18 july}
- {seg_43444; (Malak rezan street) UHAWI= 24.9; WBGT=29.1°C; [Severe Warning] worst=18 july}
- {seg_43436; (Malak rezan street) UHAWI= 24.9; WBGT=29.1°C; [Severe Warning] worst=18 july}
- {seg_43435; (Malak rezan street) UHAWI= 25.0; WBGT=29.1°C; [Severe Warning] worst= 18 july}

Figure 11. Worst performing segments output.

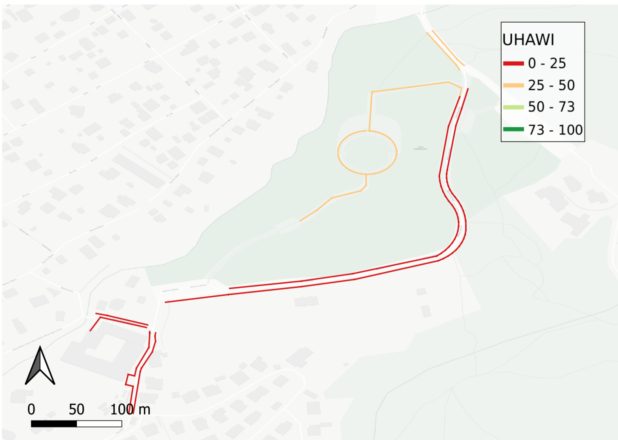

Figure 12. Illustration of the classification of ”Malak rezen” street.

for other aspects of the built environment and does not draw on sufficient background knowledge about the region. The suggested instances should be reviewed and validated by experts from the field domain.

In theory, the addition of green space produced the largest walkability gain (+8 points) (Gill et al., 2007), making it the preferred intervention for places with moderate thermal stress, but infrastructure quality remains the dominant deficit. High-albedo materials yielded the most modest absolute gain (-1 °C and no change in walkability) but remain relevant for scenarios where tree planting is structurally constrained (Akbari et al., 2009). Blue infrastructure performed intermediately on both dimensions (Manteghi et al., 2015).

Table 2 illustrates the effects of the intervention and how they change the calculated UHAWI score. For each segment, the model selects a practice from the above-defined literature to better suit its problematic region.

ID	WBGT	Scenario	After	Delta	Mechanism
9	46.27	water	53.53	7.26	Evaporation
11	46.22	greenspace	53.95	7.73	Vegetation
14	38.48	albedo	42.9	4.42	Cool roofs

Table 2. Examples of what-if scenario model output.

The results of the K-means clustering (Figure 13) on WBGT mean, WBGT peak, and worst-date profile partitioned the network into five thermally coherent zones ordered by severity,

https://doi.org/10.5194/isprs-archives-L-4-W3-2026-33-2026

with Cluster 0 – the highest thermal load zone – concentrated in the dense central urban fabric and comprising n = 44445 segments. The temporal dimensions of the clustering provided information that the thermal conditions were most severe on July 18th, when a substantially larger proportion of segments recorded WBGT values in the Severe Warning or Danger category. This weekly variation underscores the value of multi-date aggregation – a single-date assessment would either over- or undershoot chronic pedestrian heat exposure depending on the selected day, whereas with multi-date aggregation, the worst and a typical case are captured. In answer to the first part of RQ2, the segments exhibiting the most critical combined deficits are concentrated in Cluster 0 ("Malak rezen" street), characterised by a "Severe Warning" classification and a combination of high WBGT temperatures and lower UHAWI scores. After identifying the worst-performing cluster, the output of the spatial intervention plan is shown in Figure 13.

SPATIAL INTERVENTION PLAN

- *Plant Street Trees*: Expected to improve conditions by an average of +13.3 points, with a maximum impact of +15.3.
- *Add Green Space/Park*: Averaging +7.4 points and reaching up to +11.2.
- - *Phase 1 (Immediate)*: Focus on Cluster D for tree planting and green space additions, targeting the worst streets with UHAWI below 30.
- *Phase 2 (Short-term)*: Implement high-albedo building materials and water features in clusters A and B.
- *Phase 3 (Long-term)*: Expand green infrastructure city-wide. Immediate action is required on 518 streets within Cluster D (Warning-D), where mean WBGT is 25.4°C and UHAWI reaches 59.5. The top two interventions are:

  Executive Summary

  This study evaluates the urban heat environment in Sofia, Bulgaria, using 115,986 street segments over five dates at 15:00. The overall Urban Heat Air Index (UHAWI) mean is 43.9/100, with critical conditions affecting 92% of streets. Thermal clusters reveal severe heat issues in four clusters, particularly Cluster A and B, which account for 87% of the segments.

Figure 13. Intervention summary.

The LLM reasoning layer was evaluated using a four-criterion rubric-based method, since the aspects of the response cannot be fully captured by automated metrics (Hashemi et al., 2025). The scores were defined as the following: **1** - major errors, largely unusable; **2** - several important deficiencies; **3** - mostly correct with minor issues, **4** - accurate with negligible issues; **5** - excellent.

The criteria were defined as follows: **Faithfulness** - fidelity of the natural language output to the input values, **Spatial coherence** - consistency of the proposed subgroups. **Actionability** - specificity for a planner to act without interpretations. **Justification** - correct attribution of the deficit. After computing the output and intervention plan, the expert review yielded scores of 2 for faithfulness, 4 for spatial coherence, 4 for actionability, and 3 for justification quality on the 1-5 scale, as shown in Table 3. The LLM-as-a-judge cross-check, where the same model is applied to the same outputs under identical rhetoric by a larger model, shows agreement with the deficits in the human ratings.

Criteria	Human Score	LLM Score	Evidence
C1 Faithfulness	2	2	Intervention deltas are correct and match. Executive summary has errors in number calculations and naming conventions and has claims that are not proved.
C2 Spatial coherence	4	4	Group corridor correctly uses the given proximity group and is thermally consistent. The point is docked because some sections are truncated and not full.
C3 Actionability	4	3	Some of the prioritisations are inverted and contradicting.
C4 Justification	2	3	Some of the per-street segments have repetition, and the domain driver is not identified correctly.
Overall	3.25	2.75	Agreement on 2 out of 4 criteria.

Table 3. Human and LLM as an output judge.

The model scored well on spatial coherence and actionability, and moderately on justification, demonstrating that the 7B-parameter model running on consumer CPU hardware can organise pre-computed analytics into spatially sensible guidance. The faithfulness results suggest that the compute-then-reason separation still fails to prevent the model from contradicting or hallucinating beyond the deterministic analysis it receives. The compute-then-reason separation guarantees that reported numbers are correct, but it does not guarantee that the model's natural-language narration of these numbers is correct.

Qualitatively, the spatial intervention plan for the worst thermal cluster demonstrates that the reasoning, behaviours and architecture were designed to elicit the model correctly identifying proximity-grouped segments sharing the same worst-date profile as candidates for continuous cool corridors, deferred tree-planting and blue strategies for segments with adequate walkability that only need thermal focus change instead of infrastructure intervention. The model struggled to suggest practical interventions when the predetermined ones did not fit the scenario due to the architecture. The nature of the planning did not allow the model to fit in the urban context. The quantitative outputs of the pipeline remain fully trustworthy, but the interpretive layer requires verification before operational use. Taken together, this answers the research question that a locally deployed model using predetermined interventions does not produce coherent and faithful enough interventions ready for use by urban planners.

4. Conclusions and Future Work

The proposed UHAWI addresses key methodological and practical gaps by integrating thermal stress and walkability metrics into a unified, reproducible pipeline. Its design ensures applicability not only in data-rich environments but also in resource-constrained or privacy-sensitive municipal contexts, thereby supporting scalable, evidence-based urban climate adaptation.

Based on the results, the UHAWI identifies the danger zones properly and can be used to suggest cool corridors and places for heat escape and intervention in the city (RQ1). The worst-performing zones are identified, and interventions are suggested, based on predetermined operations, which should be made context-aware for future work, based on the faithfulness score (2.0) (RQ2). The reasoning layer is limited by the reason-then-compute architecture and the nature of the local-model hardware. The worst-performing zones are identified, and interventions are suggested, based on predetermined operations, which should be made context-aware for future work, based on the faithfulness score (2.0) (RQ2). The reasoning layer is limited by the reason-then-compute architecture and the nature of the local-model hardware. Better computational power can improve results, depending on the urban setting. However, privacy concerns may remain the biggest limitation (RQ3).

Future work will develop the LLM agent as a decision-support judge to approve or reject proposed street renovations and greening interventions. Evaluation will be expanded from individual scoring to a panel of planning practitioners to improve reliability and reduce scorer bias. In addition, incorporating data beyond WBGT and walkability will enhance the UHAWI formula, providing a more comprehensive assessment of urban thermal conditions.

Acknowledgement

This research is funded by the GATE project, the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155 and the programme "Research, Innovation and Digitalization for Smart Transformation" 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01, the deployEMDS project, funded by the Digital Europe Programme DIGITAL-2022-CLOUD-AI-03 under grant agreement no. 101123520; and the CU project, funded by the Digital Europe Programme DIGITAL-2023-CLOUD-AI-04 under grant agreement no. 101168005. We acknowledge access to the Discoverer supercomputer at Sofia Tech Park, Bulgaria. We thank Simeon Malinov for his continuous support.

References

Akbari, H., Menon, S., Rosenfeld, A., 2009. Global Cooling: Increasing World-Wide Urban Albedos to Offset CO2. *Climatic Change*, 94(3-4), 275–286.

Biljecki, F., Chow, Y. S., Lee, K., 2023. Quality of Crowdsourced Geospatial Building Information: A Global Assessment of OpenStreetMap Attributes. *Building and Environment*, 237, 110295.

Bowler, D. E., Buyung-Ali, L., Knight, T. M., Pullin, A. S., 2010. Urban Greening to Cool Towns and Cities: A Systematic Review of the Empirical Evidence. *Landscape and Urban Planning*, 97(3), 147–155.

Feng, J., Liu, T., Du, Y., Guo, S., Lin, Y., Li, Y., 2024. CityGPT: Empowering Urban Spatial Cognition of Large Language Models.

Giles-Corti, B., Moudon, A. V., Lowe, M., Cerin, E., Boeing, G., Frumkin, H., Salvo, D., Foster, S., Kleeman, A., Bekessy, S., De Sá, T. H., Nieuwenhuijsen, M., Higgs, C., Hinckson, E., Adlakha, D., Arundel, J., Liu, S., Oyeyemi, A. L., Nitvimol, K., Sallis, J. F., 2022. What next? Expanding Our View of City Planning and Global Health, and Implementing and Monitoring Evidence-Informed Policy. *The Lancet Global Health*, 10(6), e919-e926.

Gill, S., Handley, J., Ennos, A., Pauleit, S., 2007. Adapting Cities for Climate Change: The Role of the Green Infrastructure. *Built Environment*, 33(1), 115–133.

Hashemi, H., Eisner, J., Rosset, C., Van Durme, B., Kedzie, C., 2025. LLM-Rubric: A Multidimensional, Calibrated Approach to Automated Evaluation of Natural Language Texts.

Kumakura, E., Ashie, Y., Ueno, T., 2024. Assessing the Impact of Summer Heat on the Movement of People in Tokyo Based on Mobile Phone Location Data. *Building and Environment*, 265, 111952.

Li, Z., Xia, L., Tang, J., Xu, Y., Shi, L., Xia, L., Yin, D., Huang, C., 2024. UrbanGPT: Spatio-Temporal Large Language Models.

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., Xu, M., 2024. Small Language Models: Survey, Measurements, and Insights.

Manteghi, G., Limit, H. B., Remaz, D., 2015. Water Bodies an Urban Microclimate: A Review. *Modern Applied Science*, 9(6), p1.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., Ermon, S., 2023. GeoLLM: Extracting Geospatial Knowledge from Large Language Models.

of Biometeorology, J. S., 2023. Guidelines for the prevention of heat illness in daily life. Japanese Society of Biometeorology, Tokyo, Japan.

Petrova-Antonova, D., Murgante, B., Malinov, S., Nikolova, S., Ilieva, S., 2025a. Walkability Analysis of Sofia's Neighborhoods Powered by 15-Minute City Concept. *Cities*, 165, 106171.

Petrova-Antonova, D., Pavlov, V., Vitanova, L. L., 2025b. Assessing Walkability in Sofia: A Multi-Metric Index for Pedestrian-Friendly Cities. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 75–82.

Santamouris, M., 2015. Analyzing the Heat Island Magnitude and Characteristics in One Hundred Asian and Australian Cities and Regions. *Science of The Total Environment*, 512–513, 582–598.

U.S, E. P. A., 2021. National walkability index user guide and methodology. Technical report, U.S, Environmental Protection Agency.

Vitanova, L. L., Kusaka, H., Doan, Q.-V., Subasinghe, S., 2021. How Urban Growth Changes the Heat Island Effect and Human Thermal Sensations over the Last 100 Years and towards the Future in a European City? *Meteorological Applications*, 28(4), e2019.

Vitanova, L. L., Petrova-Antonova, D., Shirinyan, E., Khanh, D. N., Trendafilova, T. K., Subasinghe, S., Khan, A., Doan, Q. V., 2026. Fusing Urban Informatics and Weather Modelling in Digital Twins for City-Scale Heat-Health Mapping. *Computational Urban Science*, 6(1), 32.

Zheng, Y., Xu, F., Lin, Y., Santi, P., Ratti, C., Wang, Q. R., Li, Y., 2025. Urban Planning in the Era of Large Language Models. *Nature Computational Science*, 5(9), 727–736.